

The Evolution of Reciprocal Preferences*

Samuel Bowles
Herbert Gintis
Department of Economics
University of Massachusetts, Amherst

November 19, 2000

Abstract

A number of outstanding puzzles in economics may be resolved by recognizing that where members of a group benefit from mutual adherence to a social norm, agents may obey the norm and punish its violators, even when this behavior cannot be motivated by self-regarding, outcome-oriented preferences. This behavior, which we call *strong reciprocity*, is a form of altruism in that it benefits others at the expense of the individual exhibiting it. Thus where benefits and costs are measured in fitness terms and where the relevant behaviors are governed by genetic inheritance subject to natural selection, it is generally thought that, as a form of altruism, strong reciprocity cannot invade a population of self-interested types, nor can it be sustained in a stable population equilibrium. We show that strong reciprocity can invade a population of self-interested types and can be sustained in a stable population equilibrium, and we show that, under assumptions that appear to reflect the relevant social-environmental conditions, the model can account for the genetic evolution of strong reciprocity.

1 Introduction

While the assumption of self-interested action has proven a remarkably powerful behavioral foundation for the understanding of economic behavior, a number of important economic phenomena are difficult to explain on this basis. Among these are the importance of fairness motives in wage setting and other exchanges involving

*We would like to thank Christopher Boehm, Robert Boyd, Leda Cosmides, Steven Frank, Kristin Hawkes, Hillard Kaplan, Peter Richerson, Rajiv Sethi, Eric Alden Smith, E. Somanathan, Leigh Tesfatsion, and Peyton Young for their help with this paper, and the MacArthur Foundation for financial support. The authors can be contacted at bowles@econs.umass.edu, <http://www-unix.oit.umass.edu/~bowles> and hgintis@mediaone.net, <http://www-unix.oit.umass.edu/~gintis>.

strategic interaction (Blinder and Choi 1990, Bewley 2000), the extensive support for welfare programs even among those who cannot expect to be net beneficiaries (Gilens 1996, Luttmer 1998, Gilens 1999, Fong 2000, Piketty 1999) and the effectiveness, in some cases, of group incentives even where residual claimancy is shared among so large a number that the individual gain associated with one's own effort is small (Ghemawat 1995, Hansen 1997, Knez and Simester 1998). Experiments by behavioral scientists have provided further evidence that in some situations non-selfish motives are robust predictors of behavior (Fehr and Gächter 2000, Fehr and Falk 1999, Kahneman, Knetsch and Thaler 1986, Güth and Tietz 1990, Hoffman, McCabe and Smith 1998, Isaac, Walker and Williams 1994). This evidence provides sufficient reason to consider a broader range of human motivations.

We hypothesize that where members of a group benefit from mutual adherence to a social norm, individuals may obey the norm and punish its violators, even when this behavior cannot be justified in terms of selfish preferences. We call this *strong reciprocity*. We distinguish this from contingent cooperation in a repeated game, reciprocal altruism, and other forms of mutually beneficial cooperation that can be accounted for in terms of self-interest. Compelling evidence for the existence and importance of strong reciprocity comes from controlled laboratory experiments, particularly the study of public goods, common pool resource, trust, ultimatum, and other games (Fehr and Gächter 2000), from the ethnographic studies on simple societies (Knauff 1991, Boehm 1984, Boehm 1993), from historical accounts of collective action (Moore 1978, Scott 1976), as well as from everyday observation.

Strong reciprocity confers group benefits by promoting cooperation and punishing free riding. However such behavior imposes individual costs, both because strong reciprocators contribute more to the group than selfish types, and because they sustain the costs of punishing free riders. Thus where benefits and costs are measured in fitness terms and where the relevant behaviors are governed by genetic inheritance subject to natural selection, it is generally thought that, as a form of altruism, strong reciprocity cannot invade a population of self-interested types, nor can it be sustained in a stable population equilibrium. We show that this is not the case, and offer an evolutionary explanation of the phenomenon.

We do not address the empirical question concerning the degree to which observed strongly reciprocal behavior is genetically as opposed to culturally based. Rather, we answer the question: could such behavior have a genetic basis—beyond the obvious requirements on the cognitive capacities of individuals. As the late Pleistocene is the only period long enough to account for a significant development in modern human gene distributions, we base our model on the structure of interaction among members of the small hunter-gatherer bands in this period, which constitutes most of the history of *Homo sapiens*, as revealed by historical

and anthropological evidence.¹

Here we propose an explanation based on the fact that strong reciprocity supports high levels of mutual monitoring within groups, and for this reason groups with large numbers of reciprocators have superior average levels of fitness. Despite the individually costly nature of monitoring and punishing, strong reciprocity can then evolve because of the greater likelihood that reciprocators will be in groups with effective mutual monitoring, while self-interested types will occasionally be ostracized from groups, incurring a fitness cost.

Our model has several characteristics similar to other accounts of reciprocity. The behaviors we seek to explain, while formally altruistic—that is individually costly and group beneficial—are more punishing than kind, a characteristic shared by Trivers (1971), Hirshleifer and Rasmusen (1989b), Boyd and Richerson (1992), Sethi and Somanathan (1996) and Friedman and Singh (1999).² Like Binmore (1998) we use evidence on the evolution of humans in foraging bands to study the influence of reciprocity concerns on the nature of equilibria in public goods games, but unlike Binmore we explore the evolution of non-self-regarding preferences in these environments. And like Güth and Yaari (1992), Huck and Oechssler (1996), Bester and Güth (1998) and Friedman and Singh (1999), we distinguish between utility, which affects behavior, and fitness, which affects rates of reproduction.

Our approach is also distinctive in two respects. First, while most models of reciprocity use repeated interactions among pairs of agents to induce cooperative behavior (Boorman and Levitt 1980, Axelrod and Hamilton 1981, Kreps, Milgrom, Roberts and Wilson 1982, Axelrod 1984, Boyd and Lorberbaum 1987, Guttman 1996), we treat social interaction as a series of one-time events in which no new knowledge is acquired from the events of the previous periods, and we assume that relatively large groups of agents interact.

Second, our model is based on group membership rather than genetic relatedness, as in Samuelson (1983), Bergstrom and Stark (1993), and Bergstrom (1995). However, unlike other models of this type (Robson 1990, Güth and Yaari 1992, Güth 1995) we do not assume reciprocators can be distinguished from self-interested types by some phenotypic trait, nor can individuals establish reputations by their behaviors. Rather, in our model reciprocators are more likely to be in groups with other reciprocators because they avoid being ostracized for misbehavior.

If our model is to account for the evolution of strong reciprocity in humans it

¹As the mechanics of genetic determination and its associated inheritance process are not germane to our model, we leave this issue unexplored, assuming that offspring are clones of a single parent.

²Sethi and Somanathan's paper is most closely related with our work, but our reciprocators do not use weakly dominated strategies, so our model can support a positive (indeed, quite high) level of non-cooperation in equilibrium. We believe a high frequency of non-cooperation is in fact found in both simple and contemporary societies.

should capture the social and physical environment of the foraging bands that made up most of human society for most of its history. While modern accounts of these societies record considerable variety in social organization and livelihood (Kelly 1995), the widespread sharing of food, valuable information, and other sources of survival among many of these societies in the modern world is well established (Woodburn 1982, Boehm 2000). Strong reciprocity, including spontaneous sharing and the sanctioning of those who violate sharing norms, provides a parsimonious explanation. Punishing norm violators deters free riding and hence explains both sharing and working to acquire goods that later would be shared.

The evolutionary puzzle is not why group members work and share, but rather why they punish. To address this problem, we develop a team production model in which it is costly both to follow a work norm and to punish norm violators. Our model captures key characteristics of small foraging bands.³ First, groups are sufficiently small that members directly observe and interact with one another, yet sufficiently large that the problem of free riding in team production 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 voluntary 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 horticultural and later industrial societies, which justifies our treatment of individuals as homogeneous other than by reciprocator/self-interested type and by the group to which they belong. Fifth, the sharing on which our model of team production is based—either of food individually acquired or of the common work of acquiring food—is characteristic of these societies. Sixth, hunter-gather bands experience high membership turnover, justifying our abstraction from reputation effects and repeated interactions as means of norm enforcement (Rogers 1990). Sixth, the only intertemporal relationships in our model concern fitness: the individuals in our model do not invest—store food or accumulate resources—and this too is a characteristic of at least those hunter-gather bands based on what Woodburn (1982) calls an “immediate return” system of production. Finally, 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 demographic parameters of the model.⁴ This manner of treating punishment reflects a central

³We have relied on the following sources: Balikci (1970), Chagnon (1977), Lee (1979), Cashdan (1980), Woodburn (1982), Boehm (1982), Kaplan, Hill, Hawkes and Hurtado (1984), Kaplan and Hill (1985b), Kaplan and Hill (1985a), Blurton Jones (1987), Woodburn and Barnard (1988), Endicott (1988), Kent (1989), Knauff (1989), Knauff (1991), Hawkes (1992), Boehm (1993), Hawkes (1993), Damas (1972) Kelly (1995).

⁴Hirshleifer and Rasmusen (1989a) develop an interesting model of team production in which the threat of ostracism deters shirking. Because they assume that ostracizing is costly to the group but

aspect of hunter-gatherer life: since individuals can always leave the group to avoid punishment, the cost of being ostracized is the largest penalty that can be levied upon an individual group member.⁵

2 Equilibrium Working, Shirking and Punishing Within a Group

Consider a population in which agents can work alone, in which case each has baseline fitness $\phi_0 < 0$.⁶ Agents can also work cooperatively in a group, each producing an amount q at cost b (all benefits and costs are in fitness units). We assume output is shared equally by the agents, so if all group members work, each has a net group fitness benefit $q - b > 0$.

Given the equal sharing rule each agent may benefit from shirking. To model this, we suppose agent j shirks a fraction σ_j of the time, so $\sigma = \sum_{j=1}^n \sigma_j/n$ is the average rate of shirking, where n is the size of the group. The fitness value of group output is $n(1 - \sigma)q$, and since output is shared equally, each member receives $(1 - \sigma)q$. The loss to the group from a member j shirking is $q\sigma_j$, while the gain to the shirking member is the fitness cost of effort, $b(\sigma_j)$ with $b(0) = b$, $b(1) = 0$, $b'(\sigma_j) < 0$, and $b''(\sigma_j) > 0$, the same for all group members, and $q(1 - \sigma_j) > b(\sigma_j)$ for $\sigma_j \in (0, 1]$, so at every level of effort, working helps the group more than it hurts the worker.

We assume that group size n is sufficiently large that $q(1 - \sigma_j)/n < b(\sigma_j)$ for $\sigma_j \in (0, 1]$, so if there is no policing of free riders, complete shirking ($\sigma_j = 1$) would promote a member's fitness, whether or not the other members work or shirk. However we suppose that a group member can be monitored by other members of the group, and if detected shirking, can be punished. Suppose the fitness cost to a member of punishing a shirker is $c > 0$. We also assume a member shirking at rate σ_j will be punished with probability $f\sigma_j$, where f is the fraction of agents in the group who are reciprocators. Punishment consists of a shirker being ostracized from the group. We assume an ostracised agent works alone for a period of time before being readmitted to another group. We summarize the fitness cost of being ostracized as $s > 0$, an endogenous variable that will be determined jointly with the distribution of types among the solitary individuals and those in groups.

The group now faces a 'second order free rider problem': it is costly to monitor,

not to the individual their model, unlike ours, does not explain the persistence of altruistic behaviors.

⁵There being no property, individuals cannot have 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 always reserved for such serious crimes as adultery and murder.

⁶By the *fitness* of an agent we mean the number of replicas produced by the agent in one period or, equivalently, the rate of growth of the agent's offspring.

so each member would like the others to monitor, but each suffers fitness losses by doing so himself. Suppose, however, the group consists of two type of actors. The first type, whom we call *reciprocators*, work and punish shirkers with probability one. The second type, whom we call *self-interested*, maximize fitness, and therefore never punish, and work only to the extent that the expected fitness cost of working is less than the expected fitness cost of being punished.

How might the behaviors associated with reciprocators, who do not maximize fitness, have evolved under the influence of natural selection operating on genetically transmitted traits? To answer this we explore whether individuals with these preferences might persist in population equilibrium.

We assume the fraction f of reciprocators in the group is common knowledge, but on the individual level a self-interested type cannot be distinguished from a reciprocator. Moreover, since shirkers are ostracized, members do not accumulate information concerning other members' behavior in previous periods, so we assume that all group members are monitored with equal probability.⁷ Writing the cost of working $\hat{b}(\sigma_j)$ for member j as the cost of effort plus the expected cost of being ostracized, we have

$$\hat{b}(\sigma_j) = b(\sigma_j) + sf\sigma_j. \quad (1)$$

Technically, we should subtract from this cost the agent's share $q(1 - \sigma_j)/n$ of his own production, where n is the size of the group. But to simplify the model, we assume the group size is sufficiently large that this quantity can be safely ignored. Then if the shirking rate σ_n of self-interested types is fitness-maximizing, $\sigma_n(f)$ minimizes the cost of working, the first order condition for which is

$$\hat{b}'(\sigma_j) = b'(\sigma_j) + fs = 0. \quad (2)$$

Since $b'' > 0$, there is at most one solution to this equation, and it is a minimum. If we set

$$f_o = \frac{s}{b'(0)}, \quad (3)$$

then clearly $\sigma_n(f) = 1$ on the interval $[0, f_o]$. We assume that $s + b'(0) < 0$, which implies that (2) has an interior solution for $f \in (f_o, 1]$. In particular, there

⁷Formally, the fact that a group member had previously been ostracized from another group implies that the member is self-interested. However, other sources of migration among groups are likely to render the fraction of self-interested types among migrants virtually indistinguishable from the same fraction within groups. Most important among these, according to anthropological reports of contemporary hunter-gatherers, are (a) a certain fraction of groups disband spontaneously in the face of disease, drought, and other natural events, and (b) tensions among group members unrelated to production (e.g., concerning adultery and status feuding) lead members to seek new group membership. Thus most individuals and families will be migrants at some time in their lives. For these reasons, we will assume that having migrated from a different group supplies no information concerning whether an agent is a reciprocator or is self-interested.

is a minimum shirking rate $\sigma_{\min} > 0$ even when there is full monitoring ($f = 1$), given by $\sigma_{\min} = \sigma_n(1)$.

Since reciprocators never shirk, the average level of shirking is given by $\sigma(f) = (1 - f)\sigma_n(f)$. The expected contribution of each group member to the group's population in the next period is equal to its fitness minus the rate of ostracism from shirking, which is

$$\pi_n = q(1 - \sigma) - b(\sigma_n(f)) - f\sigma_n(f), \quad (4)$$

$$\pi_r = q(1 - \sigma) - b - c(1 - f)\sigma_n(f)/f. \quad (5)$$

The final term in the last equation reflects the collective nature of punishment of norm violators (Boehm 2000). We derived this as follows. The cost of ostracizing a shirker is c . The number of shirkers is $(1 - f)n\sigma_n(f)$, where group size is n . This cost is shared equally by the fn reciprocators, so the fitness cost per reciprocator of punishing is $c(1 - f)\sigma_n(f)/f$.

To derive the condition for f to be stationary, we assume that agents first reproduce, then immigration from the pool of solitary producers occurs at rate μ , followed by emigration at rate γ for reasons other than ostracism. Then for a group of size n , the number of reciprocators at the end of the period is $n(f + \pi_r f + \mu g)(1 - \gamma)$, where g is the fraction of those in the pool who are reciprocators. The size of the group at the end of the period is $n(1 + \pi_r f + \pi_n(1 - f) + \mu)(1 - \gamma)$. The quotient of these two expressions must equal f in equilibrium, which is equivalent to the condition

$$\pi_r - \pi_n = \frac{f - g}{f(1 - f)}\mu. \quad (6)$$

To determine the equilibrium value g^* of g , note that the fraction of a group leaving to join the pool is $\gamma + (1 - f)f\sigma_n(f)$, of which γf are reciprocators. Thus the fraction g of those leaving groups who are reciprocators is given by

$$\bar{g} = \frac{\gamma f}{\gamma + (1 - f)f\sigma_n(f)}. \quad (7)$$

If reciprocators and self-interested types are equally likely to be readmitted to groups, then in equilibrium the fraction g^* of those in groups who are reciprocators must be

$$g^* = \frac{\gamma f^*}{\gamma + (1 - f^*)f^*\sigma_n(f^*)}. \quad (8)$$

Thus the equilibrium condition (6) for f can be written

$$b(\sigma_n(f^*)) + f^*\sigma_n(f^*) \left(1 - \frac{\mu}{\gamma + (1 - f^*)f^*\sigma_n(f^*)} - \frac{(1 - f)c}{f^2} \right) = b.$$

When $f = 0$ the left hand side of this is $-\infty$, and when $f = 1$, the left hand side is $b(\sigma_{\min}) + \sigma_{\min}(1 - \mu/\gamma)$, which we assume is greater than b . This implies the existence of an equilibrium. We will analyze the existence and stability of equilibrium for the model as a whole later in the paper.

In equilibrium, self-interested types within groups have fitness

$$\phi_n(f^*) = q(f^*)(1 - \sigma(f^*)) - b(\sigma_n(f^*)). \quad (9)$$

Suppose an ostracized agent must wait a time t_o before being readmitted to a group. In equilibrium the agent's fitness loss from being ostracized is then $s = t_o(\phi_n(f^*) - \phi_o)$ as the reproduction foregone in a single period of ostracism is $\phi_n(f^*)$ and fitness working alone is ϕ_o . To close the model we must determine the expected amount of time t_o , which depends on the relative size of the population in groups and working independently.

Since groups admit a fraction μ of new members in each period, if p is the fraction of agents in groups, $\mu p/(1 - p)$ is the fraction of agents in the pool who are readmitted to groups in each period. Assuming all agents working independently have an equal probability of being admitted into a group, the expected time working independently is $t_o = (1 - p)/\mu p$. Thus the equilibrium condition for s is

$$s^* = \frac{1 - p^*}{\mu p^*}(\phi_n(f^*) - \phi_o), \quad (10)$$

where p^* is the equilibrium fraction of the population in cooperative groups.

To determine the fraction p of agents in groups, we assume a fraction $\gamma > 0$ of agents leave groups for reasons other than being ostracized. Then the growth rate of the pool of independent agents is given by

$$\lambda_n = \frac{p[(1 - f^*)f^*\sigma_n(f^*) + \gamma - \mu]}{1 - p} + \phi_o. \quad (11)$$

To see this, note that the numerator of the fraction is the fraction of the population that is ostracized plus the fraction that leaves groups for other reasons, minus the fraction that migrate back into groups, while the denominator is the fraction of agents working independently. Similarly, the growth rate of groups is given by

$$\lambda_g = \pi_r(f^*)f^* + \pi_n(f^*)(1 - f^*) + \mu - \gamma. \quad (12)$$

Since these two rates of growth must be equal in equilibrium, the equilibrium condition for the fraction of agents in groups is given by

$$\lambda_g = \lambda_n. \quad (13)$$

This completes the model, for the stationary value of p^* given by (12) then determines the stationary values of s^* , g^* , f^* and σ^* given by (2), (6), (8), and (10), so the relative size of the solitary pool and the groups is constant, as is the distribution of reciprocators and nonreciprocators in both the pool and the groups, as is the endogenously given cost of ostracism, and as a consequence the level of shirking by nonreciprocators in groups.

It remains to be shown that reciprocators will proliferate when rare and that the population will sustain a positive fraction of reciprocators. To see that universal self-interest is not a stable equilibrium and can be invaded by a small fraction of reciprocators in the population under plausible conditions, suppose there is a fraction ϵ of reciprocators in a very large population N of self-interested types. In each period with some positive probability a random sample of $n = \delta N$ individuals is drawn from the population to form a group, where δ is a very small positive number. The fraction of the population in groups is then $p = n/N$, so the time an individual ostracized from the group would spend back in the pool would be

$$t_o = \frac{1-p}{p} = \frac{1}{\delta} - 1, \quad (14)$$

which is a very large number. We shall show that with positive probability there will be a within group Nash equilibrium in which no agent shirks, the within-group fitnesses of reciprocators are equal, and the group grows at a positive rate. This implies that a small number of reciprocators can invade a population of self-interested types.

Assuming $f > 0$ and there is no shirking ($\sigma_n = 0$) in the group, (4), (5) and (9) show that $\pi_n = \pi_r = \phi_n = q - b > 0$, so (10) and (14) give

$$s = \left(\frac{1}{\delta} - 1 \right) (q - b),$$

which is a very large number. This implies that $f > f_o$ in (3), as long as $f > 0$ in the group, which in turn implies $\sigma_n = 0$, thus justifying our original assumption. This proves that there is a no-shirking Nash equilibrium in the group with the characteristics required for the group to grow at a strictly positive rate and with a constant fraction of reciprocators. Since such a group forms in each period with strictly positive probability, we conclude that universal self-interest is not a stable equilibrium.

This result can be confirmed by noticing that for some small value of p , λ_n is negative so the population in the pool is shrinking, while by the above reasoning λ_r must be positive. Because groups are growing and reciprocators form a larger fraction of the groups than of the pool, their fraction in the population is also growing. Thus both p and ρ are bounded away from zero.

It is not obvious, however, that the set of equations describing the interior equilibrium fraction of reciprocators has a stable interior solution for plausible parameters, yielding plausible equilibrium values. In the next section we provide a simulation that illustrates the model and shows that such a stable interior solution may exist.

3 Simulation Strong Reciprocity

We specify the cost of working as a function of the shirking level by

$$b(\sigma) = \frac{1}{\sigma + a_o} + b_o, \quad (15)$$

where a_o and b_o are determined by the conditions $b(1) = 0$ and $b(0) = b$. We find $a_o = (\sqrt{b+4} - \sqrt{b})/2\sqrt{b}$, and $b_o = (b - \sqrt{b+4}\sqrt{b})/2$. We may now solve the first order condition (2), obtaining the self-interested type's level of shirking, which turns out to be

$$\sigma_n(f) = \frac{1}{\sqrt{fs}} - \frac{\sqrt{b+4} - \sqrt{b}}{2\sqrt{b}}. \quad (16)$$

This equation holds so long as $\sigma_n(f) \leq 1$. It is easy to check that on the interval $f \in [0, 4b/s(\sqrt{b+4} + \sqrt{b})^2]$ we have $\sigma_n(f) = 1$, and (16) holds for the remainder of the unit interval.

For our simulation, we set $b = c = 0.15$, $\gamma = 0.07$, and $\mu = 0.08$. We also assumed a baseline fitness of $\phi_o = -0.02$, and we adjust the productivity of effort q so that in equilibrium, population size is approximately constant. This involved setting $q = 0.19$. We find that in equilibrium, the fraction of those in groups who are reciprocators is $f^* = 70\%$, the shirking level of self-interested types is $\sigma_n(f^*) = 10\%$, the fitness cost of ostracism is $s^* = 0.29$, the fraction of all agents in groups is $p^* = 70\%$, and the average time in the solitary pool is $t_o = 5.38$ periods. The fitness of reciprocators is given by

$$p_r \pi_r + (1 - p_r) \phi_o = 0.00,$$

as we would expect, and the fitness of non-reciprocators is given by

$$p_n \phi_n + (1 - p_n) \phi_o = 0.01.$$

This should also be zero, but the equilibrium values of f , s , λ_n and λ_g are only accurate to two decimal places, so this result can be attributed to rounding error. Additional statistics concerning this simulation are presented in Figure 1.

I. Model Parameters

Variable	Value	Description
b	0.15	Cost of Working
c	0.15	Cost of Monitoring
q	0.19	Output/Worker with No Shirking
γ	0.07	Rate of Emigration from Groups
μ	0.08	Rate of Immigration in to Groups

II. Within-Group Equilibrium (f stationary)

Variable	Value	Description
f^*	0.70	Fraction of Group who are Reciprocators
g^*	0.51	Fraction of Pool who are Reciprocators
$\lambda_g(f^*)$	0.00	Rate of Growth of Groups
$\lambda_n(f^*)$	0.00	Rate of Growth of Pool
$\sigma_n(f^*)$	0.10	Shirking Rate of Self-Interested
$\pi_r(f^*)$	0.00	Reproduction Rate of Reciprocators in Groups
$\phi_n(f^*)$	0.03	Reproduction Rate of Self-Interested in Groups
t_o^*	5.38	Expected Duration of Spell of Ostracism
s^*	0.29	Fitness Cost of Ostracism

III. Population-Wide Equilibrium

Variable	Value	Description
p^*	0.70	Fraction of Population in Groups
p_r	0.76	Fraction of Reciprocators who are in Groups
p_n	0.59	Fraction of Self-Interested who are in Groups
ρ	0.65	Fraction of Population who are Reciprocators

Figure 1: Simulation of the model with $b = c = 0.15$, $q = 0.19$, $\gamma = 0.07$, and $\mu = 0.08$.

To show that the equilibrium is stable, we note that

$$\frac{df}{dt} = \frac{f + \pi_r f + \mu g}{f + \pi_r f + \pi_n(1 - f) + \mu} - f, \quad (17)$$

and it is easy to check that

$$\frac{dp}{dt} = p(1 - p)(\lambda_g - \lambda_n). \quad (18)$$

We must add a behavioral equation specifying how self-interested agents change their beliefs concerning the cost s of ostracism with the actual cost, given by (10). We assume the rate of adjustment of s is proportional to the difference between s and its equilibrium value $s^* = t_o(\phi_n(f) - \phi_o)$; i.e.,

$$\dot{s} = -s(s - s^*). \quad (19)$$

Using (17), (18), and (19), we find that the eigenvalues of the Jacobian matrix for the dynamical system are $(-1.10, -0.28, -0.05)$. Since these eigenvalues are all negative, the equilibrium is stable.

4 The Evolution of Strong Reciprocity

Can this model illuminate a process by which strong reciprocity might have become common in human populations? Do the interactions modeled here capture the relevant aspects of the social and physical environments of *Homo sapiens sapiens* during the past 200,000 to 50,000 years?⁸ To answer this question we turn to recent and contemporary accounts of societies generally thought to resemble the foraging bands that were common during this period, among them the !Kung of Botswana and Namibia, the Ache of Paraguay, Batek of Malaysia, Hadza of Tanzania, Pandaram and Paliyan of South India, the Inuit of the Northwest territories, and the Mbuti Pygmies of Zaire. On the basis of this reading, we believe that our model may be illuminating.⁹ There is evidence that in some contemporary simple societies the lazy and the stingy are punished. Balikci (1970):177 reports the following concerning the Netsilik, an isolated tribe of Arctic hunters living on the Arctic coast:

...there is a general rule...according to which all able bodied men should contribute to hunting, and the returns of the hunt should be shared according to established custom. Any activity in exception to this rule was bound to provoke criticism, various forms of conflict, and frequently social ostracism. (176)...lazy hunters were barely tolerated by the community. They were the objects of back biting and ostracism...until the opportunity came for an open quarrel. Stingy men who shared in a niggardly manner were treated similarly. (177)

And Lee (1979):458 reports that

The most serious accusations one !Kung can level against another are the charge of stinginess and the charge of arrogance. To be stingy, or far-hearted, is to hoard one's goods jealously and secretively, guarding them "like a hyena." The corrective for this is to make the hoarder

⁸This is the time span of anatomically modern humans reported by Klein (1989):344. Foley's (1987):22 estimate is 100,000 years. The horticultural societies that eventually replaced foraging bands almost everywhere appeared 12-10,000 years ago. Even Klein's lower limit for the appearance of modern humans leaves ample time for significant change in gene distributions to have taken place under the kinds of selection pressures at work.

⁹Our main sources are listed in footnote 3. The difficulty in making inferences about simple societies during the late Pleistocene on the basis of contemporary simple societies is stressed by Foley (1987):75-78 and Kelly (1995).

give “till it hurts”; that is to make him give generously and without stint until everyone can see that he is truly cleaned out. In order to ensure compliance with this cardinal rule the !Kung browbeat each other constantly to be more generous and not to hoard.

Lethal violence among the !Kung is quite high so the costs of these conflicts must sometimes be borne by those seeking to uphold norms of sharing (Lee 1979).¹⁰ More extensive evidence of punishment of norm violators is provided by Christopher Boehm’s (1993) survey of the many studies in this area.

...intentional leveling linked to an egalitarian ethos is an immediate and probably an extremely widespread cause of human societies’ failing to develop authoritative or coercive leadership. (226)

Bruce Knauff (1991):393,395 adds:

In all ethnographically known simple societies, cooperative sharing of provisions is extended to mates, offspring, and many others within the band. ...This sharing takes place well outside the range of immediate kin, viz. among the diverse array of kin and non-kin who constitute the typical residence group of 25+ persons. Archeological evidence suggests that widespread networks facilitating diffuse access to and transfer of resources and information have been pronounced at least since the Upper Paleolithic...The strong internalization of a sharing ethic is in many respects the *sine qua non* of culture in these societies.

Using data from forty-eight surviving simple societies, Boehm (1993):228 concluded that

the primary and most immediate cause of egalitarian behavior is a moralistic determination on the part of a local group’s main political actors that no one of its members should be allowed to dominate the others.

Boehm further sought to determine whether intentional behavior (notably, social sanctioning) that had a leveling effect was widespread in such societies and more specifically whether it had any significant effects in suppressing the growth of authoritarian leadership. He found evidence that arrogant members of the group are constrained by public opinion, criticism and ridicule, disobedience, and extreme sanction:

¹⁰By contrast to the reports of Lee and Balikci, however, Endicott (1988):118 reports horror expressed by a Batek informant at the thought of exiling a member whose laziness had caused some resentment.

...assassination is reported in 11 out of the 48...behaviors that terminated relations with an overly assertive individual or removed him from a leadership role involved 38 of the 48 societies, while in an additional 28 instances the person was manipulated by social pressure...the great majority of these misbehaviors involve dominance or self-assertion. (231)

among simple foragers, ...group execution of overassertive persons seems to be rather frequent. (239)

We have modelled punishment simply as ostracism from the group. But in the ethnographic record it takes several forms, including group fissioning to minimize interacting with shirkers and the withdrawal of cooperation from shirkers who remain co-resident. Those who have violated a norm may also leave a group in anticipation of more serious punishment. Extensions of the model to include these forms of punishment are straightforward.

Our reading of the ethnographic and paleoanthropological evidence is that our model may capture the salient social and ecological conditions of the late Pleistocene. This alone is not adequate, of course, for we must also show that the model can account for the proliferation of reciprocators in a population composed of self-interested types, as our ancestral populations undoubtedly were.

Such a population, a small fraction of whom are reciprocators, we will suppose, initially work alone, but form groups for mutual protection and raising young, as is the case among primates in general. If many such groups are forming and dissolving by random draws from the population, one, by chance will have a distribution of types within the basin of attraction of f^* . It will then evolve as a social group with its equilibrium distribution of reciprocators. The superior fitness of its members will then lead to the expansion and fissioning of this group until it becomes generalized in the population.¹¹

5 Conclusion

Other cases of costly enforcement of norms relevant to the model arise because its application is considerably more general than the case of working and shirking with which we have motivated it. Suitably emended, the model covers many generic cases of adherence to group-beneficial norms, and punishment for violation of these

¹¹We do not address the manner in which a small group of reciprocators might constitute a group and establish group norms except note that the process could easily come about simply by an extension to non-kin of common within-kin group practices (Boehm 2000).

norms. The extension from team production to the sharing of food acquired individually has already been mentioned and is readily accomplished. A more ambitious extension is to the norm of monogamy, which if possible would considerably expand the scope of our model by encompassing what appears to be a quite common norm in hunter gather bands and a frequent occasion for the sanctioning of violators.

Suppose there is norm that restricts copulations to monogamous pairs, which when violated leads to strife within a group or lessens its effectiveness in acquiring food, insuring against stochastic events, or defending itself, all of which reduce fitness levels of group members. Those who violate the norm, however, enhance their fitness by an amount b . Let σ_n represent the fraction of self-interested types violating the norm of monogamy, and suppose reciprocators never violate it. If group fitness costs of violations of the norm are simply linear in σ , then in the absence of monitoring and ostracism the self-interest type's payoff in the absence of monitoring and punishment is $q(1 - \sigma) - b(1 - \sigma_n)$, while the corresponding payoff for reciprocators is $q(1 - \sigma) - b$, where $q - b$ is just the fitness level in a group uniformly conforming to the norm, and as before $q > b$, so adherence to the norm is group beneficial. We thus reproduce the working-shirking-monitoring model exactly, and therefore believe the model as we have developed it is applicable to a wide range of concrete problems of norm adherence likely to arise in small stateless groups.

It might be thought that the mechanism underlying our model might be vulnerable to the emergence of actors who, like reciprocators, never shirk ($\sigma_j = 0$), but like self-interested types, never punish violators. It is clear that a 'cheater' with this pattern of behavior would be more fit than a reciprocator, and would never be ostracized. Hence such 'cheaters' would drive out reciprocators, after which self-interested types would always shirk, thus destroying the group. We do not regard this objection as decisive for several reasons. First, a cheater type is irrational, in the sense of not choosing a best response to the other group actors. A fitness-maximizing mutant would simply mimic the self-interested type described in our model, thus choosing a strictly positive level of shirking. Nevertheless, We would expect social groups to develop mechanisms to safeguard against invasion by non-punishing mutants, and other types of irrational behavior that pose a threat to the group. One simple way to guard against such cheaters involves reciprocators punishing cheaters for not punishing in the same way they punish self-interested types for shirking, as modeled by Axelrod (1986).

Second, the two traits exhibited by reciprocators are likely to have become pleiotropically linked, the mutations effectively delinking the traits having proven non-viable due to group extinctions among those experiencing these mutations, or for other reasons outside the model. Nor is this possibility a *Deus ex machina*, since the pleiotropic linking of traits is not merely a fortuitous possibility, but in

fact is a likely evolutionary outcome in a situation where two traits separately are deleterious but together are fitness enhancing.

Third, the cognitive and affective traits required to fashion, learn, detect violations of, and wish to uphold social norms may be genetically transmitted, while the content of the norms (and in particular the linking of nonshirking and punishing) may be culturally transmitted. For example, one's unwillingness to join in the punishment of a norm violator (which according to Boehm (1993) is often collective and hence public) would itself be punished through a cultural convention. Notice that this possible cultural linking of norm adherence and the punishment of violators does not trivialize the problem, as the fundamental puzzle remains, namely how could this individually costly mélange of behaviors overcome its fitness disadvantage within groups?

In sum, we think that the model, suitably extended to cover generic norm adherence and to accommodate movement between groups as well as group dissolution and formation, may adequately account for those fitness determining individual interactions in groups during the late Pleistocene.

We do not know that a human predisposition to strong reciprocity evolved as we have described. But it might well have. Our results convince us that an evolutionary process based on genetic inheritance under the influence of natural selection is capable of accounting for the considerable extent of strong reciprocity observed in contemporary society. If we are right, the experimental, historical and other evidence of strong reciprocity may appear to be expressions of human propensities rather than puzzling behaviors inviting *ad hoc* explanation.

REFERENCES

- Axelrod, Robert, *The Evolution of Cooperation* (New York: Basic Books, 1984).
- , “An Evolutionary Approach to Norms,” *American Political Science Review* 80 (1986):1095–1111.
- and William D. Hamilton, “The Evolution of Cooperation,” *Science* 211 (1981):1390–1396.
- Balikci, Asen, *The Netsilik Eskimo* (New York: Natural History Press, 1970).
- Bergstrom, Theodore C., “On the Evolution of Altruistic Ethical Rules for Siblings,” *American Economic Review* 85,1 (March 1995):58–81.
- and Oded Stark, “How Altruism can Prevail in an Evolutionary Environment,” *American Economic Review* 83,2 (May 1993):149–155.
- Bester, Helmut and Werner Güth, “Is Altruism Evolutionarily Stable?,” *Journal of Economic Behavior and Organization* 34,2 (February 1998):193–209.

- Bewley, Truman F., *Why Wages Don't Fall During a Recession* (Cambridge: Harvard University Press, 2000).
- Binmore, Ken, *Game Theory and the Social Contract: Just Playing* (Cambridge, MA: MIT Press, 1998).
- Blinder, Alan S. and Don H. Choi, "A Shred of Evidence on Theories of Wage Stickiness," *Quarterly Journal of Economics* 105,4 (November 1990):1003–15.
- Blurton Jones, Nicholas G., "Tolerated Theft: Suggestions about the Ecology and Evolution of Sharing, Hoarding, and Scrounging," *Social Science Information* 26,1 (1987):31–54.
- Boehm, Christopher, "The Evolutionary Development of Morality as an Effect of Dominance Behavior and Conflict Interference," *Journal of Social and Biological Structures* 5 (1982):413–421.
- , *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.
- , *Hierarchy in the Forest: The Evolution of Egalitarian Behavior* (Cambridge, MA: Harvard University Press, 2000).
- Boorman, Scott A. and Paul Levitt, *The Genetics of Altruism* (New York: Academic Press, 1980).
- Boyd, Robert and J. Lorberbaum, "No Pure Strategy Is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game," *Nature* 327 (1987):58–59.
- and Peter J. Richerson, "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups," *Ethology and Sociobiology* 113 (1992):171–195.
- Cashdan, Elizabeth A., "Egalitarianism among Hunters and Gatherers," *American Anthropologist* 82 (1980):116–120.
- Chagnon, Napoleon A., *Yanomamö: The Fierce People* (New York: Holt, Rinehart and Winston, 1977).
- Damas, David, "Central Eskimo Systems of Food Sharing," *Ethnology* 11,3 (1972):220–240.
- Endicott, Kirk, "Property, Power and Conflict among the Batek of Malaysia," in T. Ingold, D. Riches, and J. Woodburn (eds.) *Hunters and Gatherers* (New York: St. Martin's Press, 1988) pp. 110–127.
- Fehr, Ernst and Armin Falk, "Wage Rigidity in a Competitive Incomplete Contract Market," *Journal of Political Economy* 107,1 (February 1999):106–134.

- and Simon Gächter, “Cooperation and Punishment,” *American Economic Review* 90,4 (September 2000).
- Foley, Robert, *Another Unique Species: Patterns in Human Evolutionary Ecology* (New York: John Wiley and Sons, 1987).
- Fong, Christina, “Social Insurance or Conditional Generosity: The Role of Beliefs about Self- and Exogenous-Determination of Incomes in Redistributive Politics,” 2000. Washington University Department of Political Science.
- Friedman, Daniel and Nirvikar Singh, “On the Viability of Vengeance,” 1999. Economics Department, UC Santa Cruz.
- Ghemawat, Pankaj, “Competitive Advantage and Internal Organization: Nucor Revisited,” *Journal of Economic and Management Strategy* 3,4 (winter 1995):685–717.
- Gilens, Martin, “‘Race Coding’ and White Opposition to Welfare,” *American Political Science Review* 90,3 (September 1996):593–604.
- , *Why Americans Hate Welfare* (University of Chicago Press, 1999).
- Güth, Werner, “An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives,” *International Journal of Game Theory* (1995):323–344.
- and Menahem E. Yaari, “Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach,” in Ulrich Witt (ed.) *Explaining process and change: Approaches to evolutionary Economics* (Ann Arbor: University of Michigan Press, 1992) pp. 23–34.
- Güth, Werner and Reinhard Tietz, “Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results,” *Journal of Economic Psychology* 11 (1990):417–449.
- Guttman, Joel M., “Rational Actors, Tit-for-Tat Types, and the Evolution of Cooperation,” *Journal of Economic Behavior and Organization* 29,1 (1996):27–56.
- Hansen, Daniel G., “Individual Responses to a Group Incentive,” *Industrial and Labor Relations Review* 51,1 (October 1997):37–49.
- Hawkes, Kristen, “Sharing and Collective Action,” in E. A. Smith and B. Winterhalder (eds.) *Evolutionary Ecology and Human Behavior* (New York: Aldine de Gruyter, 1992) pp. 269–300.
- , “Why Hunter-Gatherers Work: An Ancient Version of the Problem of Public Goods,” *Current Anthropology* 34,4 (1993):341–361.
- Hirshleifer, David and Eric Rasmusen, “Cooperation in a Repeated Prisoners’ Dilemma with Ostracism,” *Journal of Economic Behavior and Organization* 12 (1989):87–106.
- Hirshleifer, Jack and Eric Rasmusen, “Cooperation in a Repeated Prisoners’ Dilemma with Ostracism,” *Journal of Economic Behavior and Organization* 12

- (1989):87–106.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith, “Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology,” *Economic Inquiry* 36,3 (July 1998):335–352.
- Huck, Steffen and Jorg Oechssler, “The Indirect Evolutionary Approach to Explaining Fair Allocations,” 1996. Humboldt University, forthcoming in *Games and Economic Behavior*.
- Isaac, R. Mark, James M. Walker, and Arlington W. Williams, “Group Size and Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups,” *Journal of Public Economics* 54 (May 1994):1–36.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, “Fairness as a Constraint on Profit Seeking: Entitlements in the Market,” *American Economic Review* 76,4 (September 1986):728–741.
- Kaplan, Hillard and Kim Hill, “Food Sharing among Ache Foragers: Tests of Explanatory Hypotheses,” *Current Anthropology* 26,2 (1985):223–246.
- and —, “Hunting Ability and Reproductive Success among Male Ache Foragers: Preliminary Results,” *Current Anthropology* 26,1 (1985):131–133.
- , —, Kristen Hawkes, and Ana Hurtado, “Food Sharing among Ache Hunter-Gatherers of Eastern Paraguay,” *Current Anthropology* 25,1 (1984):113–115.
- Kelly, Robert L., *The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways* (Washington, DC: The Smithsonian Institution, 1995).
- Kent, Susan, “And Justice for All: The Development of Political Centralization among Newly Sedentary Foragers,” *American Anthropologist* 93,1 (1989):703–712.
- Klein, Richard G., *Human Career: Human Biological and Cultural* (Chicago: University of Chicago Press, 1989).
- Knauff, Bruce, “Sociality versus Self-interest in Human Evolution,” *Behavioral and Brain Sciences* 12,4 (1989):12–13.
- , “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,” September 1998. Graduate School of Business, University of Chicago.
- 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.
- Lee, Richard Borshay, *The !Kung San: Men, Women and Work in a Foraging Society* (Cambridge, UK: Cambridge University Press, 1979).

- Luttmer, Erzo F. P., "Group Loyalty and the Taste for Redistribution," 1998. University of Chicago Business School.
- Moore, Jr., Barrington, *Injustice: The Social Bases of Obedience and Revolt* (White Plains: M. E. Sharpe, 1978).
- Piketty, Thomas, "Attitudes Toward Income Inequality in France: Do People Really Disagree?," 1999. CEPREMAP, Paris.
- Robson, Arthur J., "Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake," *Journal of Theoretical Biology* 144 (1990):379–396.
- Rogers, Alan R., "Group Selection by Selective Emigration: The Effects of Migration and Kin Structure," *American Naturalist* 135,3 (March 1990):398–413.
- Samuelson, Paul, "Complete Genetic Models for Altruism, Kin Selection, and Like-Gene Selection," *Journal of Social and Biological Structures* 6,1 (January 1983):3–15.
- 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.
- Trivers, R. L., "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971):35–57.
- Woodburn, James, "Egalitarian Societies," *Man* 17,3 (1982):431–451.
- and Alan Barnard, "Property, Power and Ideology in Hunter-Gathering Societies: An Introduction," in T. Ingold, D. Riches, and J. Woodburn (eds.) *Hunters and Gatherers* (New York: St. Martin's Press, 1988) pp. 4–31.