University of Massachusetts - Amherst
ScholarWorks@UMass Amherst

Economics Department Working Paper Series | Economics

2000

# Strong Reciprocity and Human Sociality

Herbert Gintis

Follow this and additional works at: http://scholarworks.umass.edu/econ_workingpaper

Part of the Economics Commons

Recommended Citation
Gintis, Herbert, "Strong Reciprocity and Human Sociality" (2000). *Economics Department Working Paper Series*. 82.
http://scholarworks.umass.edu/econ_workingpaper/82

This Article is brought to you for free and open access by the Economics at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Economics Department Working Paper Series by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

**Strong Reciprocity and Human Sociality***

Herbert Gintis
Department of Economics
University of Massachusetts, Amherst
Phone: 413-586-7756
Fax: 413-586-6014
Email: hgintis@mediaone.net
Web: http://www-unix.oit.umass.edu/~gintis
Running Head: Strong Reciprocity and Human Sociality

March 11, 2000

**Abstract**

Human groups maintain a high level of sociality despite a low level of relatedness among group members. The behavioral basis of this sociality remains in doubt. This paper reviews the evidence for an empirically identifiable form of prosocial behavior in humans, which we call 'strong reciprocity,' that may in part explain human sociality. A strong reciprocator is predisposed to cooperate with others and punish non-cooperators, even when this behavior cannot be justified in terms of extended kinship or reciprocal altruism. We present a simple model, stylized but plausible, of the evolutionary emergence of strong reciprocity.

## 1 Introduction

Human groups maintain a high level of sociality despite a low level of relatedness among group members. Three types of explanation have been offered for this phenomenon: reciprocal altruism (Trivers 1971, Axelrod and Hamilton 1981), cultural group selection (Cavalli-Sforza and Feldman 1981, Boyd and Richerson 1985) and genetically-based altruism (Lumsden and Wilson 1981, Simon 1993, Wilson and Dugatkin 1997). These approaches are of course not incompatible. Reciprocal

*I would like to thank Lee Alan Dugatkin, Ernst Fehr, David Sloan Wilson, and the referees of this Journal for helpful comments, Samuel Bowles and Robert Boyd for many extended discussions of these issues, and the MacArthur Foundation for financial support. This paper is dedicated to the memory of W. D. Hamilton.

altruism is almost certainly a very important part of human sociality, and if altruism has a genetic basis, it would almost necessarily follow that there has been a coevolutionary development of genes and culture.

This paper reviews the evidence for an empirically identifiable form of prosocial behavior in humans that fits neither the extended kinship and or reciprocal altruism frameworks, and probably has a significant genetic component. We call this 'strong reciprocity.' A strong reciprocator is predisposed to cooperate with others and punish non-cooperators, even when this behavior cannot be justified in terms of self-interest or extended kinship. We present a simple yet plausible model of the evolutionary emergence of strong reciprocity.

## 2 The Conditions for Sustaining Cooperation

A group of $n$ individuals faces in each time period a 'public goods game' in which each member, by sacrificing an amount $c > 0$, contributes an amount $b > c$ shared equally by the other members of the group (all costs and benefits are in fitness units).[1]

If all members cooperate, each receives a net payoff of $b - c > 0$. However the only Nash equilibrium in this game is universal defection, in which no member contributes, and all members have baseline fitness zero (an arbitrary constant can be added to all fitnesses to account for the growth rate of the overall population). We also assume an individual not in a cooperating group has baseline fitness zero.

While cooperation is not an equilibrium outcome in a single play of this public goods game, it can be sustained under appropriate conditions if the game is repeated. Specifically, suppose a members's contribution is publicly observable, and in any period a player who fails to contribute $c$ is ostracized from the group. Suppose also that group disbands spontaneously at the end of a given period (due to war, pestilence, climate change, and the like) with probability $1-\delta$. Let $\pi$ be a member's total expected fitness when contributing, assuming all other members contribute. Then $\pi$ can be determined by noting that the current period net fitness gain is $b-c$, plus with probability $\delta$ the game is continued and again has value $\pi$ in the next

[1] For a review of the evidence concerning cooperation in nonhuman species see Dugatkin (1997). For a review of cooperation in the two-player public goods game (usually called the iterated Prisoner's Dilemma), see Dugatkin (1998). Following Axelrod and Hamilton (1981), most models deal with repeated two-person interactions, although Boyd and Richerson (1988,1992) and a few others deal with larger groups. Sethi and Somanathan (1996) is close to this paper in modeling endogenous punishment in a public goods game, but their model predicts a zero level of strong reciprocators in equilibrium.

period. Therefore we have $\pi = b - c + \delta\pi$, which gives [2]

$$\pi = \frac{b-c}{1-\delta}. \tag{1}$$

A player will contribute, then, as long as $(b-c)/(1-\delta) > b$, since by not contributing, the member earns $b$ during the current period, but is ostracized at the end of the period. Rearranging terms in this inequality, we get

Theorem 1. *Suppose c be the fitness cost to a group member of cooperating, b is the fitness gain to others in the group when a member cooperates, and $\delta$ is a discount factor representing the probability that the group will remain constituted for at least one more period.*[3] *Then cooperation can be sustained in the repeated public goods game if and only if*

$$\frac{c}{b} \leq \delta.$$

Theorem 1 is of course a completely standard result. With $n = 2$ is analogous to William Hamilton's (1964) inclusive fitness criterion (where $\delta$ represents the degree of relatedness), Robert Triver's (1971) reciprocal altruism mechanism (where $\delta = 1$), and Robert Axelrod's (1984) condition for cooperation in the repeated prisoner's dilemma. However, the explicit presence of the discount factor $\delta$ in Theorem 1 makes it clear that *however great the net benefits of cooperation, if groups disband with high probability, then cooperation among self-interested agents cannot be sustained*.[4] Moreover, periodic social crises are not implausible, since population contractions were likely common in the evolutionary history of *Homo sapiens* (Boone and Kessler 1999). The very low rate of growth of the human population over the whole prehistoric period, plus the high rate of human population growth in even poor contemporary foraging societies in good times (Keckler 1997), suggests periodic crises occurred in the past. Moreover, flattened mortality profiles of prehistoric skeletal populations indicate population crashes ranging from 10% to

[2]Equation (1) can also be derived by noting that $(b-c)\delta^n$ is the expected return in period $n$, so

$$\pi = (b-c)(1+\delta+\delta^2+\ldots) = (b-c)/(1-\delta).$$

[3]In general the discount factor $\delta$ is the ratio of the contribution to fitness of a unit payoff in the next period to a unit payoff in the current period. In addition to the probability of group dissolution, this ratio generally depends, among other things, on an individual's age and health. We abstract from these factors in this paper.

[4]To my knowledge, endogenous variation in the discount factor $\delta$, central to explaining a high frequency of strong reciprocators in this paper, has not previously been modeled. Nor has the relationship between group longevity and the prevalence of reciprocal altruism in nonhuman species been subjected to systematic empirical investigation. See however Dugatkin and Alfieri (1992).

54% at a mean rate of once per thirty years (Keckler 1997). Finally, optimal foraging models of hunter-gatherer societies often predict stable limit cycles (Belovsky 1988).

In contrast to the self-interested agents assumed in Theorem 1, a strong reciprocator cooperates and punishes noncooperators without considering the value of $\delta$; i.e., even when the probability of future interactions is low. As we shall see, when $\delta$ is low, the presence of strong reciprocators can allow the group to secure the benefits of cooperation. However strong reciprocators are altruists, since they bear surveillance and punishment costs not borne by self-interested group members, so they can persist in equilibrium only if certain conditions, which we develop below, are satisfied.

## 3 Experimental Evidence for Strong Reciprocity

An extensive body of evidence suggests that a considerable fraction of the population, in many different societies, and under many different social conditions, including complete anonymity, behave like the strong reciprocator. We here review of some of the evidence. For additional evidence, including the results of common pool resource and trust games, see Gintis (2000).

### 3.1 The Ultimatum Game

In the *ultimatum game*, under conditions of anonymity, one player, called the 'proposer,' is handed a sum of money, say \$10, and is told to offer any number of dollars, from \$1 to \$10, to the second player, who is called the 'responder.' The responder, again under conditions of anonymity, can either accept the offer, or reject it. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

There is only *one* strategy in which the self-interested responder moves optimally in response to the proposer's offer: accept whatever is offered. However when actually played, this outcome is almost never attained or even approximated. In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts (50% of the total being the modal offer), and respondents frequently reject low offers (e.g., offers below 30%). These results are obtained in experiments with stakes as high as three months' earnings. For a review of ultimatum game experiments, see Güth and Tietz (1990), Roth (1995), and Camerer and Thaler (1995).

Are these results culturally dependent? Do they have a strong genetic component, or do all 'successful' cultures transmit similar values of reciprocity to individuals? Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991) conducted ultimatum games in four different countries (United States, Yugoslavia, Japan, and Israel), and found that while the level of offers differed a bit in different countries, the probability of an offer being rejected did not. This indicates that both proposers and respondents share the same notion of what is considered cooperative behavior in sharing a windfall gain in that society, that proposers adjust their offers to reflect this common notion, and respondents punish uncooperative splits by refusing the offer. The differences in level of offers across countries, by the way, were relatively small.

When asked why they offer more than the lowest possible amount, proposers commonly say that they are afraid that respondents will consider low offers unfair and reject them. When respondents reject offers, they give virtually the same reasons for their actions. In all of the above experiments, by the way, a significant fraction of subjects (about a quarter, typically) conform to the self-interested preferences of *Homo economicus*, and it is often the self-serving behavior of this minority that, when it goes unpunished, unravels initial generosity and cooperation when the game is repeated.

### 3.2 The Public Goods Game

Another important experimental setting in which strong reciprocity has been observed is that of the *public goods game*, designed to illuminate such problems as the voluntary payment of taxes and contribution to team and community goals. Public goods experiments have been run many times, under varying conditions, beginning with the pioneering work of the sociologist G. Marwell, the psychologist R. Dawes, the political scientist J. Orbell, and the economists R. Isaac and J. Walker in the late 1970's and early 1980's.[5] The following is a common variant of the game. Ten subjects are told that \$1 will be deposited in each of their 'private accounts' as a reward for participating in each round of the experiment. For every \$1 a subject moves from his 'private account' to the 'common pool,' the experimenter will deposit \$0.50 in the private accounts of each of the subjects at the end of the game (this corresponds to $c = 0.50$ and $b = 4.50$ in our previous model). This process will be repeated ten times, and at the end, the subjects can take home whatever they have in their private accounts.

If all ten subjects are cooperative, each puts \$1 in the common pool at the end of each round, generating a public pool of \$10; the experimenter then put \$5 in

[5] For a summary of this research and an extensive bibliography, see Ledyard (1995).

the private account of each subject. After ten rounds of this, each subject has $50. Suppose, by contrast, that one subject is self-interested while the others are strong reciprocators. The self-interested one keeps his $1-per-round in his private account, whereas the cooperative ones continue to put $1 in the public pool. In this case, the self-interested subject who free rides on the cooperative contributions of others ends up with $55 at the end of the game, while the other players will end up with $45 each. But if all players are self-interested, then no one contributes to the public pool, and each ends up with $10 at the end of the game. And if one player cooperates, while the others behave in a self-interested manner, the cooperative player will have $5 at the end of the game, while the others will have $15. It is thus clear that this is indeed an iterated Prisoner's Dilemma—whatever other players do on a particular round a player's highest payoff comes from contributing nothing to the common pool.

Public goods experiments show that only a fraction of subjects conform to the self-interested model, contributing nothing to the common pool. Rather, in a one-stage public goods game, people contribute on average about half of their private account. The results in the early stages of a repeated public goods game are similar. In the middle stages of the repeated game, however, contributions begin to decay, until at the end, they are close to the self-interested level—i.e., zero.

Could we not explain the decay of public contribution by *learning*: the participants really do not understand the game at first, but once they hit upon the free-riding strategy, they apply it? This is not likely. One indication that learning does not account for the decay of cooperation is that increasing the number of rounds of play (when this is known to the players) leads to a decline in the rate of decay of cooperation (Isaac, Walker and Williams 1994). Similarly, Andreoni (1988) finds that when the whole process is repeated with the same subjects but with different group composition, the initial levels of cooperation are restored, but once again cooperation decays as the game progresses. Could the explanation lie in the fact that people cooperate less as the end of the game approaches? This explanation fails because, as described below, if we change to rules to allow player to punish non-contributors, cooperation is sustained up to and including the final round.

### 3.3 The Public Goods Game with Punishment

The most plausible explanation for the decay of cooperation is that cooperative contributors attempt to retaliate against free-riders in the only way available to them—by not contributing themselves (Andreoni 1995). Subjects often report this behavior retrospectively. More compelling, however, is the fact that when subjects are given a more constructive way of punishing defectors, they use it to sustain

cooperation even at a cost to themselves (Dawes, Orbell and Van de Kragt 1986, Sato 1987, Yamagishi 1988a,1988b,1992).

For instance, in Ostrom, Walker and Gardner (1992) subjects interacted for about 25 periods in a public goods game, and by paying a 'fee,' subjects could impose costs on other subjects by 'fining' them. Since fining costs the individual who uses it, but the benefits of increased compliance accrue to the group as a whole, the only subgame perfect Nash equilibrium in this game is for no player to pay the fee, so no player is ever punished for defecting, and all players defect by contributing nothing to the common pool. However the authors found a significant level of punishing behavior. The experiment was then repeated with subjects being allowed to communicate, without being able to make binding agreements. In the framework of the *Homo economicus* model, such communication is called *cheap talk*, and cannot lead to a distinct subgame perfect equilibrium. But in fact such communication led to almost perfect cooperation (93%) with very little sanctioning (4%).

The design of the Ostrom-Walker-Gardner study allowed individuals to engage in strategic behavior, since costly retaliation against defectors could increase cooperation in future periods, yielding a positive net return for the retaliator. What happens if we remove any possibility of retaliation being strategic? This is exactly what Fehr and Gächter (2000) studied. They set up a repeated public goods game with the possibility of costly retaliation, but they ensured that group composition changed *in every period* so subjects knew that costly retaliation could not confer any pecuniary benefit to those who punish. Nonetheless, punishment of free-riding was prevalent and gave rise to a large and sustainable increase in cooperation levels. Indeed, a high level of cooperation was generally sustained even in the last period of play, showing that the deterioration in cooperation when no punishment is allowed is not simply an end-game effect.

## 4 The Evolution of Strong Reciprocity

A critical weakness of reciprocal altruism is that when a social group is threatened with extinction or dispersal, say through war, pestilence, or famine, cooperation is most needed for survival. But the discount factor $\delta$, which is the probability of group survival for one period, decreases sharply when the group is threatened, since the probability that the group will dissolve increases. Thus *precisely when a group is most in need of prosocial behavior, cooperation based on reciprocal altruism will collapse*, since the discount factor then falls to a level rendering defection an optimal behavior for self-interested agents.

But strong reciprocity can sustain cooperation in the face of such a threat to

the group, and hence might have an evolutionary advantage in situations where groups are frequently threatened. Strong reciprocators, however, are altruists in that they increase the fitness of unrelated individuals at a cost to themselves. For, unlike self-interested agents, who cooperate and punish only if this maximizes their within-group fitness payoff, strong reciprocators cooperate even when this involves a fitness penalty. If strong reciprocity is an evolutionary adaptation, it must be a considerable benefit to a group to have strong reciprocators, and the group benefits must outweigh the individuals costs.[6]

These benefits and costs are conveniently represented in terms of Price's equation (1970), which we express as follows (Frank 1998). Suppose there are groups $i = 1, \ldots, m$, and let $q_i$ be the fraction of the population in group $i$. Let $\pi_i$ be the mean fitness of group $i$, so $\overline{\pi} = \sum_i q_i \pi_i$ is the mean fitness of the whole population. Groups grow from one period to the next in proportion to their relative fitness, so if $q_i'$ is the fraction of the population in group $i$ in the next period, then

$$q_i' = q_i \frac{\pi_i}{\overline{\pi}}.$$

Suppose there is a trait with frequency $f_i$ in group $i$, so the frequency of the trait in the whole population is $\overline{f} = \sum_i q_i f_i$. If $\pi_i'$ and $f_i'$ are the mean fitness of group $i$ and the frequency of the trait in group $i$ in the next period, respectively, then $\overline{f}' = \sum_i q_i' f_i'$, and writing $\Delta f_i = f_i' - f_i$, we have

$$\begin{aligned}
\overline{f}' - \overline{f} &= \sum q_i' f_i' - \sum q_i f_i \\
&= \sum q_i \frac{\pi_i}{\overline{\pi}} (f_i + \Delta f_i) - \sum q_i f_i \\
&= \sum q_i \left( \frac{\pi_i}{\overline{\pi}} - 1 \right) f_i + \sum q_i \frac{\pi_i}{\overline{\pi}} \Delta f_i.
\end{aligned}$$

Now writing $\Delta \overline{f} = \overline{f}' - \overline{f}$, this becomes

$$\overline{\pi} \Delta \overline{f} = \sum q_i (\pi_i - \overline{\pi}) f_i + \sum q_i \pi_i \Delta f_i. \tag{2}$$

The second term in (2) is just $\mathbf{E}[\pi \Delta f]$, the expected value of $\pi \Delta f$, over all groups, weighted by the relative size of the groups. If the trait in question renders individuals bearing it less fit than other group members, this term will be negative, since $\Delta f_i < 0$ within each group. To interpret the first term, note that the covariance between the variables $\pi$ and $f$ is given by

$$\operatorname{cov}(\pi, f) = \sum_i q_i (\pi_i - \overline{\pi})(f_i - \overline{f}),$$

[6]This model is an instance of analyzing *trait groups* in *structured demes*, to use the terminology of Wilson (1977), to which the reader can refer for a general treatment with numerous applications to behavioral ecology. See also Soltis, Boyd and Richerson (1995) and references therein.

and since $\sum_i q_i(\pi_i - \overline{\pi})\overline{f} = 0$, we can write (2) as

$$\overline{\pi}\Delta\overline{f} = \mathrm{cov}(\pi, f) + \mathbf{E}[\pi\Delta f]. \tag{3}$$

Strong reciprocity can thus persist in equilibrium if and only if $\mathrm{cov}(\pi, f) > -\mathbf{E}[\pi\Delta f]$ where $f$ is the frequency of the strong reciprocity trait and $\pi$ is group fitness.

Suppose now that in each 'good' period the group will persist into the next period with probability $\delta^*$, while in a 'bad' period, which occurs with probability $p$, the group persists with probability $\delta_* < \delta^*$ provided members cooperate, but dissolves with probability one if members do not cooperate.

At the beginning of each period, prior to members deciding whether or not to cooperate, the state of the group for that period is revealed. Let $\pi^*$ be the total fitness of a member if all members cooperate, and the state of the group is 'good,' and let $\pi_*$ be the total fitness if members cooperate and the state is 'bad.' Then, the expected fitness before the state is revealed is $\pi = p\pi_* + (1-p)\pi^*$, and using the same argument as in the derivation of (1), we have the following recursion equations:

$$\pi^* = b - c + \delta^*\pi$$
$$\pi_* = b - c + \delta_*\pi,$$

which entail

$$\pi^* = \frac{1 + p(\delta^* - \delta_*)}{1 - \delta^* + p(\delta^* - \delta_*)}(b - c) \tag{4}$$

$$\pi_* = \frac{1 - (1-p)(\delta^* - \delta_*)}{1 - \delta^* + p(\delta^* - \delta_*)}(b - c) \tag{5}$$

$$\pi = \frac{1}{1 - \delta^* + p(\delta^* - \delta_*)}(b - c). \tag{6}$$

When can cooperation be sustained? Clearly if it is worthwhile for an agent to cooperate in a bad period, it is worthwhile to cooperate in a good period, so we need only check the bad period case. The current benefit of defecting is $b$, so the condition for cooperation is $b < \delta_*\pi$. There is a Nash equilibrium in which members thus cooperate in the good state but not in the bad when the following inequalities hold:

$$\delta^*\pi > b > \delta_*\pi. \tag{7}$$

We assume these inequalities hold.

Suppose group $i$ has a fraction $f_i$ of strong reciprocators, who cooperate and punish independent of whether the state of the group is good or bad. Suppose each strong reciprocator inflicts a total amount of harm $h > 0$ on noncooperators, at a

personal cost of retaliation $c_r > 0$. Because of (7), in a bad state self-interested agents always defect unless punished by strong reciprocators. If there are $n_i$ group members, in a bad state $n_i(1-f_i)$ defect, and the total harm inflicted on those caught is $n_i f_i h$, so the harm per defector imposed by strong reciprocators is $f_i h/(1-f_i)$. The gain from defecting in (7) now becomes $b - f_i h/(1-f_i)$. Thus if the fraction $f_i$ of strong reciprocators is at least

$$f_* = \frac{b - \pi\delta_*}{b - \pi\delta_* + h}, \tag{8}$$

complete cooperation will hold. Note that $f_*$ lies strictly between zero and one. Equation (8), where $\pi$ is given by (6), leads to the following

Theorem 2. *The minimum fraction $f_*$ of strong reciprocators needed to induce cooperation is lowered by a decrease in the probability $p$ of the bad state, an increase in the probability of survival $\delta_*$ in the bad state, and/or an increase in the amount of harm $h$ per strong reciprocator inflicted upon noncooperators.*

These properties of the model have a straightforward interpretation. A decrease in $p$ raises the fitness value $\pi$ of being in a cooperative group, thus lowering the fitness gain $b - \delta_*\pi$ from defecting in the bad state, which reduces the amount of punishment needed to induce self-interested members to cooperate. An increase in $\delta_*$ also raises $\pi$, and hence lowers $b - \delta_*\pi$, with the same result. The fact that an increase in $h$ allows for cooperation with a smaller fraction of strong reciprocators is completely obvious, but is perhaps the most interesting of these properties, since probably distinguishes *Homo sapiens* from other species that group and recognizing individuals, hence for which reciprocal altruism might have fitness value.

As a result of the superior tool-making and hunting ability of *Homo sapiens*, humans have the capacity to impose high costs on others (high $h$) at a low cost to themselves (low $c_r$) both absolutely and relative to other species. While size, strength, and vigor generally determine the outcome of animal disputes, often at great cost even to the winner, in human societies even a small number of attackers can defeat the most formidable single enemy at very low fitness cost through the use of coordination, stealth and deadly weapons. Bingham (1999) has stressed the importance of the superior abilities of humans in clubbing and throwing projectiles as compared with other primates, citing Goodall (1964), Plooij (1978) on the relative advantage of humans, and Darlington (1975), Fifer (1987), and Isaac (1987) on the importance of these traits in human evolution. Calvin (1983) argues that the same neural machinery for rapid manual-brachial movements that allow for precision stone-throwing may have developed secondary uses for language reception and production. Theorem 2 suggests one reason why these factors favor the evolution of strong reciprocity.

If $f_i < f_*$ there will be no cooperation in a bad period (we continue to assume the parameters of the model are such that there is always cooperation in the good period). In this situation the group disbands. Using the same argument as that leading to (1), we see that the fitness $\pi_s$ of members of such noncooperative groups satisfies the recursion equation $\pi_s = (1-p)(b-c+\delta^*\pi_s)$, so

$$\pi_s = \frac{1-p}{1-(1-p)\delta^*}(b-c). \tag{9}$$

Our assumption that there is always cooperation in the good state requires that $\delta^*\pi_s > b$, which becomes

$$\frac{\delta^*(1-p)}{1-(1-p)\delta^*}(b-c) > b.$$

Note that the relative fitness benefit from being in a cooperative group is

$$\Delta\pi = \pi - \pi_s = p\pi\frac{1-(1-p)(\delta^*-\delta_*)}{1-(1-p)\delta^*} > 0. \tag{10}$$

For example, suppose $\delta^* = 0.95$, so the expected duration of a group exposed only to 'good' states is 20 years, suppose $p = 0.10$, so a 'bad' period occurs in one year out of ten, and suppose $\delta_* = 0.25$, so a cooperating group survives with 25% probability in a 'bad' period. Then, $\Delta\pi/\pi = 0.255$; i.e., the cooperating group enjoys a 25.5% fitness advantage over the noncooperating group.

We suppose that the fraction of strong reciprocators in a group is common knowledge, and strong reciprocators punish defectors only in groups where $f_i \geq f_*$, and in doing so they each incur the fixed fitness cost of retaliation $c_r$. We shall interpret $c_r$ as a surveillance cost, and since punishment is unnecessary except in 'bad' periods, strong reciprocators will incur this cost only with probability $p$, so the expected fitness cost of being a strong reciprocator is $pc_r$.[7]

We will use Price's equation to chart the dynamics of strong reciprocity, which in this case says the change $\Delta\overline{f}$ in the fraction of strong reciprocators in the population is given by

$$\Delta\overline{f} = \frac{1}{\overline{\pi}}\mathrm{cov}(\pi, f) + \frac{1}{\overline{\pi}}\mathbf{E}[\pi\,\Delta f], \tag{11}$$

[7]An alternative, perhaps more plausible, pair of assumptions is that $c_r$ is expended only when noncooperation is actually detected, and there is some source of stochasticity (for instance imperfect signaling or variable agent behavior) that leads to a positive level of punishment even in cooperative groups. The treatment of $c_r$ as a surveillance cost is simpler and leads to the identical result that strong reciprocators incur positive costs even in a cooperative equilibrium.

where $\overline{\pi}$ is the mean fitness of the population. Let $q_f$ be the fraction of the population in cooperative groups, so

$$q_f = \sum_{f_i \geq f_*} q_i, \tag{12}$$

The fitness of each member of a group with $f_i \geq f_*$ (resp. $f_i < f_*$) is $\pi$ (resp. $\pi_s$), so the average fitness is $\overline{\pi} = q_f \pi + (1 - q_f)\pi_s$. We then have

$$\frac{1}{\overline{\pi}} \mathbf{E}[\pi \, \Delta f] = \sum_{f_i \geq f^*} q_i f_i \frac{\pi}{\overline{\pi}} (-p c_r). \tag{13}$$

Algebraic manipulation gives

$$\frac{\pi}{\overline{\pi}} = \frac{1 - \delta^*(1-p)}{1 - \delta^*(1-p) - p(1-q_f)(1 - (\delta^* - \delta_*)(1-p))},$$

so if we let $f_c = \sum_{f_i \geq f^*} q_i f_i / q_f$, which is the mean fraction of strong reciprocators in cooperative groups, then (13) becomes

$$\frac{1}{\overline{\pi}} \mathbf{E}[\pi \, \Delta x] = -\frac{c_r f_c p q_f (1 - \delta^*(1-p))}{1 - \delta^*(1-p) - p(1-q_f)(1 - (\delta^* - \delta_*)(1-p))}. \tag{14}$$

To evaluate the covariance term, we define $f_s = \sum_{f_i < f^*} q_i f_i / (1 - q_f)$, which is the mean frequency of strong reciprocators in noncooperative groups. Then we have

$$\frac{1}{\overline{\pi}} \mathrm{cov}(\pi_i, f_i) = \frac{(f_c - f_s) p q_f (1 - q_f)(1 - (\delta^* - \delta_*)(1-p))}{1 - \delta^*(1-p) - p(1-q_f)(1 - (\delta^* - \delta_*)(1-p))}. \tag{15}$$

The condition for the increase in strong reciprocity is that the sum of (14) and (15) be positive, which for $q_f > 0$ reduces to

$$\left(1 + \frac{\delta_*(1-p)}{1 - \delta^*(1-p)}\right) \frac{f_c - f_s}{f_c} (1 - q_f) > c_r. \tag{16}$$

Note that $0 < q_f < 1$ implies $0 \leq f_s < f_c$, so we have the following

Theorem 3. *Suppose the discount factor is $\delta^*$ in a good period and $\delta_*(< \delta^*)$ in a bad period, and bad periods occur with probability $p > 0$. Suppose (7) holds, so there is cooperation in the good but not the bad periods in groups in which the fraction of strong reciprocators is less than $f_*$, given by (8). Then if the fraction of strong reciprocators in cooperative groups is strictly positive ($q_f > 0$), (16) is the condition for an increase in the fraction of strong reciprocators in the population.*

Let $\overline{f} = f_s(1 - q_f) + f_c q_f$, which is the frequency of strong reciprocity in the whole population. To close the model and thus determine the equilibrium value of $\overline{f}$, we must develop a plausible mechanism for the assignment of individuals to groups, thereby determining $f_c$ and $f_s$ as functions of $\overline{f}$. We shall adopt the conservative assumption that new groups form by the assignment of self-interested individuals and strong reciprocators in proportion to their frequency in the population, so that there is no assortative interaction in the formation of new groups.[8]

For simplicity, we assume a *fixed size founder process*, in which newly formed groups are of a fixed size $k$, and the number of such groups is effectively infinite, so that the frequency of strong reciprocators in a group is given by the binomial distribution; i.e., we assume sampling with replacement in the assignment of individuals to groups.[9] The probability $p_k$ that a newly formed group satisfies $f \geq f_*$ is then given by

$$p_k = \sum_{r \geq f_* k}^{k} \binom{k}{r} \overline{f}^r (1 - \overline{f})^{k-r}, \tag{17}$$

the frequency of strong reciprocators in such groups is given by

$$f_c = \frac{1}{k p_k} \sum_{r \geq f_* k}^{k} r \binom{k}{r} \overline{f}^r (1 - \overline{f})^{k-r}, \tag{18}$$

and the frequency of strong reciprocators in groups with $f < f_*$ is given by

$$f_s = (\overline{f} - f_c q_f)/(1 - q_f), \tag{19}$$

where $q_f$ is given by (12). It follows that (16) cannot be satisfied when $\overline{f} = 1$, since in this case $q_f = 1$. On the other hand, $\overline{f} \geq q_f f_c \geq q_f f_*$, so when $\overline{f}$ is small, so is $q_f$. Then (19) shows that when $\overline{f}$ is small, so is $f_s$. But $f_c \geq f_*$, so both the second and third terms in (16) approach unity for small $\overline{f}$. This proves

Theorem 4. *Under the conditions of Theorem 3, and assuming a fixed size founder process, in newly-forming groups self-interested agents can always invade a population of strong reciprocators, and when the cost $c_r$ of punishing noncooperators is sufficiently low, a small fraction $\overline{f}$ of strong reciprocators can always invade a population of self-interested agents.*

[8]It is generally understood, of course, that the maintenance of altruistic behavior depends on assortative interactions. William Hamilton (1975) first noted that kin selection is based on assortative interactions. Others who have contributed to the theory of assortative interactions include Wilson (1977), Boyd (1982), Michod (1982), Wade (1985), and Boyd and Richerson (1993). Assortative interactions in our model take the form of groups with a high frequency of strong reciprocators lasting longer than other groups.

[9]For a more general analysis of this case, see Cohen and Eshel (1976).

Theorems 2 and 4 suggest the central importance of the amount of harm $h$ an agent can inflict on noncooperators and the cost $c_r$ the agent incurs in so doing. As long as there is a positive fraction of strong reciprocators in the population, (8) shows that sufficiently large $h$ implies $q_f > 0$, where Theorem 2 applies. Theorem 2 then asserts that for sufficiently low cost of retaliation $c_r$, strong reciprocators can invade a population of self-interested agents.[10] Under no condition, however, can strong reciprocators drive self-interested agents to extinction, since (16) is necessarily violated when $q_f$ is near unity.

Simulating this model (I used Mathematica 3.0) allows us to assess the plausibility of the parameters involved and the nature of the equilibrium fraction $\overline{f}^*$ of strong reciprocators in the population. In equilibrium, (16) must hold as an equality, since the fraction of strong reciprocators in newly formed groups must be equal to that of the population as a whole.

Equations (17) and (18) allow us to estimate the left hand side of (16). The first term on the left hand side is a number greater than unity that, for plausible values of the parameters, lies between 1.0 and 4.0. For instance if $\delta^* = 0.95$, $\delta_* \leq 0.25$, and $p \geq 0.10$, this factor has a minimum of 1.00 and a maximum of 2.47. The lower curve in Figure 1 shows the equilibrium fraction $\overline{f}^*$ of strong reciprocators for values of $c_r$ from 0.05 to 1, when $\delta^* = 0.95$, $\delta_* = 0.10$, $p = 0.10$, there are forty members per group, and $f_* = 3/8$ must be strong reciprocators to induce cooperation in the bad state. The upper curve shows the same relationship when there are eight members per group. The latter curve would be relevant if groups are composed of a small number of 'families,' and the strong reciprocity characteristic is highly heritable within families. It is clear from Figure 1 that the incidence of strong reciprocity can be much higher, especially when the cost of retaliation $c_r$ is high, when family assortative interaction occurs.[11]

[10] Our model thus strongly supports Bingham's (1999) stress on physical factors in explaining cooperation among humans. Bingham makes the stronger claim that human cooperation is based on 'coalitional enforcement' by self-interested agents. This claim is doubtful because coalitional enforcement is a form of reciprocal altruism, which we have shown fails when there is a high probability of group dissolution. Moreover, as we have seen, human revenge and retaliation does not follow the logic of self-interested is behavior.

[11] Models of assortative interaction taking families as a behavioral unit include Wilson (1977), Boyd (1982), Michod (1982), Wade (1985) and Boyd and Richerson (1993). The argument that hunter-gatherer groups in both recent and Pleistocene periods have consisted of a small coalitions of families is made by Kaplan and Hill (1985), Blurton-Jones (1987), Knauft (1991), Boehm (1993), and Hawkes (1993).

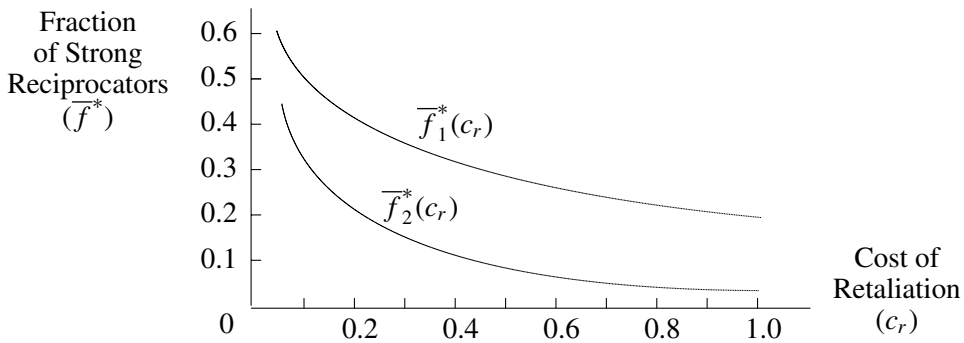


**Figure 1:** The Equilibrium Fraction of Strong Reciprocating Families: A Computer Simulation

## 5 Conclusion

Reciprocal altruism leads to a high level of cooperation in human societies, and many behavioral scientists believe that reciprocal altruism is sufficient to explain human sociality. Economists are particularly favorable to this belief, since reciprocal altruism is a behavior supported by the so-called *rational actor model*, which much of economic analysis is presumes.

However laboratory experiments, conducted in many different social settings by different research groups, consistently show that people tend to behave prosocially and punish antisocial behavior, at a cost to themselves, even when the probability of future interactions is extremely low, or zero. We call this *strong reciprocity*, in contrast with the *weak reciprocity* associated with reciprocal altruism, because the former behavior is robust in the face of changes in the probability of future interaction.

Strong reciprocity is a form of altruism, in that it benefits group members at a cost to the strong reciprocators themselves. This papers shows that there is a plausible evolutionary model supporting the emergence of strong reciprocity. This model based on the notion that societies periodically experience extinction-threatening events, and reciprocal altruism will fail to motivate self-interested individuals in such periods, thus exacerbating the threat and increasing the likelihood of group extinction. If the fraction of strong reciprocators is sufficiently high, even self-interested agents can be induced to cooperate in such situations, thus lowering the probability of group extinction.

## References

Andreoni, James, "Why Free Ride? Strategies and Learning in Public Good Experiments," *Journal of Public Economics* 37 (1988):291–304.

— , "Cooperation in Public Goods Experiments: Kindness or Confusion," *American Economic Review* 85,4 (1995):891–904.

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

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

Belovsky, G., "An Optimal Foraging-Based Model of Hunter-Gatherer Population Dynamics," *Journal of Anthropological Archaeology* 7 (1988):329–372.

Bingham, Paul M., "Human Uniqueness: A General Theory," *Quarterly Review of Biology* 74,2 (June 1999):133–169.

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

Boone, James L. and Karen L. Kessler, "More Status or More Children? Social Status, Fertility Reduction, and Long-Term Fitness," *Evolution and Human Behavior* 20,4 (July 1999):257–277.

Boyd, Robert, "Density Dependent Mortality and the Evolution of Social Behavior," *Animal Behavior* 30 (1982):972–982.

— and Peter J. Richerson, *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

— and — , "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.

— and — , "Effect of Phenotypic Variation on Kin Selection," *Proceedings of the National Academy of Sciences* 77 (1993):7506–7509.

Calvin, William H., "A Stone's Throw and its Launch Window: Timing Precision and its Implications for Language and Hominid Brains," *Journal of Theoretical Biology* 104 (1983):121–135.

Camerer, Colin and Richard Thaler, "Ultimatums, Dictators, and Manners," *Journal of Economic Perspectives* 9,2 (1995):209–219.

Cavalli-Sforza, Luigi L. and Marcus W. Feldman, *Cultural Transmission and Evolution* (Princeton, NJ: Princeton University Press, 1981).

Cohen, Dan and Elan Eshel, "On the Founder Effect and the Evolution of Altruistic Traits," *Theoretical Population Biology* 10 (1976):276–302.

Darlington, P. J., "Group Selection, Altruism, Reinforcement and Throwing in Human Evolution," *Proceedings of the National Academy of Sciences* 72 (1975):3748–52.

Dawes, Robyn M., John M. Orbell, and J. C. Van de Kragt, "Organizing Groups for Collective Action," *American Political Science Review* 80 (December 1986):1171–1185.

Dugatkin, Lee Alan, *Cooperation among Animals* (New York: Oxford University Press, 1997).

— , "Game Theory and Cooperation," in Lee Alan Dugatkin and Hudson Kern Reeve (eds.) *Game Theory and Animal Behavior* (Oxford: Oxford University Press, 1998) pp. 38–63.

— and M. Alfieri, "Interpopulational Difference in the Cooperative Strategy Used During Predator Inspection in the Guppy," *Evolutionary Ecology* 6 (1992):519–526.

Fehr, Ernst and Simon Gächter, "Cooperation and Punishment," *American Economic Review (*2000). forthcoming.

Fifer, F. C., "The Adoption of Bipedalism by the Hominids: a New Hypothesis," *Human Evolution* 2 (1987):135–47.

Frank, Steven A., *Foundations of Social Evolution* (Princeton: Princeton University Press, 1998).

Gintis, Herbert, *Game Theory Evolving* (Princeton, NJ: Princeton University Press, 2000).

Goodall, Jane, "Tool-using and Aimed Throwing in a Community of Free-Living Chimpanzees," *Nature* 201 (1964):1264–1266.

Güth, Werner and Reinhard Tietz, "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results," *Journal of Economic Psychology* 11 (1990):417–449.

Hamilton, W. D., "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology* 37 (1964):1–16,17–52.

— , "Innate Social Aptitudes of Man: an Approach from Evolutionary Genetics," in Robin Fox (ed.) *Biosocial Anthropology* (New York: John Wiley and Sons, 1975) pp. 115–132.

Hawkes, Kristen, "Why Hunter-Gatherers Work: An Ancient Version of the Problem of Public Goods," *Current Anthropology* 34,4 (1993):341–361.

Isaac, B., "Throwing and Human Evolution," *African Archeological Review* 5 (1987):3–17.

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.

Kaplan, Hillard and Kim Hill, "Hunting Ability and Reproductive Success among Male Ache Foragers: Preliminary Results," *Current Anthropology* 26,1 (1985):131–133.

Keckler, C. N. W., "Catastrophic Mortality in Simulations of Forager Age-of-Death: Where Did all the Humans Go?," in R. Paine (ed.) *Integrating Archaeological Demography: Multidisciplinary Approaches to Prehistoric Populations. Center for Archaeological Investigations, Occasional Papers No. 24* (Carbondale, IL: Southern Illinois Univerisity Press, 1997) pp. 205–228.

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

Ledyard, J. O., "Public Goods: A Survey of Experimental Research," in J. H. Kagel and A. E. Roth (eds.) *The Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press, 1995) pp. 111–194.

Lumsden, C. J. and E. O. Wilson, *Genes, Mind, and Culture: The Coevolutionary Process* (Cambridge, MA: Harvard University Press, 1981).

Michod, R., "The Theory of Kin Selection," *Annual Review of Ecological Systems* 13 (1982):23–55.

Ostrom, Elinor, 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.

Plooij, F. X., "Tool-using during Chimpanzees' Bushpig Hunt," *Carnivore* 1 (1978):103–106.

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

Roth, Alvin, "Bargaining Experiments," in John Kagel and Alvin Roth (eds.) *The Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press, 1995).

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review* 81,5 (December 1991):1068–1095.

Sato, Kaori, "Distribution and the Cost of Maintaining Common Property Resources," *Journal of Experimental Social Psychology* 23 (January 1987):19–31.

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

Simon, Herbert A., "Altruism and Economics," *American Economic Review* 83,2 (May 1993):156–61.

Soltis, Joseph, Robert Boyd, and Peter Richerson, "Can Group-functional Behaviors Evolve by Cultural Group Selection: An Empirical Test," *Current Anthropology* 36,3 (June 1995):473–483.

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

Wade, Michael J., "Soft Selection, Hard Selection, Kin Selection and Group Selection," *American Naturalist* 125,1 (January 1985):61–73.

Wilson, David Sloan, "Structure Demes and the Evolution of Group-advantageous Traits," *American Naturalist* 111 (1977):157–185.

— and Lee A. Dugatkin, "Group Selection and Assortative Interactions," *American Naturalist* 149,2 (1997):336–351.

Yamagishi, Toshio, "The Provision of a Sanctioning System in the United States and Japan," *Social Psychology Quarterly* 51,3 (1988):265–271.

— , "Seriousness of Social Dilemmas and the Provision of a Sanctioning System," *Social Psychology Quarterly* 51,1 (1988):32–42.

— , "Group Size and the Provision of a Sanctioning System in a Social Dilemma," in W.B.G. Liebrand, David M. Messick, and H.A.M. Wilke (eds.) *Social Dilemmas: Theoretical Issues and Research Findings* (Oxford: Pergamon Press, 1992) pp. 267–287.