J. theor. Biol. (2000) **206,** 169–179
doi:10.1006/jtbi.2000.2111, available online at http://www.idealibrary.com on IDEAL®

Strong Reciprocity and Human Sociality

HERBERT GINTIS*

Department of Economics, University of Massachusetts, Amherst, U.S.A.

(*Received on* 15 *June* 1999, *Accepted in revised form on* 22 *May* 2000)

Human groups maintain a high level of sociality despite a low level of relatedness among group members. 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 approaches have been offered to explain this phenomenon: reciprocal altruism (Trivers, 1971; Axelrod & Hamilton, 1981), cultural group selection (Cavalli-Sforza *et al*., 1981; Boyd & Richerson, 1985) and genetically based altruism (Lumsden & Wilson, 1981; Simon, 1993; Wilson & Dugatkin, 1997; Sober & Wilson, 1998). These approaches are complementary and doubtless all contribute to the explanation of human sociality. The analysis of altruism, however, has tended to argue the plausibility of altruism in general, rather than isolating particular human traits that might have emerged from a group selection process.

This paper reviews the evidence for one such trait—an empirically identifiable form of prosocial behavior in humans that 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, extended kinship, or reciprocal altruism. 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).*

*E-mail: hgintis@mediaone.net, web: http://www-unix.oit.umass.edu/~gintis

*For a review of the evidence concerning cooperation in non-humans and humans, see Dugatkin (1997) and Dugatkin (1998), respectively. Following Axelrod & Hamilton (1981), most models deal with repeated two-person interactions, although Boyd & Richerson (1988, 1992) and a few others deal with larger groups. Sethi & Somanathan (1996) is close to this paper in modeling endogenous punishment in a public goods game, but their model predicts the absence of punishers in equilibrium, a result at variance with observed behavior in human society.

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 member'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 π be a member's total expected fitness when contributing, assuming all other members contribute. Then π can be determined by noting that the current period net fitness gain is $b - c$, plus with probability δ the game is continued and again has value π in the next period. Therefore, we have $\pi = b - c + \delta\pi$, which gives†

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

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 is the fitness cost to a group member of cooperating, b is the fitness gain to others in the group when a member cooperates, and δ is a discount factor representing the probability that the group will remain constituted for at least one more period.‡ Then cooperation can be sustained in the repeated public goods game if and only if*

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

Theorem 1 is of course a completely standard result. With $n = 2$ is analogous to Hamilton's (1964) inclusive fitness criterion (where δ represents the degree of relatedness), Triver's (1971) reciprocal altruism mechanism (where $\delta = 1$), and Axelrod's (1984) condition for cooperation in the repeated prisoner's dilemma. However, the explicit presence of the discount factor δ 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.*§ Moreover, periodic social crises are not implausible, since population contractions were likely common in the evolutionary history of *Homo sapiens* (Boone & Kessler, 1999). The very low rate of growth of the human population over the whole pre-historic 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 pre-historic skeletal populations indicate population crashes ranging from 10 to 54% at a mean rate of once in 30 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 non-cooperators without considering the value of δ, i.e. even when the probability of future interactions is low. As we shall see, when δ 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

† Equation (1) can also be derived by noting that $(b - c)\delta^n$ is the expected return in period n, so we have $\pi = (b - c)(1 + \delta + \delta^2 + \cdots) = (b - c)/(1 - \delta)$.

‡ In general, the discount factor δ 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.

§ To my knowledge, endogenous variation in the discount factor δ, 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 non-human species been subjected to systematic empirical investigation. See, however, Dugatkin & Alfieri (1992).

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 laboratory evidence concerning the *public goods game* as modeled in the previous section. For additional evidence, including the results of dictator, ultimatum, common pool resource and trust games, see Güth & Tietz (1990), Roth (1995), and Camerer & Thaler (1995), and for analytical models, see Gintis (2000).

The public goods game is a direct test of strong reciprocity, and is designed to illuminate such behaviors as contributing to team and community goals, as well as punishing non-contributors. Public goods experiments were first undertaken by the sociologist G. Marwell, the psychologist R. Dawes, the political scientist J. Orbell, and the economists R. Isaac and J. Walker in the late 1970s and early 1980s [see Ledyard (1995) for a summary of this research and an extensive bibliography]. The following is a typical public good game, using a protocol studied by Fehr & Gächter (2000).

Each round of the game consists of each subject being grouped with three other subjects under conditions of strict anonymity. Each subject is then given 20 "points", redeemable at the end of the experimental session for real money. Each subject then places some number of the 20 points in a "common account", and the remainder in the subject's "private account". The experimenter then tells the subjects how many points were contributed to the common account, and adds to the private account of each subject 40% of the total amount in the common account. It follows that if a subject contributes his whole 20 points to the common account, each player will receive eight points at the end of the round. In effect, by cooperating a player loses 12 points but his teammates gain 24 points. In terms of our model of the previous section, $c = 12$ and $b = 24$. The experiment continues for exactly ten rounds, and the subjects are informed of this fact at the start of the experiment.

Clearly, full free riding is a dominant strategy in the public goods game. Public goods experiments, however, show that only a fraction of subjects conform to the self-interested model, contributing nothing to the common account. Rather, subjects begin by contributing on average about half of their endowment to the public account. The level of contributions decays over the course of the ten rounds, until in the final rounds most players are behaving in a self-interested manner (Dawes & Thaler, 1988; Ledyard, 1995). In a meta-study of 12 public goods experiments Fehr & Schmidt (1999) found that in the early rounds, average and median contribution levels ranged from 40 to 60% of the endowment, but in the final period 73% of all individuals ($N = 1042$) contributed nothing, and many of the remaining players contributed close to zero. These results are not compatible with the self-interested actor model, which predicts zero contribution on all rounds, though they might be predicted by a reciprocal altruism model, since the chance to reciprocate declines as end of the experiment approaches. However, we shall see that this is not in fact the explanation of moderate but deteriorating levels of cooperation in the public goods game.

The explanation of the decay of cooperation offered by subjects when debriefed after the experiment is that cooperative subjects became angry at others who contributed less than themselves, and retaliated against free-riding low contributors in the only way available to them—by lowering their own contributions (Andreoni, 1995). Experimental evidence supports this interpretation. When subjects are allowed to punish non-contributors, they do so at a cost to themselves (Dawes *et al.*, 1986; Sato, 1987; Yamagishi, 1988a, b, 1992).

For instance, in Ostrom *et al.* (1992) subjects interacted for 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 Nash equilibrium in this game that does not depend on incredible threats 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 design of the Ostrom–Walker–Gardner study allowed individuals to engage in strategic behavior, since costly punishment of defectors could increase cooperation in future periods, yielding a positive net return for the retaliator. Fehr & Gächter (2000) set up an experimental situation in which the possibility of strategic punishment was removed. They used a ten-round public goods game with costly punishment, employing three different methods of assigning members to groups.|| There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the *Personal* treatment, the four subjects remained in the same group for all ten periods. Under the *Stranger* treatment, the subjects were randomly reassigned after each round. Finally, under the *Perfect Stranger* treatment the subjects were randomly reassigned and assured that they would never meet another subject more than once (in this case, the number of rounds had to be reduced from ten to six to accommodate the size of the subject pool). Subjects earned an average of about $35 for an experimental session.

Fehr & Gächter (2000) performed their experiment for ten rounds with punishment and ten rounds without. Their results are illustrated in Fig. 1. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the partner game, despite strict anonymity, cooperation increases almost to full cooperation, even on the final round. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games.

The contrast between the partner effect and the two stranger effects is worth noting. In the latter case, punishment prevented the deterioration of cooperation, whereas in the former case punishment led to an increase in participation over time, until near full cooperation was achieved. This result suggest that subjects are motivated by the personal desire to punish free riders (the stranger treatment), but are even more strongly motivated when there is an identifiable group, to which they belong, whose cooperative effort is impaired by free riding (the partner treatment). The prosociality of strong reciprocity is thus more strongly manifested, the more coherent and permanent the group in question.

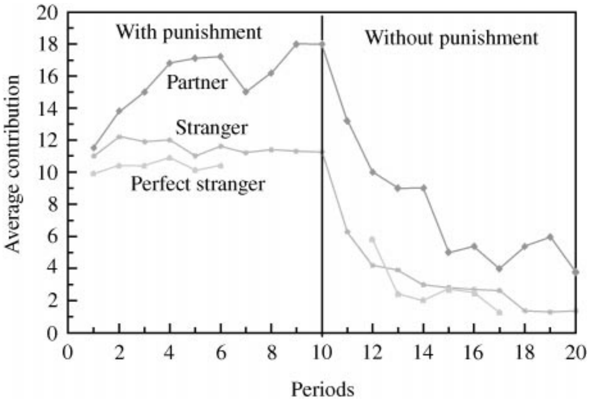

FIG. 1. Average contributions over time in the partner, stranger, and perfect stranger treatments when the punishment conditions is played first (adapted from Fehr & Gächter, 2000).

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 δ, 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

|| Imposing a punishment was quite costly in this experiment. Low levels of punishment (one or two points) was equally costly to punisher and punishee, with higher levels of punishment being relatively more costly to the punisher. Imposing a ten-point punishment—the highest level permitted—cost the punisher 30 points. As we argue below, in human societies we expect the costs of punishing to be quite low compared to the costs of being punished. Thus, this experiment is strongly biased against finding strong reciprocity.

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 outweight the individuals costs.¶

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 π_i be the mean fitness of group i, so $\bar{\pi} = \sum_i q_i \pi_i$ 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

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

Suppose there is a trait with frequency f_i in group i, so the frequency of the trait in the whole population is $\bar{f} = \sum_i q_i f_i$. If π_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 $\bar{f}' = \sum_i q_i' f_i'$, and writing $\Delta f_i = f_i' - f_i$, we have

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

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

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

The second term in eqn (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 π and f is given by

$$\mathrm{cov}(\pi, f) = \sum_i q_i (\pi_i - \bar{\pi})(f_i - \bar{f}).$$

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

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

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 π is group fitness.

Suppose now that in each "good" period the group will persist into the next period with probability δ^*, 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 π^* be the total fitness of a member if all members cooperate, and the state of the group is "good", and let π_* 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 eqn (1), we have the following recursion equations:

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

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

which entail

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

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

¶This model is an instance of analysing *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 of behavioral ecology. See also Soltis *et al.* (1995) and references therein.

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

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 c, so the condition for cooperation is $c < \delta_* \pi$. There is a Nash equilibrium in which members thus cooperate in the good state but not in the bad when the following inequalities hold:

$$\delta^* \pi > c > \delta_* \pi. \quad (7)$$

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 non-cooperators, at a personal cost of retaliation $c_r > 0$. Because of eqn (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 eqn (7) now becomes $c - f_i h/(1 - f_i)$. Thus, if the fraction f_i of strong reciprocators is at least

$$f_* = \frac{c - \pi\delta_*}{c - \pi\delta_* + h}, \quad (8)$$

complete cooperation will hold. Note that f_* lies strictly between zero and one. Equation (8), where π is given by eqn (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 δ_* in the bad state, and/or an increase in the amount of harm h per strong reciprocator inflicted upon non-cooperators.*

These properties of the model have a straightforward interpretation. A decrease in p raises the fitness value π of being in a cooperative group, thus lowering the fitness gain $c - \delta_* \pi$ from defecting in the bad state, which reduces the amount of punishment needed to induce self-interested members to cooperate. An increase in δ_* also raises π, and hence lowers $c - \delta_* \pi$, with the same result.

The fact that an increase in h allows for cooperation with a smaller fraction of strong reciprocators is completely obvious from eqn (8), but is perhaps the most interesting of these properties since, as a result of the superior tool-making and hunting ability of *Homo sapiens*, the ability to inflict costly punishment (high h) at a low cost to the punisher (low c_r), probably distinguishes humans from other species that live in groups and recognizing individuals, hence for which reciprocal altruism might occur. While size, strength, and vigor generally determine the outcome of animal disputes, victory often involving 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 to the attackers 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) and 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 humans are unique in possessing the same neural machinery for rapid manual–brachial movements that allow for precision stone-throwing. 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 eqn (1), we see that the fitness π_s of members of such non-cooperative groups satisfies the recursion equation $\pi_s = (1 - p)(b - c + \delta^* \pi_s)$, so

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

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

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

For example, suppose $p = 0.95$, 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 non-cooperating group.

We suppose that the fraction of strong reciprocators in a group is common knowledge, and strong reciprocators punish defectors only in groups where $f_i \geqslant f_*$, 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.††

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

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

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

$$q_f = \sum_{f_i \geqslant f_*} q_i, \quad (12)$$

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

$$\frac{1}{\bar{\pi}}\mathbf{E}[\pi\Delta f] = \sum_{f_i \geqslant f^*} q_i f_i \frac{\pi}{\bar{\pi}}(-pc_r). \quad (13)$$

Algebraic manipulation gives

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

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

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

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 non-cooperative groups. Then we have

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

** This common knowledge assumption is strong, but it could be dropped by assuming that the single-stage public goods game is played several times in each time period. The level of punishment in the first stage would then correctly signal the frequency of strong reciprocity in the group, so that the common knowledge assumption would hold for the remaining stages. This alternative has the added benefit of being a more plausible representation of human cooperation, and it provided a natural interpretation of c_r as the cost of punishing in the first stage. On the other hand, in this case, stronger assumptions would be needed to conclude that a small fraction of strong reciprocators could invade a population of self-interested types, since strong reciprocators are now less fit that self-interested types in uncooperative groups. We shall forego this more sophisticated model as it would unnecessarily complicate our exposition.

†† An alternative, perhaps more plausible, pair of assumptions is that c_r is expended only when non-cooperation is actually detected, and either the public goods game is multistage, as in the previous footnote, or 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.

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

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

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

Theorem 3. *Suppose the discount factor is δ^* in a good period and $\delta_*(< \delta^*)$ in a bad period, and bad periods occur with probability $p > 0$. Suppose eqn* (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 eqn* (8)*. Then if the fraction of strong reciprocators in cooperative groups is strictly positive* ($q_f > 0$), *eqn* (16) *is the condition for an increase in the fraction of strong reciprocators in the population.*

Let $\bar{f} = f_s(1 - q_f) + f_c q_f$, which is the frequency of strong reciprocity in the whole population. To close the model and thus determine the equilibrium value of $\bar{f}$, we must develop a plausible mechanism for the assignment of individuals to groups, thereby determining f_c and f_s as functions of $\bar{f}$. 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.‡‡

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.§§ The probability p_k that a newly formed group satisfies $f \geqslant f_*$ is then given by

$$p_k = \sum_{r \geqslant f_* k}^{k} \binom{k}{r} \bar{f}^r (1 - \bar{f})^{k-r}, \quad (17)$$

the frequency of strong reciprocators in such groups is given by

$$f_c = \frac{1}{k p_k} \sum_{r \geqslant f_* k}^{k} r \binom{k}{r} \bar{f}^r (1 - \bar{f})^{k-r}, \quad (18)$$

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

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

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

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 non-cooperators is sufficiently low, a small fraction $\bar{f}$ of strong reciprocators can always invade a population of self-interested agents.*

Theorems 2 and 4 suggest the central importance of the amount of harm h an agent can inflict on non-cooperators and the cost c_r the agent incurs doing so. As long as there is a positive fraction of strong reciprocators in the population, eqn (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

‡‡ 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 & Richerson (1993). Assortative interactions in our model take the form of groups with a high frequency of strong reciprocators lasting longer than other groups.

§§ For a more general analysis of this case, see Cohen & Eshel (1976). Note that the assumption of sampling with replacement assumes that population size is effectively infinite, so there are an infinite number of strong reciprocators as long as $\bar{f} > 0$.

population of self-interested agents.[‖‖] Under no condition, however, can strong reciprocators drive self-interested agents to extinction, since eqn (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 $\bar{f}^*$ of strong reciprocators in the population. In equilibrium, eqn (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 (l.h.s.) of eqn (16). The first term on the l.h.s. is a number greater than unity that, for plausible values of the parameters, lies between 1.0 and 4.0. For instance, if $\delta^* = 0.95$, $\delta_* \leqslant 0.25$, and $p \geqslant 0.10$, this factor has a minimum of 1.00 and a maximum of 2.47. The lower curve in Fig. 2 shows the equilibrium fraction $\bar{f}^*$ 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 40 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 Fig. 2 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.[¶¶]

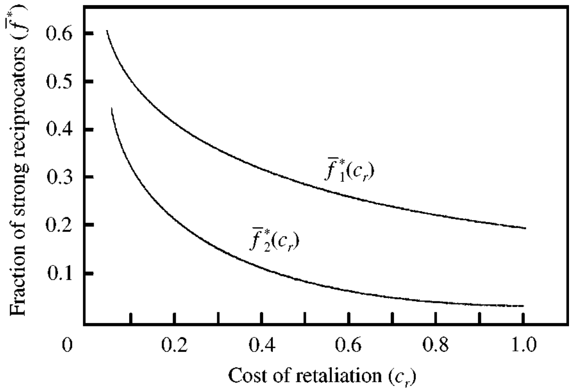

FIG. 2. The equilibrium fraction of strong reciprocating families: a computer simulation.

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*, in which much of the economic analysis is presumed.

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.

I have stressed the laboratory experiments in this paper because the controlled environment of the laboratory is conducive to isolating the strong reciprocity motive from other bases for cooperation and punishment. It would be remiss, however, not to mention the prevalence of strong reciprocity in the everyday operation of human society. Two categories of behavior immediately come to mind. First, in many circumstances people retaliate against others, at considerable personal cost, when the possibility of gains through future interaction is remote or zero. Victims of crime, for instance, spend time and energy ensuring that the perpetrators are apprehended

[‖‖] 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 behavior.

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

and receive harsh sentences, and jilted or betrayed lovers retaliate at great personal cost—often reducing their biological fitness to zero.

A second manifestation of strong reciprocity is evident in the propensity of humans to engage in episodic collective action towards transforming social norms and political regimes. Movements for civil rights, civil liberties, and political democracy in authoritarian states are responsible for creating modern liberal democracies, yet participation in such movements cannot usually be explained in terms of self-interest or reciprocal altruism (Bowles & Gintis, 1986). Non-participators in collective action, such as "scab workers" during a trade union strike, or "traitors" in a civil war, are spontaneously ostracized and punished, while guerrillas and underground freedom fighters are widely supported, often at great cost to the supporters. Without strong reciprocity, then, human society would likely be quite differently organized than it is, and we likely would be considerably less successful as a species.

Strong reciprocity is a form of altruism, in that it benefits group members at a cost to the strong reciprocators themselves. This paper 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.

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.

REFERENCES

ANDREONI, J. (1995). Cooperation in public goods experiments: kindness or confusion. *Am. Econ. Rev.* **85,** 891–904.

AXELROD, R. (1984). *The Evolution of Cooperation.* New York: Basic Books.

AXELROD, R. & HAMILTON, W. D. (1981). The evolution of cooperation. *Science* **211,** 1390–1396.

BELOVSKY, G. (1988). An optimal foraging-based model of hunter–gatherer population dynamics. *J. Anthropol. Archaeol.* 329–372.

BINGHAM, P. M. (1999). Human uniqueness: a general theory. *Quart. Rev. Biol.* **74,** 133–169.

BLURTON-JONES, N. G. (1987). Tolerated theft: suggestions about the ecology and evolution of sharing, hoarding, and scrounging. *Soc. Sci. Inf.* **26,** 31–54.

BOEHM, C. (1993). Egalitarian behavior and reverse dominance hierarchy. *Curr. Anthropol.* **34,** 227–254.

BOONE, J. L. & KAREN, L. K. (1999). More status or more children? Social status, fertility reduction, and long-term fitness. *Evol. Human Behav.* **20,** 257–277.

BOWLES, S. & GINTIS, H. (1986). *Democracy and Capitalism: Property, Community, and the Contradictions of Modern Social Thought.* New York: Basic Books.

BOYD, R. (1982). Density dependent mortality and the evolution of social behavior. *Animal Behav.* **30,** 972–982.

BOYD, R. & RICHERSON, P. J. (1985). *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

BOYD, R. & RICHERSON, P. J. (1988). The evolution of reciprocity in sizable groups. *J. theor. Biol.* **132,** 337–356.

BOYD, R. & RICHERSON, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizeable groups. *Ethol. Sociobiol.* **113,** 171–195.

BOYD, R. & RICHERSON, P. J. (1993). Effect of phenotypic variation of kin selection. *Proc. Nat. Acad. Sci. U.S.A.* **77,** 7506–7509.

CALVIN, W. H. (1983). A stone's throw and its launch window: timing precision and its implications for language and hominid brains. *J. theor. Biol.* **104,** 121–135.

CAMERER, C. & THALER, R. (1995). Ultimatums, dictators, and manners. *J. Econ. Perspect.* **9,** 209–219.

CAVALLI-SFORZA, LUIGI, L. & FELDMAN, M. W. (1981). *Cultural Transmission and Evolution.* Princeton, NJ: Princeton University Press.

COHEN, D. & ESHEL, I. (1976). On the founder effect and the evolution of altruistic traits. *Theor. Popul. Biol.* **10,** 276–302.

DARLINGTON, P. J. (1975). Group selection, altruism, reinforcement and throwing in human evolution. *Proc. Nat. Acad. Sci. U.S.A.* **72,** 3748–3752.

DAWES, R. M., ORBELL, J. M. & VAN DE KRAGT, J. C. (1986). Organizing groups for collective action. *Am. Pol. Sci. Rev.* **80,** 1171–1185.

DUGATKIN, L. A. (1997). *Cooperation among Animals.* New York: Oxford University Press.

DUGATKIN, L. A. (1998). Game theory and cooperation. In: *Game Theory and Animal Behavior* (Dugatkin, L. A. & Reeve, H. K., eds), pp. 38–63. Oxford: Oxford University Press.

DUGATKIN, L. A. & ALFIERI, M. (1992). Interpopulational difference in the cooperative strategy used during predator inspection in the guppy. *Evol. Ecol.* **6,** 519–526.

FEHR, E. & SCHMIDT, K. M. (1999). A theory of fairness, competition, and cooperation. *Quart. J. Econom.* **114,** 817–868.

FEHR, E. & GÄCHTER, S. (2000). Cooperation and punishment. *Am. Econom. Rev.* (in press).

FIFER, F. C. (1987). The adoption of bipedalism by the hominids: a new hypothesis. *Human Evol.* **2,** 135–147.

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

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

Goodall, J. (1964). Tool-using and aimed throwing in a community of free-living chimpanzees. *Nature* **201,** 1264–1266.

Güth, W. & Tietz, R. (1990). Ultimatum bargaining behavior: a survey and comparison of experimental results. *J. Econom. Psychol.* **11,** 417–449.

Hamilton, W. D. (1964). The genetical evolution of social behavior. *J. theor. Biol.* **37,** 1–16, 17–52.

Hamilton, W. D. (1975). Innate social aptitudes of man: an approach from evolutionary genetics. In: *Biosocial Anthropology* (Fox, R., ed.), pp. 115–132. New York: John Wiley & Sons.

Hawkes, K. (1993). Why hunter–gatherers work: an ancient version of the problem of public goods. *Curr. Anthropol.* **34,** 341–361.

Isaac, B. (1987). Throwing and human evolution. *African Archeol. Rev.* **5,** 3–17.

Kaplan, H. & Hill, K. (1985). Hunting ability and reproductive success among male ache foragers: preliminary results. *Curr. Anthropol.* **26,** 131–133.

Keckler, C. N. W. (1997). Catastrophic mortality in simulations of forager age-of-death: where did all the humans go? In: *Integrating Archaeological Demography*: *Multidisciplinary Approaches to Prehistoric Populations. Center for Archaeological Investigations*, *Occasional Papers*, *No.* 24. (Paine, R., ed.), pp. 205–228. Carbondale, IL: Southern Illinois University Press.

Knauft, B. (1991). Violence and sociality in human evolution. *Curr. Anthropol.* **32,** 391–428.

Ledyard, J. O. (1995). Public goods: a survey of experimental research. In: *The Handbook of Experimental Economics*. pp. 111–194. (Kagel, J. H. & Roth, A. E. eds), Princeton, NJ: Princeton University Press.

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

Robyn Dawes, M. & Thaler, R. (1988). Cooperation. *J. Econom. Perspect.* **2,** 187–197.

Michod, R. (1982). The theory of kin selection. *Ann. Rev. Ecol. Systems* **13,** 23–55.

Ostrom, E., Walker, J. & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *Am. Pol. Sci. Rev.* **86,** 404–417.

Plooij, F. X. (1978). Tool-using during chimpanzees bushpig hunt. *Carnivore* **1,** 103–106.

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

Roth, A. (1995). Bargaining experiments. In: *The Handbook of Experimental Economics* (Kagel, J. & Roth, A., eds). Princeton, NJ: Princeton University Press.

Sato, K. (1987). Distribution and the cost of maintaining common property resources. *J. Expl. Soc. Psychol.* **23,** 19–31.

Sethi, R. & Somanathan, E. (1996). The evolution of social norms in common property resource use. *Am. Econom. Rev.* **86,** 766–788.

Simon, H. A. (1993). Altruism and economics. *Am. Econom. Rev.* **83,** 156–161.

Sober, E. & Wilson, D. S. (1998). *Unto Others*: *The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Soltis, J., Boyd, R. & Richerson, P. (1995). Can group-functional behaviors evolve by cultural group selection: an empirical test. *Curr. Anthropol.* **36,** 473–483.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quart. Rev. Biol.* **46,** 35–57.

Wade, M. J. (1985). Soft selection, hard selection, kin selection and group selection. *Am. Nat.* **125,** 61–73.

Wilson, D. S. (1977). Structure demes and the evolution of group-advantageous traits. *Am. Nat.* **111,** 157–185.

Wilson, D. S. & Dugatkin, L. A. (1997). Group selection and assortative interactions. *Am. Nat.* **149,** 336–351.

Yamagishi, T. (1988a). The provision of a sanctioning system in the United States and Japan. *Soc. Psychol. Quart.* **51,** 265–271.

Yamagishi, T. (1988b). Seriousness of social dilemmas and the provision of a sanctioning system. *Soc. Psychol. Quart.* **51,** 32–42.

Yamagishi, T. (1992). Group size and the provision of a sanctioning system in a social dilemma. In: *Social Dilemmas*: *Theoretical Issues and Research Findings* (Liebrand, W. B. G., Messick, D. M. & Wilke, H. A. M., eds), pp. 267–287. Oxford: Pergamon Press.