The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations

Samuel Bowles
Herbert Gintis*

July 22, 2003

Abstract

How do human groups maintain a high level of cooperation despite a low level of genetic relatedness among group members? We suggest that many humans have a predisposition to punish those who violate group-beneficial norms, even when this imposes a fitness cost on the punisher. Such altruistic punishment is widely observed to sustain high levels of cooperation in behavioral experiments and in natural settings.

We offer a model of cooperation and punishment that we call *strong reciprocity*: where members of a group benefit from mutual adherence to a social norm, strong reciprocators obey the norm and punish its violators, even though as a result they receive lower payoffs than other group members, such as selfish agents who violate the norm and do not punish, and pure cooperators who adhere to the norm but free-ride by never punishing. Our agent-based simulations show that, under assumptions approximating likely human environments over the 100,000 years prior to the domestication of animals and plants, the proliferation of strong reciprocators when initially rare is highly likely, and that substantial frequencies of all three behavioral types can be sustained in a population. As a result, high levels of cooperation are sustained. OUr results do not require that group members be related or that group extinctions occur.

*Forthcoming in *Theoretical Population Biology*. We would like to thank Christopher Boehm, Robert Boyd, Jung-Kyoo Choi, Leda Cosmides, Marcus Feldman, Steven Frank, Alan Grafen, Kristin Hawkes, Hillard Kaplan, Olof Leimar, Peter Richerson, Rajiv Sethi, Eric Alden Smith, E. Somanathan, Leigh Tesfatsion, Polly Wiessner, and Peyton Young for their help with this paper, the John D. and Catherine T. MacArthur Foundation and the Santa Fe Institute for financial support. Institutional affiliations: Santa Fe Institute and University of Siena (Bowles), Santa Fe Institute (Gintis). The authors can be contacted at bowles@santafe.edu, http://www.santafe.edu/~bowles and hgintis@comcast.net, http://www-unix.oit.umass.edu/~gintis.

1 Introduction

How do human groups maintain a high level of cooperation despite a low level of genetic relatedness among group members? The hypothesis we explore is that cooperation is maintained because many humans have a predisposition to punish those who violate group-beneficial norms, even when this reduces their fitness relative to other group members. Compelling evidence for the existence and importance of such altruistic punishment comes from controlled laboratory experiments, particularly the study of public goods, common pool resource, ultimatum, and other games (Yamagishi 1986, Ostrom, Walker and Gardner 1992, Fehr and Gächter 2002), from ethnographic studies of simple societies (Knauft, 1991; Boehm, 1984,1993), from historical accounts of collective action (Moore 1978, Scott 1976, Wood 2003), as well as from everyday observation.

Several plausible resolutions to the evolutionarily puzzle posed by altruistic punishment have been offered. If group extinctions are sufficiently common, altruistic punishment may evolve through the contribution of norm adherence to group survival (Boyd, Gintis, Bowles and Richerson 2003). Also, those engaging in punishment of norm violators may reap fitness benefits if their punishment is treated as a costly signal of some underlying but unobservable quality as a mate, coalition partner, or opponent (Gintis, Smith and Bowles 2001). Here we explore a different mechanism in which neither signaling nor group extinctions plays a role. Rather, punishment takes the form of ostracism or shunning, and those punished in this manner suffer fitness costs.

We hypothesize that where members of a group benefit from mutual adherence to a norm, individuals may obey the norm and punish its violators, even when this behavior incurs fitness costs by comparison to other group members who either do not obey the norm or do not punish norm violators, or both. We call this *strong reciprocity*. Strong reciprocity is altruistic, conferring group benefits by promoting cooperation, while imposing upon the reciprocator the cost of punishing shirkers.

It might be thought that *Cooperators*, who unconditionally cooperate but never punish would outcompete Reciprocators, who bear the cost of punishing norm violators. But as Cooperators replace Reciprocators in a group, the fraction of Selfish agents increases, and with Reciprocators less frequent in the group, the Selfish agents increasingly shirk, thereby attaining higher fitness than the Cooperators, and eventually replacing them. We model this dynamic below (see Figure 1). For this reason, Cooperators do not displace Reciprocators in the population as a whole. Moreover, in groups in which a low level of shirking is established, the expected costs of punishing others become quite low, so the within-group selection pressures operating against Reciprocators is weak.

We base our model on the structure of interaction among members of the mobile

hunter-gatherer bands in the late Pleistocene. Modern accounts of these societies record considerable variety in social organization and livelihood (Kelly 1995, Binford 2001). But widespread participation in joint projects such as hunting and common defense as well as the sharing of food, valuable information, and other sources of survival among many of these societies in the modern world is well established. A good case can be made that these cooperative projects were at least as important among our Late Pleistocene ancestors as they are among modern mobile foraging bands (Boehm 2002).

Our model therefore reflects the following empirical considerations. First, groups are sufficiently small that members directly observe and interact with one another, yet sufficiently large that the problem of shirking in contributing to public goods is present. Second, there is no centralized structure of governance (state, judicial system, Big Man, or other) so the enforcement of norms depends on the participation of peers. Third, there are many unrelated individuals, so altruism cannot be explained by inclusive fitness. Fourth, status differences are quite limited, especially by comparison to agricultural and later industrial societies, which justifies our treatment of individuals as homogeneous other than by behavioral type and by the group to which they belong. Fifth, the sharing on which our model is based—either of food individually acquired or of the common work of acquiring food, for example—is characteristic of these societies. Sixth, the individuals in our model do not store food or accumulate resources. This, too, is a characteristic of at least those hunter-gather bands based on what Woodburn (1982) calls an "immediate return" system of production.

Seventh, we take the major form of punishment to be *ostracism* and we treat the cost of being ostracized as endogenously determined by the amount of punishment and the evolving demographic structure of the population. This manner of treating punishment reflects a central aspect of hunter-gatherer life: since individuals can often leave the group to avoid punishment, the cost of being ostracized is among the more serious penalties that can be levied upon an individual group member.[1] Finally, behavioral heterogeneity is an emergent property of populations in our model, one that corresponds to what we know from the ethnographic record of foraging bands, as well as from the experimental evidence on both hunter gatherers (Henrich, Boyd, Bowles, Camerer, Fehr and Gintis 2003) and modern market based societies (Loewenstein, Thompson and Bazerman 1989, Andreoni and Miller 2002).

[1] There being relatively little individually held property, individuals cannot be severely punished by having their wealth confiscated; there being no fixed residence, individuals cannot be jailed or otherwise confined; extreme physical harm can be meted out against norm violators, but such measures are generally reserved for such serious crimes as adultery and murder.

2 Working, Shirking and Punishing Within a Group

Consider a population in which agents can live and work alone, in which case each has fitness $\phi_o < 0$. We assume reproduction is haploid. By the *fitness* of an agent, we mean the expected number offspring produced by the agent in one period minus the probability the agent dies in that period. Agents can also work cooperatively in a group, each producing an amount b at cost c (all benefits and costs are in fitness units). We assume that output of the group is shared equally by the agents, so if all group members work, each has a net group fitness benefit $b - c > 0$.

The group consists of three type of actors. The first type, whom we call *Reciprocators*, work unconditionally and punish shirkers. The second type, whom we call *Selfish*, maximize fitness. They never punish shirkers, and work only to the extent that the expected fitness cost of working exceeds the expected fitness cost of being punished. The third type, whom we call *Cooperators* work unconditionally but never punish shirkers.

Parents pass on their type to their offspring with probability $1 - \epsilon$, and with probability $\epsilon/2$, an offspring takes on each of the other two types. We call ϵ the *rate of mutation*. Also, with probability $1 - \epsilon$, Selfish agents inherit the estimate of $s > 0$ (the cost of being ostracized) from their parents. With probability ϵ an offspring is a mutant whose s is drawn from a uniform distribution on $[0, 1]$. Thus, s is an endogenous variable. Selfish agents with an s that is different from the objective fitness cost s^* of being ostracized may shirk too little or too much, leading to suboptimal fitness, a selective pressure for a modification of s.

We assume an ostracized agent works alone for a period of time before being readmitted to a group. The objective cost of ostracism therefore depends on the length of a spell of solitary living that occurs after leaving the group, and this changes over time as the distribution of the population between those in groups and those living alone changes. So s^* is also endogenous. In our model, Selfish agents with very high values of s behave exactly like Cooperators, except that if there are zero Reciprocators in the group, they do not work.

Suppose a Selfish agent shirks (that is, does not work) a fraction σ_s of the time, so the average rate of shirking is given by $\sigma = (1 - f_r - f_c)\sigma_s$, where f_r is the fraction of the group who are Reciprocators, and f_c is the fraction who are Cooperators. The fitness value of group output is $n(1 - \sigma)b$, where n is the size of the group. Since output is shared equally, each member receives $(1 - \sigma)b$. The loss to the group from a Selfish agent shirking is $b\sigma_s$. The *fitness cost of working function*, which can be written as $\lambda(\omega_s)$, where $\omega_s \equiv 1 - \sigma_s$, is increasing and convex in its argument (i.e., $\lambda', \lambda'' > 0$, with $\lambda(1) = c$ and $\lambda(0) = 0$). Expending effort always benefits the group more than it costs the workers, so $\omega_s b > \lambda(\omega_s)$ for $\sigma_s \in (0, 1]$. Thus, at every level of effort, ω_s, working helps the group more than it hurts the worker.

Further, we assume that the cost of effort function is such that in the absence of punishment of norm violators, the members face a public goods problem (i.e., an n-player prisoner's dilemma), in which the dominant strategy is to contribute very little or nothing.

We model punishment as follows. The fitness cost to a Reciprocator of punishing a shirker is $c_p > 0$. A member shirking at rate σ_s will be punished with probability $f_r\sigma_s$. Punishment consists of being ostracized from the group.

Selfish agents, given their individual assessment s of the cost of being ostracized, and with the knowledge that there is a fraction f_r of Reciprocators in their group, choose a level of shirking, σ_s, to maximize expected fitness.[2]

Writing the expected fitness cost of working, $g(\sigma_s)$, as the cost of effort plus the expected cost of being ostracized, plus the agent's share in the loss of output associated with one's own shirking, we have[3]

$$g(\sigma_s) = \lambda(1-\sigma_s) + sf_r\sigma_s + \sigma_s b/n, \tag{1}$$

Then, Selfish agents select σ_s^*, namely, that which minimizes the cost of working (1). Assuming an interior solution, this is given by

$$g'(\sigma_s^*) = \lambda'(1-\sigma_s^*) + f_r s + b/n = 0, \tag{2}$$

requiring the shirker to equate the marginal fitness benefits of expending less effort on work (the first and third terms) with the marginal costs of greater shirking, namely the increased likelihood of bearing the fitness cost of ostracism (the second term). Since $\lambda''(\omega_s) > 0$, there is at most one solution to this equation, and it is a minimum. This first order condition (2) shows that Selfish agents who inherit a large s (i.e., who believe the cost of ostracism is very high) will shirk less, as also will those in groups with a larger fraction of Reciprocators.

[2]Either of two informational assumptions justify our model. The first is that the type of an agent is unknown, but the fraction of Reciprocators is known to Selfish agents. The second is that the identity of Reciprocators is know to Selfish agents, but this cannot affect the probability of being caught shirking. It does not matter for our model whether or not Reciprocators can distinguish between Cooperators and Selfish agents, or whether they know the expected cost of ostracism used by a Selfish agent to determine the agent's shirking level. All of these informational assumptions can be weakened, at the cost of increased model complexity, but none is crucial to its operation.

[3]We obtain the second term by assuming each Reciprocator randomly chooses one agent to monitor in each period. Then the probability of being ostracized is

$$\sigma_s\left[1-(1-1/(n-1))^{f_r n}\right].$$

We approximate this by $\sigma_s f_r$, for simplicity. The approximation is very good for $f_r < 0.4$, but for higher f_r, ours is an overestimate. For instance, when $f_r = 1$, the actual cost is $0.66s\sigma_s$, but our estimate is $s\sigma_s$. This is a harmless but useful approximation, since when $f_r > 0.4$, there is always universal cooperation in our simulations, so Selfish agents rarely pay the cost of being ostracized.

The expected contribution of each group member to the group's population in the next period is equal to the member's fitness minus (for the Selfish agents) the likelihood of ostracism. This gives

$$\pi_s = (1-\sigma)b - \lambda(1-\sigma_s) - f_r\sigma_s, \tag{3}$$

$$\pi_c = (1-\sigma)b - c, \tag{4}$$

$$\pi_r = (1-\sigma)b - c - c_p(1-f_r-f_c)\sigma_s, \tag{5}$$

where the subscripts s, c, and r refer to Selfish agents, Cooperators, and Reciprocators. The final term in the expression for π_r follows because each Reciprocator chooses a random agent to monitor; this agent is Selfish with probability $(1-f_r-f_c)$ and this agent shirks with probability σ_s.

We assume that at the end of each period, groups admit a number of new members equal to a fraction μ of their existing numbers. Candidates for immigration into groups are the pool of solitary individuals, plus a fraction γ of current group members who want to emigrate for exogenous reasons (e.g., to find a mate, or to end a personal dispute with another group member). If the number of candidates exceeds the number of places, a random sample of these candidates emigrate to groups. We assume that group members that desired to emigrate but did not find a receptive group remain in their current group.

The number of groups and the total population is fixed. A simulation starts with all groups of the same size, which we call the *initial group size*. Individual group size will, of course, change from period to period. It is plausible that there is an optimal group size, excessive deviation from which leads to efficiency losses. Rather than explicitly building such losses into our model, we assume that if a group falls below some $n_{\min}$ in size, it disbands, the remaining members migrating to the pool. We further assume that the vacated site is repopulated by migrants randomly chosen from the most populous remaining groups, restoring the initial group size. These two procedures ensure that extremes in group size are avoided (details of the simulation structure are presented in the Appendix).

We can gain more insight into the dynamics of this system by choosing a specific function $\lambda(1-\sigma)$ satisfying the conditions $\lambda(1)=c$, $\lambda(0)=0$, $\lambda'(1-\sigma)<0$, and $\lambda''(1-\sigma)>0$. Extensive simulations suggest that the exact form of this function is unimportant, and the simplest function satisfying these conditions is

$$\lambda(1-\sigma) = c(1-\sigma)^2. \tag{6}$$

Using (6) and (2), it is easy to check that Selfish agents then shirk according to the function

$$\sigma_s(f_r) = \begin{cases} 1 - \frac{f_r s n - b}{2cn} & \text{for } f_r \le f_r^{\max} = \frac{2c+b/n}{s} \\ 0 & \text{for } f_r > f_r^{\max}. \end{cases} \tag{7}$$

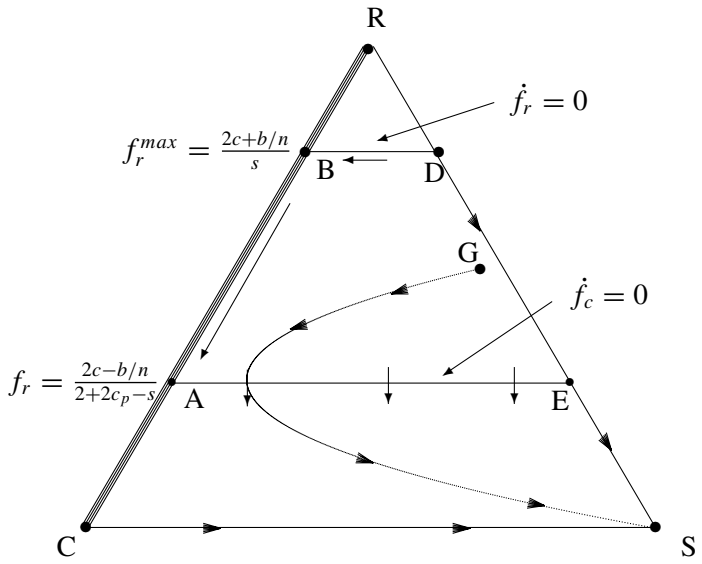

Figure 1: Within Group Dynamics with Ostracism but No Migration. This diagram is based on $s > 2c$ (the cost of being ostracized is greater than twice the cost of working for one period). The value of f_r for which $\dot{f}_r = 0$ and the value of f_r for which $\dot{f}_c = 0$ are based on the cost of effort function (6) below.

A phase diagram of within-group dynamics, abstracting from ostracism and group dissolution, appears in Figure 1. In this figure, each point in the simplex is a distribution of behavioral types in the population. The vertex S refers to the all-shirking group composition ($f_r = f_c = 0$), vertex R refers to the all-Reciprocator composition ($f_r = 1$), and C refers to the all-Cooperator case ($f_c = 1$). For the parameter values illustrated in the figure, there is some fraction of Reciprocators $f_r^{max} = (2c + b/n)/s$ (the line BD) at and above which Selfish agents do not shirk at all, but below which they shirk at a strictly positive rate. Thus in the triangle RBD all three types have equal payoffs, equal to $b - c$. Along the whole CR segment, Reciprocators and Cooperators do equally well, since there are no Selfish agents to punish (the payoffs are again $b-c$). Along the CS segment, Reciprocators are absent, so Selfish agents do better than Cooperators. On the segment DS, Selfish agents do better than Reciprocators. Selfish agents optimize, so when $\sigma_s \in (0, 1)$, we know from (2) that an increase in f_r holding the frequency of Selfish agents constant, must entail a decline in σ_s. This, along with equation (7), means that lowering the fraction of Selfish agents increases the payoff to Reciprocators relative to Selfish agents in the area CBDS. Moreover, Cooperators always have higher payoffs than Reciprocators in the interior of this area. We conclude that the only asymptotically

stable equilibrium of the system is the all-Selfish point S, and its basin of attraction consists of all interior points below the line BD in Figure 1.

Clearly, then, if cooperation is to be sustained, it must be because Cooperators, who undermine the cooperative equilibrium by driving out Reciprocators, must themselves be harmed by shirking Selfish agents when Reciprocators are rare. Figure 1 illustrates exactly this process, with Selfish agents proliferating at the expense of both Cooperators and Reciprocators once the frequency of Reciprocators falls below the line AE in the figure. We shall show by simulation that this is indeed the case for a wide range of plausible parameter values.

Figure 2 shows the group average payoffs as a function of the distribution of types within a group. The curved lines are iso-group-average-fitness loci, showing clearly that average group fitness increases as we move away from the unique stable equilibrium S.

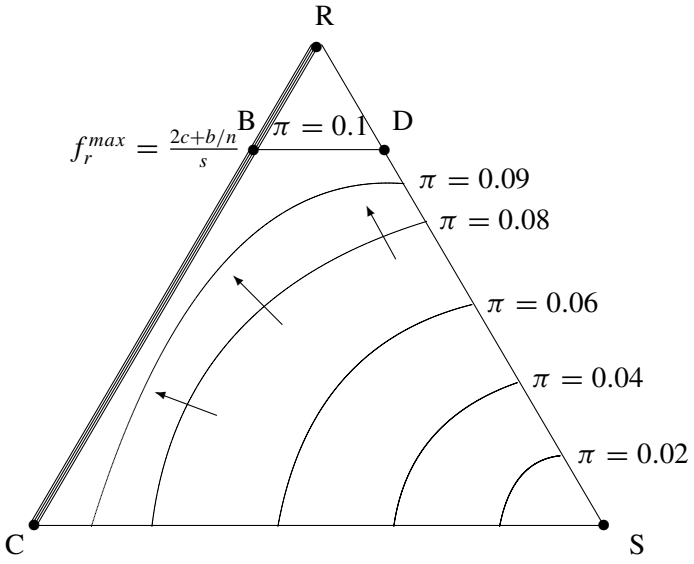

Figure 2: Average Payoffs and Group Composition. This diagram is based on $b = 0.2$, $c = c_p = 0.1$, and $s = 0.3$. The arrows point towards increasing payoffs. For all frequencies in BDR, $\pi = 0.1$

In the next section we consider the evolution of a number of groups, each occupying a site allowing cooperative production as described above. The resulting model is too complex to admit an analytical solution, so we provide a series of simulations that illustrate its characteristics over an appropriate range of parameter values.

3 Simulating Strong Reciprocity

For our baseline simulation, we set up twenty groups, each starting out with twenty members, and an empty pool. We set the initial frequency of Selfish agents at 100%, and we assigned each Selfish agent a cost of being ostracized (s), using a random number drawn from a uniform distribution on the interval [0, 1]. We assume that immigration into a group occurs at a rate of 12% per generation, which we take to be four simulation periods. Therefore the immigration rate is $\mu = 0.03$.

Value	Description
0.2	Output per Agent, no Shirking (b)
0.1	Cost of Working, no Shirking (c)
0.1	Cost of Punishing (c_p)
0.05	Emigration Rate (γ)
0.03	Immigration Rate (μ)
20	Initial Group Size (n)
20	Number of Groups
-0.1	Fitness in Pool (ϕ_0)
6	Minimum Group Size
[0, 1]	Initially Seeded Expected Cost of Ostracism (s) Uniformly Distributed on this Interval
0.01	Mutation Rate (ϵ)

Table 1: Baseline Parameters. These parameters are used in all simulations, unless otherwise noted. All simulations start with a homogeneous population of Selfish agents.

We assume a desired emigration rate of $\gamma = 0.05$ per period, so more agents would like to emigrate from groups than are admitted into groups. We set the fitness cost of being in the pool to be equal to the cost of working (as Figure 6 shows, this figure is not at all critical). We also set individual productivity, net of the cost of working, equal to the cost of punishing a shirker. Simulations show that it matters little how large this parameter is, and we set it to 0.1 in our baseline simulations. Since we assume the cost of working is $c = 0.1$, the cost of punishing becomes $c_p = 0.1$ and individual productivity is $b = 0.2$. The reason absolute productivity does not matter is that we cull or augment the population randomly at the end of each period to maintain a constant population size. Effectively, fitness figures are therefore relative. The baseline parameters are listed in Table 1.

Using these parameters, Figure 3 shows the evolution of the distribution of agent types and the average shirking rate for the whole population in the initial 3,000 periods of a typical run. The results shown are backward-moving averages over 100 periods. To determine the typical behavior of the model, we ran the simulation 25 times with the baseline parameters for 30,000 periods, and calculated

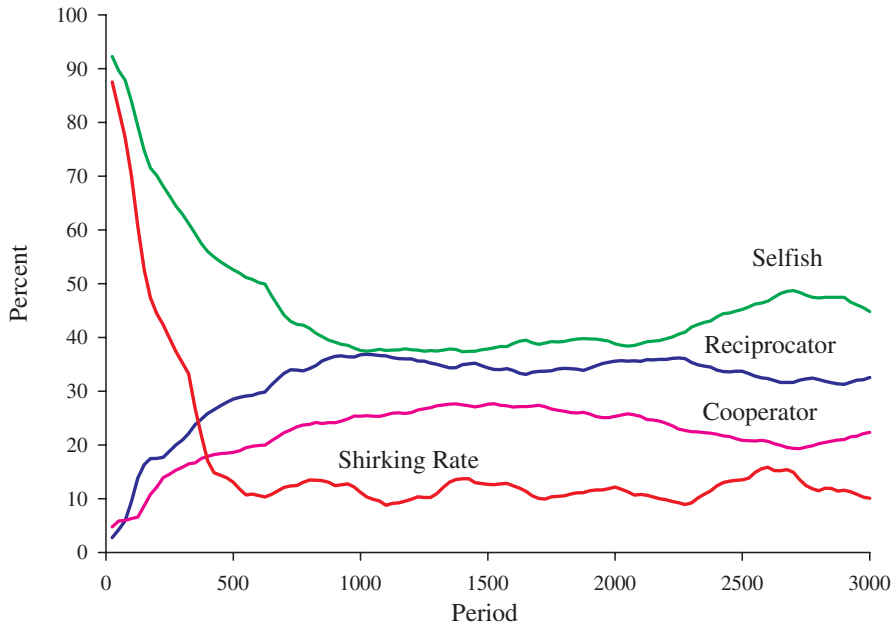

Figure 3: Initial Periods of a Single Simulation Run. As is clear from the figure, and as in all our simulations, the initial population consisted of Selfish agents alone. The baseline parameters are as in Table 1.

the average fraction of each type and the average shirking rate, averaged over the last 1000 periods. These averages are reported in Table 2. There was remarkably little variation across the runs–in all cases the standard error of the frequencies reported is less than 1.14 percentage points.

Why does the long-run behavior of the system involve roughly equal numbers of the three types of agents? Table 2 shows that over this period, Cooperators were slightly more likely (0.48%) than Reciprocators (0.38%) to find themselves in the pool of solitary agents, while Selfish agents were more than an order of magnitude more likely to be in this position (10%). This is in part because Cooperators are more likely than Reciprocators to be in disbanded groups (by a ratio of 1.21 to 1), and the Selfish agents are more likely than Reciprocators to be in disbanded groups (by a ratio of 3.4 to 1). But ostracism, not disbanding, is overwhelmingly important in populating the pool with Selfish agents. Indeed, we have found that even when groups are not disbanded until a single agent is left in the group (at which time it is

Long-run Values	
Value	Description
37.2%	Fraction of Reciprocators
24.6%	Fraction of Cooperators
38.2%	Fraction of Selfish Agents
11.1%	Average Shirking Rate
4%	Fraction of Population in Pool
0.38%	Fraction of Reciprocators in Pool
0.48%	Fraction of Cooperators in Pool
10%	Fraction of Selfish agents in Pool
4%	Fraction of Pool who are Reciprocators
3%	Fraction of Pool who are Cooperators
93%	Fraction of Pool who are Selfish
1.21	Ratio of Cooperators to Reciprocators in Disbanded Groups
3.4	Ratio of Selfish Agents to Reciprocators in Disbanded Groups

Table 2: Long Run Simulation Statistics. The baseline parameters are as in Table 1. The statistics represent the average of the last 1000 periods of a 50,000 period simulation, averaged over 25 simulations.

no better than solitary production), similar long-run values of the major variables obtain. Thus, the dispersion of members of very small groups, while empirically realistic, is not crucial to the model's workings.

To illustrate that the dynamics depicted in Figures 1 and 2 are operative, in Figure 4 we trace a bit of the history of a single group, from "birth" in period 320 (repopulation of a site after another group had disbanded) to "death" in period 390 (disbanding because group size fell below the minimum sustainable group size). This group began with roughly equal numbers of the three agent types, but early on, Selfish agents were driven out by the Reciprocators and were replaced by Cooperators. Starting in period 330, Cooperators began displacing Reciprocators until, about period 340, the group consisted of many Cooperators and a few Selfish agents (i.e., near the C vertex in Figure 1). With few Reciprocators present, the expected cost of shirking fell dramatically, shirking rose, and the Selfish therefore outperformed the Cooperators. The numbers of Selfish therefore grew rapidly from period 340 to 350, displacing Cooperators. This process reproduces the dynamic illustrated by the curved arrow QS in Figure 1. From this point, average group fitness was low (the group composition placed it close to the S vertex in Figure 2) and hence, despite a small chance infusion of immigrant Reciprocators, the group loses members until it disbands in period 390. Figure 4 also shows the effect of these demographic movements on the shirking rate and group size. We see that the

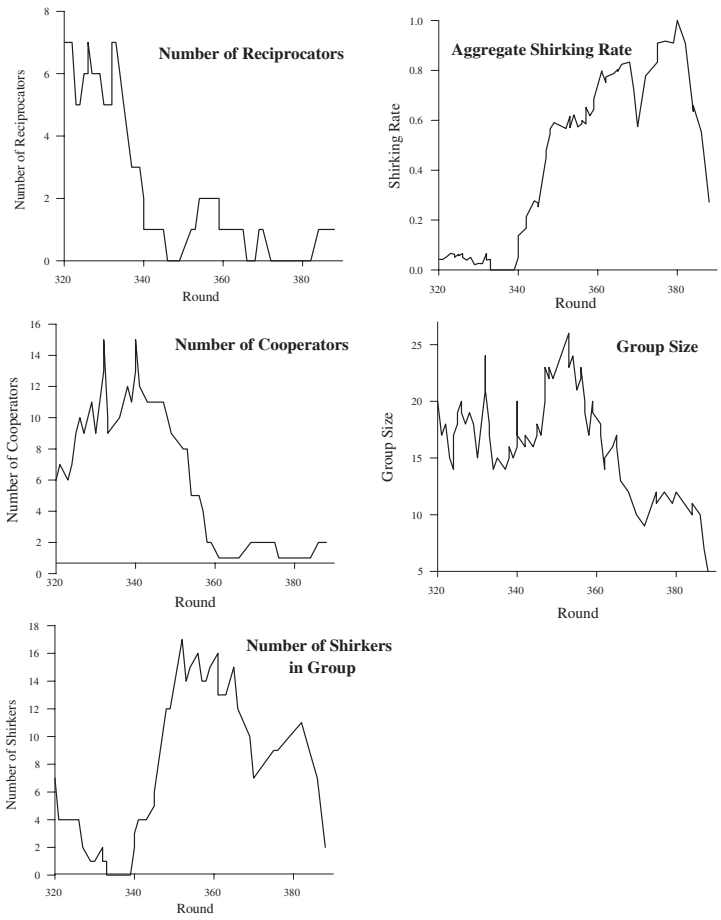

Figure 4: Snapshot of the Life History of a Single Group, from Repopulation of a Site to Disbanding: Membership Composition, Size and Efficiency. The baseline parameters are as in Table 1.

shirking rate remains very low until period 340, and then climbs steadily until just prior to the end of the group's life. Near the end there is an increase in group fitness because a couple of Reciprocators migrate into the group and the Selfish respond by shirking less, but the high level of shirking combined with the high cost of deterring it through ostracism reduces group size below the minimum before it can recover.

Our interpretation of this group's life course and demise illustrates the population-level forces preventing Cooperators from driving out Reciprocators. We can formalize this reasoning considerably, and subject it to an empirical test. Cooperators in our model are *negative altruists*, who have a fitness advantage within groups, offset by the lower average fitness of groups with a high frequency of Cooperators. If the group effect is sufficiently strong. Cooperators do not drive out Reciprocators in the population, even though they eliminate Reciprocators from the groups in which they coexist. We can use Price's equation (Price 1970) for the analysis of multi-level selection processes to demonstrate this.

Suppose there are groups $i = 1, \ldots m$, and let q_i be the fraction of the total population in group i. Suppose π_i average fitness of members of group i, and let f_c^i, f_r^i, f_s^i be the fraction of Cooperators, Reciprocators, and Selfish in group i. Price's equation for the change in the fraction of the entire population who are Cooperators, Δf_c, can then be written as

$$\bar{\pi}\,\Delta \bar{f_c} = \sum q_i(\pi_i - \bar{\pi})\, f_c^i + \sum q_i \pi_i f_c^i \pi_c^i, \tag{8}$$

where $\pi_i = f_c^i \pi_c^i + f_r^i \pi_r^i + f_s^i \pi_s^i$ and $\bar{\pi} = \sum q_i \pi_i$.

The first summation in (8) is the covariance between average group fitness and the share of Cooperators in the group. The interpretation offered above and the group history depicted in Figure 4 suggests this term will be negative, since groups with a high frequency of Cooperators are also likely to have a high frequency of Selfish agents who engage in high levels of shirking. These groups thus will have lower than average mean fitness. The second summation is the expected change in within-group Cooperator fitness, which we expect to be positive since, unless Selfish agents are entirely absent, Cooperators free-ride on the altruistic punishment of shirkers by Reciprocators. Our model is too complex to solve Price's equation analytically, but we can show that our description of Cooperator fitness is accurate, since when we use values from the simulation in (8), the two summations have the expected signs, and approximately offset each other, over many thousands of rounds of simulation, thereby explaining why f_c oscillates around a stationary long term average (as illustrated in Figure 3).

Figure 5 show the movement of the terms of the Price equation (8) for this simulation. As predicted, the covariance between average group fitness and the share of Cooperators is negative, the expected change in within-group Cooperator

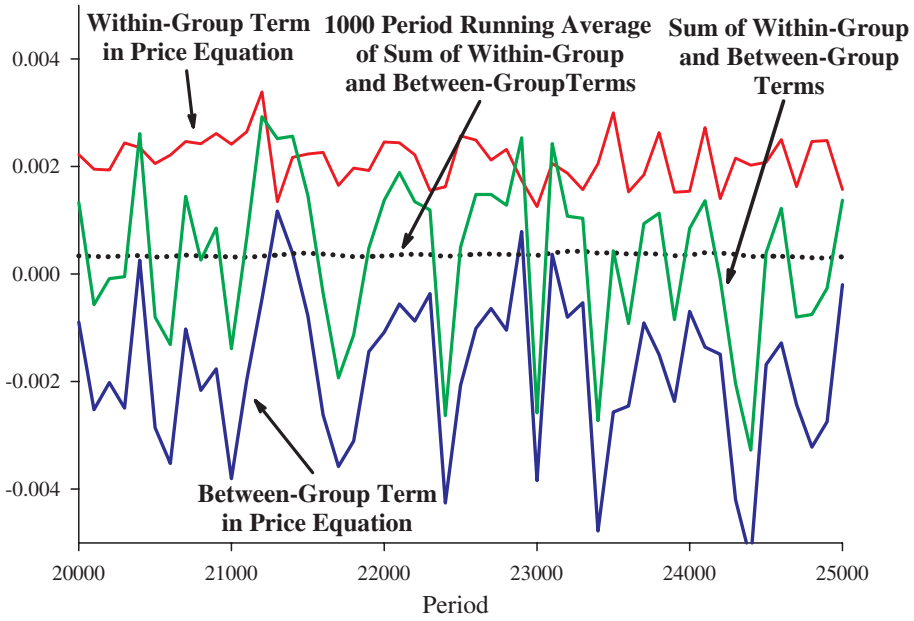

Figure 5: Price's Equation Applied to a Dynamic Simulation. The baseline parameters are as in Table 1. This figure shows a 5000 period segment of a 30,000 period run. Measurements were taken every 100 periods. The within-group term measures the tendency of Cooperators to enjoy a fitness advantage within groups due to the fact that they do not punish, while the between-group term measures the fitness disadvantage of Cooperators across groups due to the fact that members of groups with a large fraction of Cooperators have low mean fitness. The sum of the two terms is highly variable over a small number of periods, but average to zero, since the fraction of Cooperators is stable in the long run. Over the 5,000 period segment depicted, the Fraction of Cooperators rose slightly, which accounts for the small positive bias in the long-run average of within- and between-group terms.

fitness is positive, and the two are approximately offsetting over many thousands of periods.

4 Variations and Extensions

In further simulations, we explore a large parameter space with two objectives. First, we will check that the simulation results respond in plausible ways to parameter shifts, thus illuminating the causal structure of the model. Second, we will confirm that the model works for values approximating human ancestral environments.

The first three panels of Figure 6 show that the model responds in the expected way to changes in parameter values, and that high levels of cooperation are sustained for a quite large parameter space. These panels show the average shirking value, plotted against the fitness of agents per period in the pool of solitary agents, the cost to Reciprocators of punishing shirkers, and initial group size. These figures are the average of the last 1000 periods of ten runs of 30,000 periods. In each case, we see that the shirking rate, which is the best aggregate measure of group cooperation, is small or moderate in size over a large range of parameters, and moves in the expected direction with parameter changes. Our results are quite insensitive to variations in other parameters, including the immigration rate, the emigration rate, and minimum group size. In each case, we find cooperation to be robust over a wide range of parameters.

Our simulations shows that the ostracism mechanism promotes high levels of cooperation in groups and that a substantial fraction of the population are Reciprocators in the long term steady state of this model under a wide range of parameter values. But how could Reciprocators come into existence *de novo*? We speculate that it could have emerged through a rather trivial modification of fitness-enhancing behaviors. For example, a good case could be made that strong reciprocity among kin could emerge and proliferate through as a form of kin-based altruism, and then be generalized to unrelated individuals. Another possibility is that strong reciprocity arose through a modification of the individual fitness enhancing strategy of reciprocal altruism. In this case the modification is trivial: simply ignore the future payoffs to current behavior. Like the extension of kin-based strong reciprocity to non-kin, the mutation or mutations to convert reciprocal altruist strategy to a strong reciprocal one involves a reduction in complex discrimination rather than an increase in complexity. These strategy conversions thus might occur with high probability.

As we have seen (Figure 3), the model supports the emergence of strong reciprocity after as few as 500 periods. This occurs because at the rate of mutation assumed, it does not take many periods before at least one group will have enough Reciprocators to implement a high level of cooperation. When this occurs, the co-

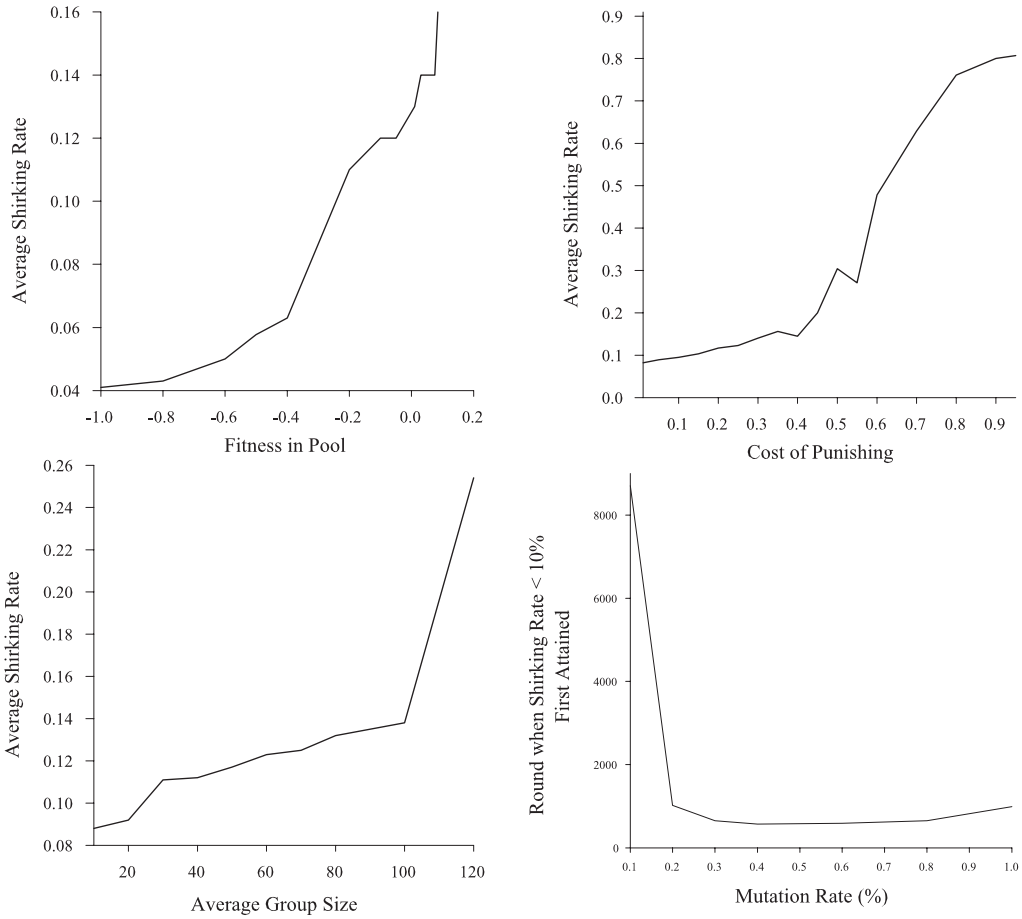

Figure 6: The Effect on Shirking of the assumed values of Fitness in Pool, Cost of Punishing, Group Size, and the Mutation Rate. The baseline parameters are otherwise as in Table 1. For average group size larger than about 120, Reciprocators rarely invade an initially Selfish population. The Mutation Rate panel shows, for a population of 50 groups, the average over 25 runs of the number of periods until the shirking rate, averaged over previous 1000 periods, first fell below ten percent.

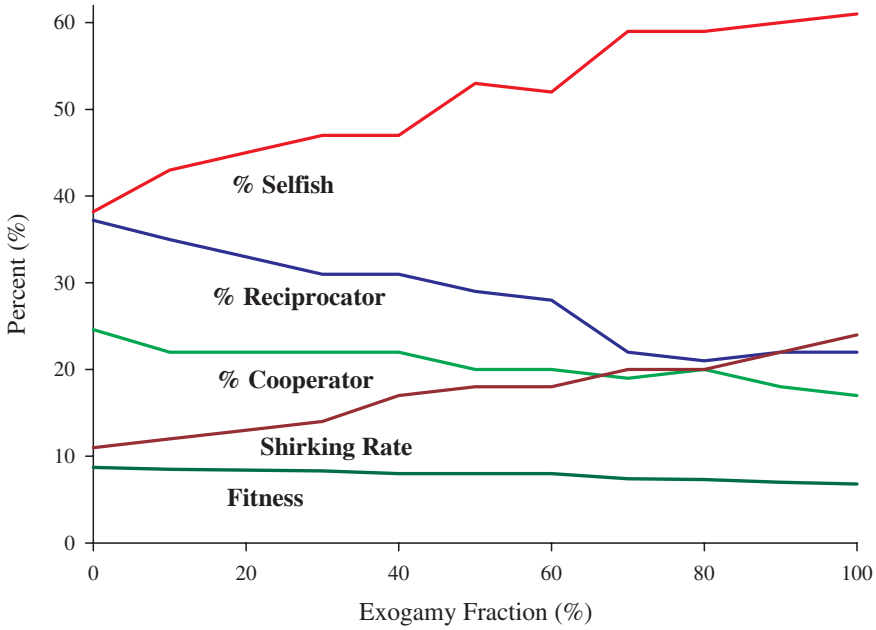

Figure 7: Degree of Exogamy and Degree of Cooperation. The baseline parameters are as in Table 1.

operative groups grows in size, and as a result it seeds other groups by migration and repopulates the sites of disbanded groups. The rate at which this process takes place obviously depends on the number of groups in the population. In the fourth panel of Figure 6 we present data in a population with 50 groups. A high level of cooperation is achieved (on the average) after about 1500 periods, for mutation rates as low as $\epsilon = 0.002$. Simulations with just 20 groups, by contrast, show that for mutation rates lower than 0.002, it takes on the average 30,000 periods before the shirking rate first falls below ten percent. We think that our simulations with 50 groups probably overstate the obstacles to the emergence of strong reciprocity for the simple reason that in order to proliferate the behavior need only become common in a single group, and there were far more than 50 bands of early foraging humans.

Since offspring of group members remain in the group until they either migrate, die, or are ostracized, our model is also likely to have some kin selection involved

in both strengthening the altruism of Reciprocators and damping the free-riding of Cooperators. In a haploid model, the relatedness of two agents is either unity if they are in the same clonal group, and zero otherwise. Thus, the within-group relatedness of a single group member is equal to the number of other agents in the group who are in the same clonal family, divided by group size minus one. Average within-group relatedness is the mean of this statistic over all agents in groups. In our baseline simulation, we find the degree of relatedness of group member to be about 10%, varying from 7% to 14% at different times within a single simulation. These figures move in the expected direction when migration and mutation rates are varied.

To assess quantitatively the contribution of kin selection to the maintenance of cooperation in the model, we allowed the fraction of offspring whose first period of life is within the parental group to vary from 0% to 100%, the remaining offspring being randomly distributed among all groups. A major reason for offspring to migrate to other groups is exogamy, which in small groups may be necessary to avoid inbreeding. So we define the *degree of exogamy* to be the fraction x of offspring who leave the group for random reassignment to other groups, with the remaining fraction $1 - x$ staying in the parental group unless reassigned through ostracism or migration (those leaving for 'exogamy reasons' were simply added to the migration flows mentioned above).

Figure 7 shows that the propensity of offspring to remain in the parental group is an important factor in promoting cooperation and reciprocity. When all offspring stay in the parental group, the average shirking level, using the same parameters as in Table 1, is approximately 11%, whereas when all offspring are randomly located to groups, the average shirking rate rises to 24%. In a plausible intermediate case, that of 50% exogamy, the average shirking rate is about 18%. This variation occurs, as is clear from the figure, because the long-run fraction of Reciprocators depends upon the degree of exogamy: with zero exogamy, this rate is about 37%, whereas with 100% exogamy, there are only 22% Reciprocators in the long run. However, while showing that within-group relatedness has the predicted effects on the population frequencies, Figure 7 also makes it clear that within group relatedness is not necessary to sustain high levels of cooperation in groups.

5 Discussion

Explanations of the evolution of cooperation among unrelated humans sometimes fail to explain why similar behaviors are seldom observed in other animals. Our model, however, relies on cognitive, linguistic, and other capacities unique to our species. The *moralistic aggression* (Robert Trivers' apt term) that arguably pro-

vides the motivational underpinnings of altruistic punishment requires uniquely human cognitive and linguistic abilities in the formulation of behavioral norms, the achievement of group consensus that the norms ought to be followed, the communication of their violations, and the coordination of the often collective nature of the punishment of miscreants. Additionally, uniquely human capacities to inflict punishment at a distance, through projectile weapons, reduce the cost of ostracizing a norm violator.

Our proposed explanation of human cooperation contrasts with the more standard interpretation stressing *reciprocal altruism* (Trivers 1971, Axelrod and Hamilton 1981). The canonical status of this view notwithstanding, there is little evidence that cooperation in prisoners' dilemma type situations among non-human animals is explained by the opportunities for inflicting costs on non-cooperators offered by repeated interactions (Stephens, McLinn and Stevens 2002). Thus, if our speculation that strong reciprocity emerged through a modification of reciprocal altruist behaviors is correct, this provides another reason why strong reciprocity might be uniquely human, given that reciprocal altruism appears to be very rare in other species.

Among humans however, we do not doubt the importance of repeated interactions and other structures that reward cooperators with higher fitness or other payoffs, rendering seemingly selfish acts a form of mutualism. While an important part of the explanation of human cooperation, there are several reasons for doubting the adequacy of this explanation. First, reciprocal altruism fails when a social group is threatened with dissolution, since members who sacrifice now on behalf of group members do not have a high probability of being repaid in the future (Gintis 2000).

Second, many human interactions in the relevant evolutionary context took the form of n-person public goods games—food sharing and other co-insurance, as well as common defense—rather than dyadic interactions. Even if repeated with high probability, n-person public goods (or common pool resource) interactions make cooperation difficult to sustain by means of the standard tit-for-tat and other reciprocal behaviors, as suggested by Joshi (1987), Boyd and Richerson (1988), and Choi (2003). Alexander (1987) has proposed a more general "indirect reciprocity" mechanism more amenable to large group interactions, and this has been formalized by Nowak and Sigmund (1998). According to this model, agents reward and punish other agents who have defected in pairwise interaction with third parties. While indirect reciprocity may well be an important source of cooperation in humans, for reasons given by Leimar and Hammerstein (2001), we doubt that indirect reciprocity can be sustained in a population of self-interested agents. Indirect reciprocity is more likely promoted, as in our model, by strong reciprocators who reward prosocial behavior and punish antisocial behavior even when this behavior reduces within-group fitness. For an empirical study supporting this idea, see dn Urs Fischbacher

(2002).

Third, the contemporary study of human behavior has documented a large class of prosocial behaviors inexplicable in terms of reciprocal altruism. For instance, there is extensive support for income redistribution in advanced industrial economies, even among those who cannot expect to be net beneficiaries (Fong, Bowles and Gintis 2002). Group incentives for large work teams are often effective motivators even when the opportunity for reciprocation is absent and the benefits of cooperation are so widely shared that a self-interested group member would gain from free-riding on the effort of others (Ghemawat 1995, Hansen 1997, Knez and Simester in press). Finally, laboratory and field experiments show that non-selfish motives are frequently robust predictors of behavior, even in one-shot, anonymous setting. This research has been summarized in Ostrom (1998) and Fehr and Gächter (2000) for industrial societies, and Henrich, Boyd, Bowles, Camerer, Fehr, Gintis, and McElreath (2001) hunter-gatherer and other small scale societies.

Our model differs from other explanations of cooperation among unrelated individuals in several ways. Most models of reciprocity treat interactions among pairs of agents (Boorman and Levitt 1980, Axelrod and Hamilton 1981, Kreps, Milgrom, Roberts and Wilson 1982, Axelrod 1984). Since strong reciprocity is exhibited in such collective situations as group food-sharing and defense, these models not suited to explaining this phenomenon. By contrast, we model n-agent groups (where n is on the order of ten to 100) in a series of production periods that are effectively one-shot, since the only inter-period influences are those involving the biological and cultural reproduction of new agents. Moreover, in contrast to other models of cooperation in groups, (Robson 1990, Nowak, Page and Sigmund 2000, Wedekind and Milinski 2000), we assume Reciprocators cannot gain from being phenotypically identified as such, or by establishing a reputation for reciprocation across production periods.[4]

Nor in our model can Reciprocators use their altruistic behavior as a costly signal of superior fitness (Zahavi 1975, Bliege Bird, Smith and Bird 2001, Gintis et al. 2001). Finally, while most models of strong reciprocity depend on group extinctions (Gintis 2000, Boyd et al. 2003, Bowles, Choi and Hopfensitz 2003), ours does not.

Our approach is related to the model of Aviles, Abbot and Cutter (2002), as applied to tree-killing bark beetles (Raffa and Berryman 1987) and other species in

[4]Even though our stage game is a one-shot, it may in fact involve behaviors that take place over several, or even many years (e.g., a hunting season, of which one period in our model may comprise several). Our treatment of being "detected shirking" is compatible with individuals' building reputations during the course of the game that at some point trigger punishment. Our "probability of being detected shirking" is a summary description of this process, which may occur over time even in a one-shot game.

which there are strong fitness benefits associated with social living. They posit a minimum group size and positive fitness effects of group size for smaller groups. While these conventional Allee effects play no role in our model, they are approximated by our assumption living in cooperative groups confers fitness benefits and that solitary individuals bear fitness costs. Like Trivers (1971), Hirshleifer and Rasmusen (1989), Boyd and Richerson (1992) Sethi and Somanathan (1996), and Friedman and Singh (2001), we stress the importance of altruistic punishment. Hirshleifer and Rasmusen (1989) and Friedman and Singh (2001) develop models of team production in which the threat of ostracism deters shirking. But, because they assume that ostracizing is not costly to the individual ostracizer, their models, unlike ours, do not explain the persistence of altruistic behaviors. Our paper is most closely related to Sethi and Somanathan (1996). But, in Sethi and Somanathan's model the equilibrium frequency of Reciprocators is zero and cooperation is complete. By contrast, our model supports a positive (indeed, quite high) fraction of Reciprocators and a significant level of non-cooperation in the long run.

We think that our model, suitably extended, can capture the environments that may have supported high levels of cooperation among our ancestors living in mobile foraging bands during the late Pleistocene. We do not know that a human predisposition to strong reciprocity evolved as we have described. But our simulations suggest that it could have.

6 Appendix: Simulation Details

Initialization specifies the following parameters (baseline values in parenthesis): initial number of agents per group (20), number of groups (20), initial fraction ρ_r of population who are Reciprocators (0), initial fraction ρ_c of population who are Cooperators (0), cost c of cooperating (.1), cost c_p of punishing (.1), gain b from cooperation (.2), fraction γ of members of groups seeking to emigrate to another group (.05), fraction μ of groups size allowed as immigrants(.03), fitness in pool ϕ_0 (-.01), mutation rate (.01), fraction ρ_c^0 of Reciprocators in seeded groups (.8), minimum group size (6), total number of rounds (30000), range of uniform distribution for subjective cost s of being ostracized ([0,1]).

The initial agents are then created, and assigned a random value of s. These agents are then randomly assigned to groups, with a fraction ρ_r being Reciprocators and a fraction ρ_c begin Cooperators.

A period of play consists of (a) the frequency of each type of agent in each group determines the shirking choice of each Selfish agent, and from this the average shirking level for each group; (b) within-group fitness payoffs are assigned, using (3); (c) agents reproduce and die according to these fitnesses; offspring are copies of

their single parent with probability 1-e, and are random mutants with probability e (a mutant has equal probability of being either type different from his parent, and has a random draw from the uniform distribution of subjective cost of being ostracized; (d) shirkers are ostracized with the appropriate probabilities; (e) immigration to groups from the pool and other groups takes place; a maximum number of immigrants equal to a fraction μ of current group members is determined, and set of potential emigrants, consisting of all pool members, plus a fraction γ of group members, is specified; if this set of candidates for immigration is smaller than the maximum number of immigrants, all are randomly assigned to groups in proportion to their size; more likely, this set will be considerably larger than the maximum number of immigrants, in which case a randomly chosen subset is specified, and this subset is similarly distributed among groups; (f) if the total population has grown, agents are randomly killed to restore the simulation total population size; given the baseline parameters, this usually involved killing 5% to 8% of the population per period; (g) groups that are less than minimum group size are disbanded, and their members are sent to the pool; the site is repopulated by migration of members randomly drawn from the largest group, until the source group reaches initial group size, and so on for all groups larger than initial group size; if the site is still not fully population, agents are randomly assigned to the site from the pool.

References

Alexander, R. D., *The Biology of Moral Systems* (New York: Aldine, 1987).

Andreoni, James and John H. Miller, "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica* 70,2 (2002):737–753.

Aviles, Leticia, Patrick Abbot, and Asher D. Cutter, "Propulation Ecology, Nonlinear Dynamics, and Social Evolution. I. Associations among Nonrelatives," *The American Naturalist* 159,2 (February 2002):115–128.

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

— and William D. Hamilton, "The Evolution of Cooperation," *Science* 211 (1981):1390–1396.

Binford, Lewis, *Constructing Frames of Reference: An Analytical Method for Archeological Theory Using Hunter-Gatherer and Environmental Data Sets* (Berkeley, CA: University of California Press, 2001).

Bliege Bird, Rebecca L., Eric A. Smith, and Douglas W. Bird, "The Hunting Handicap: Costly Signaling in Human Foraging Strategies," *Behavioral Ecology and Sociobiology* 50 (2001):9–19.

Boehm, Christopher, *Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies* (Philadelphia, PA: University of Pennsylvania Press, 1984).

— , "Egalitarian Behavior and Reverse Dominance Hierarchy," *Current Anthropology* 34,3 (June 1993):227–254.

— , "Variance Reduction and the Evolution of Social Control," 2002. Department of Anthropology, University of Southern California.

Boorman, Scott A. and Paul Levitt, *The Genetics of Altruism* (New York: Academic Press, 1980).

Bowles, Samuel, Jung-kyoo Choi, and Astrid Hopfensitz, "The Co-evolution of Individual Behaviors and Social Institutions," *Journal of Theoretical Biology* 223 (2003):135–147.

Boyd, Robert and Peter J. Richerson, "The Evolution of Reciprocity in Sizable Groups," *Journal of Theoretical Biology* 132 (1988):337–356.

— and — , "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups," *Ethology and Sociobiology* 113 (1992):171–195.

— , Herbert Gintis, Samuel Bowles, and Peter J. Richerson, "Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences* 100,6 (March 2003):3531–3535.

Choi, Jung-Kyoo, "Is Repetition Sufficient to Support Retaliation?," 2003. Santa Fe Institute.

dn Urs Fischbacher, Dirk Engelman, "Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game," 2002. Institute for Empirical Research in Economics Working Paper.

Fehr, Ernst and Simon Gächter, "Cooperation and Punishment," *American Economic Review* 90,4 (September 2000):980–994.

— and — , "Altruistic Punishment in Humans," *Nature* 415 (10 January 2002):137–140.

Fong, Christina M., Samuel Bowles, and Herbert Gintis, "Reciprocity and the Welfare State," in Jean Mercier-Ythier, Serge Kolm, and Louis-André Gérard-Varet (eds.) *Handbook on the Economics of Giving, Reciprocity and Altruism* (Amsterdam: Elsevier, 2002).

Friedman, Daniel and Nirvikar Singh, "Evolution and Negative Reciprocity," in Y. Aruka (ed.) *Evolutionary Controversies in Economics* (New York: Springer, 2001).

Ghemawat, Pankaj, "Competitive Advantage and Internal Organization: Nucor Revisited," *Journal of Economic and Management Strategy* 3,4 (winter 1995):685–717.

Gintis, Herbert, "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology* 206 (2000):169–179.

— , Eric Alden Smith, and Samuel Bowles, "Costly Signaling and Cooperation," *Journal of Theoretical Biology* 213 (2001):103–119.

Hansen, Daniel G., "Individual Responses to a Group Incentive," *Industrial and Labor Relations Review* 51,1 (October 1997):37–49.

Henrich, Joe, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, *Foundations of Human Sociality: Ethnography and Experiments in Fifteen Small-scale Societies* (Oxford: Oxford University Press, 2003).

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath, "Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies," *American Economic Review* 91 (May 2001):73–78.

Hirshleifer, David and Eric Rasmusen, "Cooperation in a Repeated Prisoners' Dilemma with Ostracism," *Journal of Economic Behavior and Organization* 12 (1989):87–106.

Joshi, N. V., "Evolution of Cooperation by Reciprocation Within Structured Demes," *Journal of Genetics* 66,1 (1987):69–84.

Kelly, Robert L., *The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways* (Washington, DC: The Smithsonian Institution, 1995).

Knauft, Bruce, "Violence and Sociality in Human Evolution," *Current Anthropology* 32,4 (August–October 1991):391–428.

Knez, Marc and Duncan Simester, "Firm-Wide Incentives and Mutual Monitoring at Continental Airlines," *Journal of Labor Economics (*in press).

Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory* 27 (1982):245–252.

Leimar, O. and P. Hammerstein, "Evolution of Cooperation through Indirect Reciprocity," *Proc. Royal. Soc. Lond. B* 268 (2001):745–753.

Loewenstein, George F., Leigh Thompson, and Max H. Bazerman, "Social Utility and Decision Making in Interpersonal Contexts," *Journal of Personality and Social Psychology* 57,3 (1989):426–441.

Moore, Jr., Barrington, *Injustice: The Social Bases of Obedience and Revolt* (White Plains: M. E. Sharpe, 1978).

Nowak, Martin A. and Karl Sigmund, "Evolution of Indirect Reciprocity by Image Scoring," *Nature* 393 (1998):573–577.

— , Karen M. Page, and Karl Sigmund, "Fairness Versus Reason in the Ultimatum Game," *Science* 289 (8 September 2000):1773–1775.

Ostrom, Elinor, "A Behavioral Approach to the Rational Choice Theory of Collective Action," *American Political Science Review* 92,1 (March 1998):1–21.

— , James Walker, and Roy Gardner, "Covenants with and without a Sword: Self-Governance Is Possible," *American Political Science Review* 86,2 (June 1992):404–417.

Price, G. R., "Selection and Covariance," *Nature* 227 (1970):520–521.

Raffa, K. F. and A. A. Berryman, "Interacting Selective Pressures in Conifer-bark Beetle Systems: a Basis for Reciprocal Adaptations?," *The American Naturalist* 129 (1987):234–262.

Robson, Arthur J., "Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake," *Journal of Theoretical Biology* 144 (1990):379–396.

Scott, James C., *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia* (New Haven, CT: Yale University Press, 1976).

Sethi, Rajiv and E. Somanathan, "The Evolution of Social Norms in Common Property Resource Use," *American Economic Review* 86,4 (September 1996):766–788.

Stephens, W., C. M. McLinn, and J. R. Stevens, "Discounting and Reciprocity in an Iterated Prisoner's Dilemma," *Science* 298 (13 December 2002):2216–2218.

Trivers, R. L., "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971):35–57.

Wedekind, Claus and Manfred Milinski, "Cooperation Through Image Scoring in Humans," *Science* 289 (5 May 2000):850–852.

Wood, Elisabeth Jean, *Insurgent Collective Action and Civil War in El Salvador* (Cambridge,: Cambridge University Press, 2003).

Woodburn, James, "Egalitarian Societies," *Man* 17,3 (1982):431–451.

Yamagishi, Toshio, "The Provision of a Sanctioning System as a Public Good," *Journal of Personality and Social Psychology* 51 (1986):110–116.

Zahavi, Amotz, "Mate Selection—A Selection for Handicap," *Journal of Theoretical Biology* 53 (1975):205–214.