ELSEVIER

Evolution and Human Behavior 24 (2003) 153–172

Evolution and Human Behavior

Explaining altruistic behavior in humans

Herbert Gintis[a,b,*], Samuel Bowles[a,b], Robert Boyd[c], Ernst Fehr[d]

[a]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[b]Department of Economics, University of Massachusetts, Amherst, MA 01003, USA
[c]Department of Anthropology, University of California at Los Angeles, 405 Hilgard Avenue, Box 951361 Los Angeles, CA 90095-1361, USA
[d]University of Zürich, Blümlisalpstrae 10 CH-8006 Zürich, Switzerland

Received 15 February 2002; received in revised form 29 November 2002

Abstract

Recent experimental research has revealed forms of human behavior involving interaction among unrelated individuals that have proven difficult to explain in terms of kin or reciprocal altruism. One such trait, *strong reciprocity* is a predisposition to cooperate with others and to punish those who violate the norms of cooperation, at personal cost, even when it is implausible to expect that these costs will be repaid. We present evidence supporting strong reciprocity as a schema for predicting and understanding altruism in humans. We show that under conditions plausibly characteristic of the early stages of human evolution, a small number of strong reciprocators could invade a population of self-regarding types, and strong reciprocity is an evolutionary stable strategy. Although most of the evidence we report is based on behavioral experiments, the same behaviors are regularly described in everyday life, for example, in wage setting by firms, tax compliance, and cooperation in the protection of local environmental public goods. © 2003 Elsevier Science Inc. All rights reserved.

Keywords: Altruism; Reciprocity; Experimental games; Evolution of cooperation

1. Introduction

The explanatory power of inclusive fitness theory and reciprocal altruism (Hamilton, 1964; Trivers, 1971; Williams, 1966) convinced a generation of researchers that what appears to be

[*] Corresponding author. Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.
E-mail address: hgintis@attbi.com (H. Gintis).

1090-5138/03/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S1090-5138(02)00157-5

altruism — personal sacrifice on behalf of others — is really just long-run self-interest. Richard Dawkins (1976, 1989), for instance, struck a responsive chord when, in *The Selfish Gene*, he confidently asserted "We are survival machines — robot vehicles blindly programmed to preserve the selfish molecules known as genes.. . .This gene selfishness will usually give rise to selfishness in individual behavior." Dawkins allows for morality in social life, but it must be socially imposed on a fundamentally selfish agent. "Let us try to *teach* generosity and altruism," he advises, "because we are born selfish." Yet, even social morality, according to R. D. Alexander, the most influential ethicist working in the William–Hamilton tradition, can only superficially transcend selfishness. In *The Biology of Moral Systems*, Alexander (1987, p. 3) asserts, "ethics, morality, human conduct, and the human psyche are to be understood only if societies are seen as collections of individuals seeking their own self-interest." In a similar state of explanatory euphoria, Ghiselin (1974, p. 247) claims "No hint of genuine charity ameliorates our vision of society, once sentimentalism has been laid aside. What passes for cooperation turns out to be a mixture of opportunism and exploitation. . .Scratch an altruist, and watch a hypocrite bleed."

However, recent experimental research has revealed forms of human behavior involving interaction among unrelated individuals that cannot be explained in terms of self-interest. One such trait, which we call *strong reciprocity* (Gintis, 2000b; Henrich et al., 2001), is a predisposition to cooperate with others and to punish those who violate the norms of cooperation, at personal cost, even when it is implausible to expect that these costs will be repaid either by others or at a later date.

In this article, we present empirical evidence supporting strong reciprocity as a schema for explaining important forms of altruism in humans. We then explain why, under conditions plausibly characteristic of the early stages of human evolution, a small fraction of strong reciprocators could invade a population of self-regarding types, and why strong reciprocity is an evolutionarily stable strategy. Although most of the evidence we report is based on behavioral experiments, the same behaviors are regularly observed in everyday life, for example in wage setting by firms (Bewley, 2000), tax compliance (Andreoni, Erard, & Feinstein, 1998), and cooperation in the protection of local environmental public goods (Acheson, 1988; Ostrom, 1990).

In supporting the importance of strong reciprocity, we of course do not deny the importance of either kin altruism (Hamilton, 1964) or reciprocal altruism (Trivers, 1971). Both are beyond doubt potent forces in human motivation. We do believe, however, that the evolutionary success of our species and the moral sentiments that have led people to value freedom, equality, and representative government are predicated upon strong reciprocity and related motivations that go beyond inclusive fitness and reciprocal altruism.

We wish to avoid three common misunderstandings of our argument. First, many contemporary researchers reject our critique of Dawkins, Alexander, and others in the "selfish gene" school by asserting that their pronouncements should not be taken at face value. Rather, they say, references to phenotypic behavior as "selfish" should be understood as asserting that the underlying genetic structures are subject to Darwinian evolutionary forces. Yet, these authors understood that their assertions were likely to be taken at face value, rather than being dramatic circumlocutions expressing completely unexceptionable propositions. It is only plausible, then, to suggest that they meant them, that it was plausible at the time to make such statements, but that they are now seen to be incorrect.

Second, we are often interpreted as rejecting the "gene-centered" approach to modeling human behavior. In fact, our results in no way contradict the standard population biology approach to genetic and cultural change. A gene that promotes self-sacrifice on behalf of others will die out unless those helped carry the mutant gene or otherwise promote its spread. In a population without structured social interactions of agents, behaviors of the type found in our experiments and depicted in our models could not have evolved. However, multilevel selection and gene–culture coevolutionary models support cooperative behavior among nonkin (Bowles, Choi, & Hopfensitz, in press; Feldman, Cavalli-Sforza, & Peck, 1985; Gintis, 2000, in press-a, in press-b; Henrich & Boyd, 2001; Sober & Wilson, 1998). These models, some of which are discussed below, are not vulnerable to the classic critiques of group selection by Dawkins (1976), Maynard Smith (1976), Rogers (1990), Williams (1966), and others.

Third, we are often told that the behavior we describe can in fact be explained by standard individual selection, kin selection, and reciprocal altruism models applied to the ancestral natural and social environment to which our species was subject during the period of its evolutionary emergence, where anonymous, one-shot interactions were supposedly extremely rare. Strong reciprocity in contemporary environments, according to this view, is a maladaption. We think this alternative is unlikely, and address the issues in Section 8.

2. Experimental evidence: Strong reciprocity in the labor market

In Fehr, Gächter, and Kirchsteiger (1997), the experimenters divided a group of 141 subjects (college students who had agreed to participate in order to earn money) into a set of "employers" and a larger set of "employees." The rules of the game are as follows. If an employer hires an employee who provides effort e and receives a wage w, they employer's payoff π is 100 times the effort e, minus the wage w that he must pay the employee ($\pi = 100w - e$), where the wage is between 0 and 100 ($0 \leq w \leq 100$) and the effort between 0.1 and 1 ($0.1 \leq e \leq 1$). The payoff u to the employee is then the wage he receives, minus a "cost of effort," $c(e)$ ($u - w - c(e)$). The cost of effort schedule $e(e)$ is constructed by the experimenters such that supplying effort $e = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 cost the employee $e(e) = 0$, 1, 2, 4, 6, 8, 10, 12, 15, and 18, respectively. All payoffs are converted into real money that the subjects are paid at the end of the experimental session.

The sequence of actions is as follows. The employer first offers a "contract" specifying a wage w and a desired amount of effort e^*. A contract is made with the first employee who agrees to these terms. An employer can make a contract (w,e^*) with at most one employee. The employee who agrees to these terms receives the wage w and supplies an effort level e, which need not equal the contracted effort e^*. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0,1,1]$, with impunity. Although subjects may play this game several times with different partners, each employer–employee interaction is a one-shot (nonrepeated) event. Moreover, the identity of the interacting partners is never revealed.

If employees are self-interested, they will choose the zero-cost effort level, $e - 0.1$, no matter what wage is offered them. Knowing this employers will never pay more than the minimum necessary to get the employee to accept a contract, which is 1 (assuming only integral wage offers are permitted). The employee will accept this offer, and will set $e=0.1$. Since $e(0)=0$, the employee's payoff is $u=1$. The employer's payoff is $\pi=0.1\times 100-1=9$.

In fact, however, this self-interested, outcome rarely occurred in this experiment. The average net payoff to employees was $u=35$, and the more generous the employer's wage offer to the employee, the higher the effort provided. In effect, employers presumed the strong reciprocity predispositions of the employees, making quite generous wage offers and receive higher effort, as a means to increase both their own and the employee's payoff, as depicted in Fig. 1. Similar results have been observed in Fehr, Kirchsteiger, and Riedl (1993, 1998).

Fig. 1 also shows that although most employees are strong reciprocators at any wage rate, there still is a significant gap between the amount of effort agreed upon and the amount actually delivered. This is not because there are a few "bad apples" among the set of employees, but because only 26% of employees delivered the level of effort they promised! We conclude that strong reciprocators are inclined to compromise their morality to some extent, just as we might expect from daily experience.

The above evidence is compatible with the notion that the employers are purely self-interested, since their beneficent behavior vis-à-vis their employees was effective in increasing employer profits. To see if employers are also strong reciprocators, following this round of experiments, the authors extended the game by allowing the employers to respond reciprocally to the *actual effort choices* of their workers. At a cost of 1, an employer could *increase* or *decrease* his employee's payoff by 2.5. If employers were self-interested, they

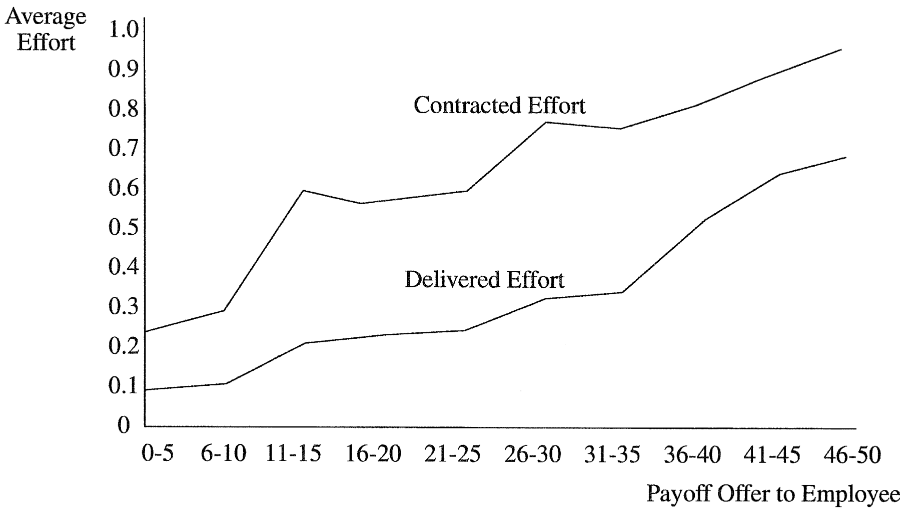

Fig. 1.

would of course do neither, since they would not interact with the same worker a second time. However, 68% of the time, employers punished employees who did not fulfill their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded 41% of employees who *exactly* fulfilled their contracts. Moreover, employees *expected* this behavior on the part of their employers, as shown by the fact that their effort levels *increased significantly* when their bosses gained the power to punish and reward them. Underfulfilling contracts dropped from 83% to 26% of the exchanges and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in the net payoffs to all subjects, even when the payoff reductions resulting from employer punishment of employees are taken into account. Several researchers have predicted this general behavior based on general real-life social observation and field studies, including Akerlof (1982), Blau (1964), and Homans (1961). The laboratory results show that this behavior has a motivational basis in strong reciprocity and not simply long-term material self-interest.

We conclude from this study that the subjects who assume the role of "employee" conform to internalized standards of reciprocity, even when they know there are no material repercussions from behaving in a self-interested manner. Moreover, subjects who assume the role of "employer" expect this behavior and are rewarded for acting accordingly. Finally, "employers" draw upon the internalized norm of rewarding good and punishing bad behavior when they are permitted to punish and "employees" expect this behavior and adjust their own effort levels accordingly.

3. Experimental evidence: The ultimatum game

In the ultimatum game, under conditions of anonymity, two players are shown a sum of money, say $10. One of the players, called the "proposer," is instructed to offer any number of dollars, from $1 to $10, to the second player, who is called the "responder." The proposer can make only one offer. The responder, again under conditions of anonymity, can either accept or reject this offer. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

Since the game is played only once and the players do not know each other's identity, a self-interested responder will accept any positive amount of money. Knowing this, a self-interested proposer will offer the minimum possible amount, $1, and this will be accepted. However, when actually played, *the self-interested outcome is never attained and never 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 generally being the modal offer) and respondents frequently reject offers below 30% (Camerer & Thaler, 1995; Güth & Tietz, 1990; Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991).

The ultimatum game has been played around the world, but mostly with university students. We find a great deal of individual variability. For instance, in all of the above experiments a significant fraction of subjects (about a quarter, typically) behave in a self-interested manner. But, among student subjects, average performance is strikingly uniform from country to country.

To expand the diversity of cultural and economic circumstances of experimental subjects, Henrich et al. (2001) undertook a large cross-cultural study of behavior in various games including the ultimatum game. Twelve experienced field researchers, working in 12 countries on four continents, recruited subjects from 15 small-scale societies exhibiting a wide variety of economic and cultural conditions. These societies consisted of three foraging groups (the Hadza of East Africa, the Au and Gnau of Papua New Guinea, and the Lamalera of Indonesia), six slash-and-burn horticulturists (the Aché, Machiguenga, Quichua, and Achuar of South America and the Tsimane and Orma of East Africa), four nomadic herding groups (the Turguud, Mongols, and Kazakhs of Central Asia, and the Sangu of East Africa), and two sedentary, small-scale agricultural societies (the Mapuche of South America and Zimbabwe farmers in Africa).

We can summarize our results as follows.

(a) The canonical model of self-interested behavior is not supported in *any* society studied. In the ultimatum game, for example, in all societies either respondents, or proposers, or both, behaved in a reciprocal manner.

(b) There is considerably more behavioral variability across groups than had been found in previous cross-cultural research. While mean ultimatum game offers in experiments with the student subjects are typically between 43% and 48%, the mean offers from proposers in our sample ranged from 26% to 58%. While modal ultimatum game offers are consistently 50% among university students, sample modes with these data ranged from 15% to 50%. In some groups, rejections were extremely rare, even in the presence of very low offers, while in others, rejection rates were substantial, including frequent rejections of *hyperfair* offers (i.e., offers above 50%). By contrast, the most common behavior for the Machiguenga was to offer zero. The mean offer was 22%. The Aché and Tsimane distributions resemble American distributions, but with very low rejection rates. The Orma and Huinca (non-Mapuche Chileans living among the Mapuche) have modal offers near the center of the distribution, but show secondary peaks at full cooperation.

(c) Differences among societies in “market integration” and “cooperation in production” explain a substantial portion of the behavioral variation between groups: The higher the degree of market integration and the higher the payoffs to cooperation, the greater the level of cooperation and sharing in experimental games. The societies were rank-ordered in five categories — “market integration” (how often do people buy and sell, or work for a wage), “cooperation in production” (is production collective or individual), plus “anonymity” (how prevalent are anonymous roles and transactions), “privacy” (how easily can people keep their activities secret), and “complexity” (how much centralized decision-making occurs above the level of the household). Using statistical regression analysis, only the first two characteristics, market integration and cooperation in production, were significant, and they together accounted for 66 of the variation among societies in mean ultimatum game offers.

(d) Individual-level economic and demographic variables did not explain behavior either within or across groups.

(e) The nature and degree of cooperation and punishment in the experiments was generally consistent with economic patterns of everyday life in these societies.

In a number of cases, the parallels between experimental game play and the structure of daily life were quite striking. Nor was this relationship lost on the subjects themselves. Here are some examples.

- The Orma immediately recognized that the public goods game was similar to the *harambee*, a locally initiated contribution that households make when a community decides to construct a road or school. They dubbed the experiment "the harambee game" and gave generously (mean 58% with 25% maximal contributors).
- Among the Au and Gnau, many proposers offered more than half the pie, and many of these "hyperfair" offers were rejected! This reflects the Melanesian culture of status-seeking through gift giving. Making a large gift is a bid for social dominance in everyday life in these societies, and rejecting the gift is a rejection of being subordinate.
- Among the whale hunting Lamalera, 63% of the proposers in the ultimatum game divided the pie equally and most of those who did not offered more than 50% (the mean offer was 57%). In real life, a large catch, always the product of cooperation among many individual whalers, is meticulously divided into predesignated parts and carefully distributed among the members of the community.
- Among the Aché, 79% of proposers offered either 40% or 50% and 16% offered more than 50%, with no rejected offers. In daily life, the Aché regularly share meat, which is being distributed equally among all other households, irrespective of which hunter made the kill.
- The Hadza, unlike the Aché, made low offers and high rejection rates in the ultimatum game. This reflects the tendency of these small-scale foragers to share meat, but a high level of conflict and frequent attempts of hunters to hide their catch from the group.
- Both the Machiguenga and Tsimane made low ultimatum game offers, and there were virtually no rejections. These groups exhibit little cooperation, exchange, or sharing beyond the family unit. Ethnographically, both show little fear of social sanctions and care little about "public opinion."
- The Mapuche's social relations are characterized by mutual suspicion, envy, and fear of being envied. This pattern is consistent with the Mapuche's postgame interviews in the ultimatum game. Mapuche proposers rarely claimed that their offers were influenced by fairness, but rather a by fear of rejection. Even proposers who made hyperfair offers claimed that they feared rare spiteful responders, who would be willing to reject even 50/50 offers.

4. Experimental evidence: The public goods game

The *public goods game* has been analyzed in a series of papers by the social psychologist Yamagishi (1986, 1988), by the political scientist Ostrom, Walker, and Gardner (1992), and by economists Fehr and Gächter (Fehr & Gächter, 2000, 2002; Gächter & Fehr, 1999). These researchers uniformly found the groups exhibit a much higher rate of cooperation than can be expected assuming the standard economic model of the self-interested actor, and this is especially the case when subjects are given the option of incurring a cost to themselves in order to punish free riders.

A typical public goods game consists of a number of rounds, say 10. The subjects are told the total number of rounds, as well as all other aspects of the game. The subjects are paid their winnings in real money at the end of the session. In each round, each subject is grouped with several other subjects — say three others — under conditions of strict anonymity. Each subject is then given a certain number of "points," say 20, redeemable at the end of the experimental session for real money. Each subject then places some fraction of his 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 some fraction, say 40%, of the total amount in the common account. Therefore, if a subject contributes his whole 20 points to the common account, each of the four group members will receive eight points at the end of the round. In effect, by putting the whole endowment into the common account, a player loses 12 points but the other three group members gain in total 24 ($=8 \times 3$) points. The players keep whatever is in their private account at the end of the round.

A self-interested player will contribute nothing to the common account. However, only a fraction of subjects in fact conform to the self-interest model. 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 10 rounds, until in the final rounds most players are behaving in a self-interested manner (Dawes & Thaler, 1988; Ledyard, 1995). In a metastudy of 12 public goods experiments, Fehr and 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 the end of the experiment approaches. However, 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 noncontributors, they do so at a cost to themselves (Dawes, Orbell, & Van de Kragt, 1986; Sato, 1987; Yamagishi, 1988a, 1988b, 1992). For instance, in Ostrom et al. (1992), subjects interacted for 25 periods in a public goods game, and paying a "fee," subjects could impose costs on other subjects but "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.

These studies 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 punisher. Fehr and Gächter (2000) set up an experimental situation in which *the possibility of strategic punishment was removed*. They used 6 and 10 round public goods games with group size of four and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the *partner* treatment, the four subjects remained in the same group for all 10 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 the same subject more than once. Subjects earned an average of about $35 for an experimental session.

Fehr and Gächter (2000) performed their experiment for 10 rounds with punishment and 10 rounds without (for additional experimental results and their analysis, see Bowles and Gintis, 2002; Fehr and Gächter, 2002). Their results are illustrated in Fig. 2. 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 in cooperation rates between the partner and the two stranger treatments is worth noting, because the strength of punishment is roughly the same across all treatments. This suggests that the credibility of the punishment threat is greater in the partner treatment because in this treatment the punished subjects are certain that, once they have been punished in previous

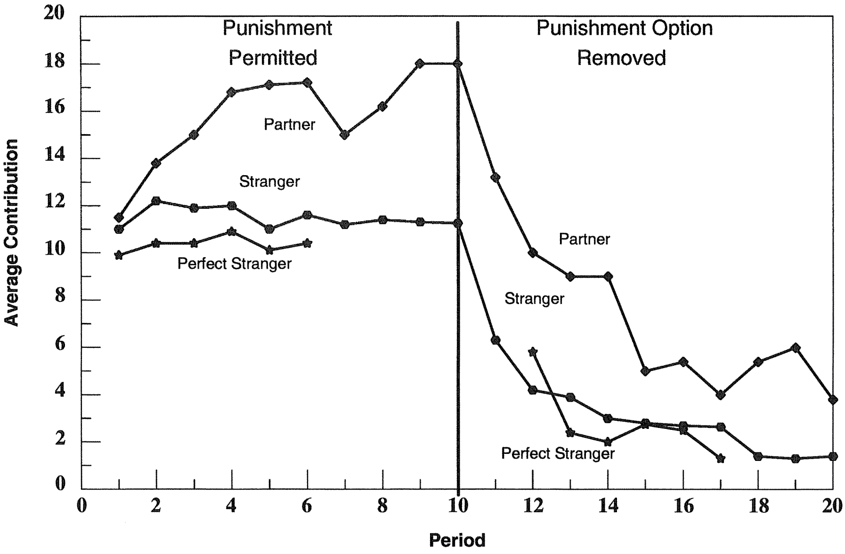

Fig. 2.

rounds, the punishing subjects are in their group. The prosociality impact of strong reciprocity on cooperation is thus more strongly manifested, the more coherent and permanent the group in question.

5. Experimental evidence: Intentions or outcomes?

One key fact missing from the above presentation is a specification of the relationship between contributing and punishing. The strong reciprocity interpretation suggests that high contributors will be high punishers, and punishees will be below-average contributors. This prediction is borne out in Fehr and Gächter (2002), wherein 75% of the punishment acts were executed by above-average contributors, and the most important variable in predicting how much one player punished another was the difference between the punisher's own contribution and the punishee's contribution.

Another key question in interpreting the above evidence is: Do reciprocators respond to fair or unfair *intentions* or do they respond to fair or unfair *outcomes*? The model of strong reciprocity unambiguously favors intentions over outcomes. To answer this question, Falk, Fehr, and Fischbacher (2002) ran two versions of the "moonlighting game" — an intention treatment (I-treatment), wherein a player's intentions could not be deduced from his actions, and a no-intention treatment (NI-treatment), wherein a player's intentions could. They provide clear and unambiguous evidence for the behavioral relevance of intentions in the domain of both negatively and positively reciprocal behavior.

The moonlighting game consists of two stages. At the beginning of the game, both players are endowed with 12 points. At the first stage, Player A chooses an action $a \in \{-6, -5, \ldots, 5, 6\}$. If A chooses $a > 0$, he gives Player B a tokens, while if he chooses $a < 0$, he takes away $|a|$ tokens from B. In case $a \geq 0$, the experimenter triples a so that B receives $3a$. After B observes a, he can choose an action $b \in \{-6, -5, \ldots, 17, 18\}$. If $b \geq 0$, B gives the amount b to A. If $b < 0$, B loses $|b|$, and A loses $|3b|$. Since A can give and take while B can reward or sanction, this game allows for both positively and negatively reciprocal behavior. Each subject plays the game only once.

If the B players are self-interested, they will all choose $b = 0$, neither rewarding nor punishing their A partners, since the game is played only once. Knowing this, if the A players are self-interested, they will all choose $a = -6$, which maximizes their payoff. In the I-treatment, A players are allowed to choose a, whereas in the NI-treatment, A players choice is determined by a roll of a pair of dice. If the players are not self-interested and care only about the fairness of the outcomes and not intentions, there will be no difference in the behavior of the B players across the I- and the NI-treatments. Moreover, if the A players believe their B partners care only about outcomes, their behavior will not differ across the two treatments. If the B players care only about the intentions of their A partners, they will never reward or punish in the NI-treatment, but they will reward partners who choose high a and punish partners who choose low a.

The experimenters' main result was that the behavior of Player B in the I-treatment is substantially different from the behavior in the NI-treatment, indicating that the attribution of fairness intentions is behaviorally important. Indeed, A players who gave to B players were generally rewarded by B players in the I-treatment much more that in NI-treatment (significant at the $P < .01$ level), and A players who took from B players were generally punished by B players in the I-treatment much more than in the NI-treatment (significant at the $P < .01$ level).

Turning to individual patterns of behavior, in the I-treatment, no agent behaved purely selfishly (i.e., no agent set $b - 0$ independent of a), whereas in the NI-treatment, 30 behaved purely selfishly. Conversely, in the I-treatment, 76 of subjects rewarded or sanctioned their partner, whereas in the NI-treatment, only 39 of subjects rewarded or sanctioned. We conclude that most agents are motivated by the intentionality of their partners, but a significant fraction care about the outcome, either exclusively or in addition to the intention of the partner.

6. The evolutionary stability of strong reciprocity

Gintis (2000b) developed an analytical model showing that under plausible conditions strong reciprocity can emerge from reciprocal altruism through group selection. The article models cooperation as a repeated n-person public goods game (see Section 4) in which, under normal conditions, if agents are sufficiently forward-looking, cooperation can be sustained by the threat of ostracism (Fudenberg & Maskin, 1986; Gintis, 2000a). However, when the group is threatened with extinction or dispersal, say through war, pestilence, or famine, cooperation is most needed for survival. During such critical periods, which were common in the evolutionary history of our species, future gains from cooperation become very uncertain, since the probability that the group will dissolve becomes high. The threat of ostracism then carries little weight, and cooperation cannot be maintained if agents are self-interested. Thus, precisely when a group is most in need of prosocial behavior, cooperation based on reciprocal altruism will collapse.

But a small number of strong reciprocators, who punish defectors whether or not it is in their long-term interest, can dramatically improve the survival chances of human groups. Moreover, among species that live in groups and recognizing individuals, humans are unique in their capacity to formulate and communicate rules of behavior and to inflict heavy punishment at low cost to the punisher (Bingham, 1999), as a result of their superior tool-making and hunting ability (Darlington, 1975; Fifer, 1987; Goodall, 1964; Isaac, 1987; Plooij, 1978). Under these conditions, strong reciprocators can invade a population of self-regarding types. This is because even if strong reciprocators form a small fraction of the population, at least occasionally they will form a sufficient fraction of a group that cooperation can be maintained in bad times. Such a group will then outcompete other self-interested groups, and the fraction of strong reciprocators will grow. This will continue until an equilibrium fraction of strong reciprocators is attained.

While the above results can be obtained analytically, there is no easily interpretable mathematical expression for the equilibrium fraction of strong reciprocators. A computer simulation, however, is quite revealing. For instance, suppose in good times a group has a

95 chance of surviving one period, while in bad times (which occur 1 period out of 10), the group only has a 25 chance of surviving. Then the lower curve in Fig. 3 shows the equilibrium fraction $\bar{f}*$ of strong reciprocators as the cost of retaliation (c_r) varies and there are 40 members per group. 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 transmittable within families. Note that a very small fraction of strong reciprocators can ensure cooperation, but the lower the cost of retaliation, the larger the equilibrium frequency of strong reciprocators.

This model highlights a key adaptive feature of strong reciprocity — its independence from the probability of future interactions — but it presumes that reciprocal altruism explains cooperation in normal times, when the probability of future interactions is high. However, reciprocal altruism does not work well in large groups (Boyd & Richerson, 1988; Choi, 2002; Joshi, 1987; Taylor, 1976). This is because when one withdraws cooperation in retaliation for the defection of a single group member, one inflicts punishment on all members, defectors and cooperators alike. The only evolutionarily stable strategy in the n-person public goods game is to cooperate as long as all others cooperate and to defect otherwise. For any payoff-monotonic dynamic, the basin of attraction of this equilibrium becomes very small as group size rises, so the formation of groups with a sufficient number of conditional cooperators is very unlikely and as a result, such an outcome may be easily disrupted by idiosyncratic play, imperfect information about the play of others, or other stochastic events. As a result, if group size is large, such an equilibrium is unlikely to be arrived at over reasonable historical time scales. Moreover, the only equilibrium is a "knife-edge" that collapses if just one member deviates.

To inject more realism in an evolutionary model of strong reciprocity, Henrich and Boyd (2001) developed a model in which norms for cooperation and punishment are acquired via payoff-biased transmission (imitate the successful) and conformist transmission (imitate high frequency behavior). They show that if two stages of punishment are permitted, then an arbitrarily small amount of conformist transmission will stabilize cooperative behavior by stabilizing punishment. They then explain how, once cooperation is stabilized in one group, it may spread through a multigroup population via cultural group selection. Once cooperation is

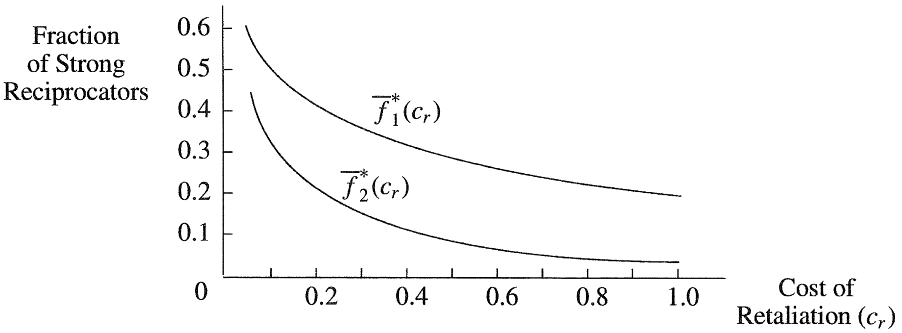

Fig. 3.

prevalent, they show how prosocial genes favoring cooperation and punishment may invade in the wake of cultural group selection, for instance, because such genes decrease an individual's chance of suffering costly punishment.

This analysis reveals a deep asymmetry between altruistic cooperation and altruistic punishment, explored further in Boyd, Gintis, Bowles, and Richerson (2002), who show that altruistic punishment allows cooperation in quite larger groups because the payoff disadvantage of altruistic cooperators relative to defectors is independent of the frequency of defectors in the population, while the cost disadvantage of those engaged in altruistic punishment declines as defectors become rare. Thus, when altruistic punishers are common, selection pressures operating against them are weak. The fact that punishers experience only a small disadvantage when defectors are rare means that weak within-group evolutionary forces, such as conformist transmission, can stabilize punishment and allow cooperation to persist. Computer simulations show that selection among groups leads to the evolution of altruistic punishment when it could not maintain altruistic cooperation.

7. The coevolution of institutions and behaviors

If group selection is part of the explanation of the evolutionary success of cooperative individual behaviors, then it is likely that group-level characteristics — such as relatively small group size, limited migration, or frequent intergroup conflicts — that enhance group selection pressures coevolved with cooperative behaviors. Thus, group-level characteristics and individual behaviors may have synergistic effects. This being the case, cooperation is based in part on the distinctive capacities of humans to construct institutional environments that limit within-group competition and reduce phenotypic variation within groups, thus heightening the relative importance of between-group competition, and hence allowing individually costly but in-group-beneficial behaviors to coevolve with these supporting environments through a process of interdemic group selection.

The idea that the suppression of within-group competition may be a strong influence on evolutionary dynamics has been widely recognized in eusocial insects and other species. Boehm (1982) and Eibl-Eibesfeldt (1982) first applied this reasoning to human evolution, exploring the role of culturally transmitted practices that reduce phenotypic variation within groups. Examples of such practices are leveling institutions, such as monogamy and food sharing among nonkin, namely those that reduce within-group differences in reproductive fitness or material well-being. By reducing within-group differences in individual success, such structures may have attenuated within-group genetic or cultural selection operating against individually costly but group-beneficial practices, thus giving the groups adopting them advantages in intergroup contests. Group-level institutions thus are constructed environments capable of imparting distinctive direction and pace to the process of biological evolution and cultural change. Hence, the evolutionary success of social institutions that reduce phenotypic variation within groups may be explained by the fact that they retard selection pressures working against in-group-beneficial individual traits and the fact that high frequencies of bearers of these traits reduces the likelihood of group extinctions.

We have modeled an evolutionary dynamic along these lines, exploring the possibility that intergroup contests play a decisive role in group-level selection. Our models assume that genetically and culturally transmitted individual behaviors, as well as culturally transmitted group-level characteristics, are subject to selection (Bowles, 2001; Bowles, Choi, & Hopfensitz, 2002). We show that intergroup conflicts may explain the evolutionary success of both (a) altruistic forms of human sociality towards nonkin and (b) group-level institutional structures such as food sharing and monogamy that have emerged and diffused repeatedly in a wide variety of ecologies during the course of human history. In-group-beneficial behaviors may evolve if (a) they inflict sufficient costs on out-group individuals and (b) group-level institutions limit the individual costs of these behaviors and thereby attenuate within-group selection against these behaviors.

Our simulations show that if group-level institutions implementing resource sharing or nonrandom pairing among group members are permitted to evolve, group-beneficial individual traits coevolve along with these institutions, even where the latter impose significant costs on the groups adopting them. These results hold for specifications in which cooperative individual behaviors and social institutions are initially absent in the population. In the absence of these group-level institutions, however, group-beneficial traits evolve only when intergroup conflicts are very frequent, groups are small, and migration rates are low. Thus, the evolutionary success of cooperative behaviors in the relevant environments during the first 90,000 years of anatomically modern human existence may have been a consequence of distinctive human capacities in social institution-building (Boyd & Richerson, in press).

8. Alternative explanations of strong reciprocity

We have argued that a variety of experimental data support our strong reciprocity model, and that strong reciprocity is adaptive in the sense of emerging from a gene–culture coevolutionary process that appears plausible in light of current and paleoanthropological evidence. The first of these claims is disputed by Price, Cosmides, and Tooby (2002), who present an alternative explanation of the data. The second claim is disputed by many who claim that altruistic cooperation and punishment in experimental games is a maladaptive response, arising from the fact that the experimental situations facing subjects in experimental games have no counterpart in either human evolutionary history or current everyday life, and hence humans have developed no adaptive genetic/cultural response to these situations. In its strongest form, this critique claims that the altruistic behavior we have described is thus of little importance in understanding human behavior in natural settings. We will deal with these two critiques in turn.

8.1. Strong reciprocity vs. reducing fitness deficits

Price et al. (2002) provide a model of self-interested behavior that they believe capable of explaining behavior in experimental games. Specifically, they assert (p. 221), “the motivation to punish free riders was designed for preventing the emergence of fitness advantages for free riders over contributors.''

We shall suggest that (a) the Price et al. model is not compatible with the existing experimental data and (b) the Price et al. model is not evolutionarily stable.

The Price et al. model asserts that punitive behavior is a response to payoff differentials, rather than to a breech of reciprocity norms. Experimenters have tested whether the desire to punish appears in response to payoff differentials, and the results have been almost uniformly negative in properly controlled settings. For a thorough discussion of this issue, see Falk, Fehr, and Fischbacher (2001).

For instance, suppose we restrict the proposer in an ultimatum game with a total pie of \$10 to offering the respondent either \$2 or \$8, and the respondent knows this restriction. Then, whereas in the unrestricted case many respondents will often reject the \$2 offer, in the restricted case almost all respondents accept the offer. This is not compatible with the notion that respondents act to reduce fitness differentials, for this is accomplished equally in the restricted and unrestricted cases. But the observed behavior is compatible with the notion that respondents act to punish ungenerous proposers, since in the restricted case, there is no way for the proposer to be generous without being extremely, indeed unreasonably, generous — in this case by keeping only \$2 for himself.

In another experiment (Blount, 1995), the offer in an ultimatum game was generated by a computer rather than the proposer, and this fact was known by the respondent. In this case, even very low offers were very rarely rejected. This is compatible with strong reciprocity, since the respondent has no incentive to punish a proposer who was not all responsible for the low offer. However, the Price et al. model predicts that low offers will be rejected whether generated by a computer or the proposer, since both give rise to the same fitness advantage of the proposer.

Indeed, more generally, in the standard ultimatum game (Section 3), a respondent who accepts any share of the pie less than a full 50 is incurring a relative fitness loss, yet almost all respondents accept offers of 40 of the pie. Similarly, in the employer–employee game (Section 2), employers voluntary shift resources to employees.

Turning to the question of evolutionary stability, the Price et al. model is based on the notion that selection should favor punitive sentiments because such sentiments would reduce the relative fitness advantage of defectors. As the authors assert, selection depends on relative, not absolute, fitness, so ''spiteful'' behavior that reduces the bearer's fitness can theoretically increase in frequency if it reduces the fitness of others by even more. But — and this is a *very important* but — the relevant others include not simply their targets, but others who cooperate enough to avoid punishment but never bear the costs of punishing. These ''stay out of trouble'' actors will have higher fitness than the punishers (or their targets) and will expand their share of the population at the expense of the other actors.

In anthropological terms, the relevant others will thus be the *deme* to which the individual belongs (the individuals who live and reproduce in the evolving population), not the individual's particular *social group* (the particular individuals with whom the individual interacts). For this reason, selection favors spiteful behavior only in very small populations in which the social group and the deme are of similar size (Hamilton, 1970) and in eusocial species (Foster, Wenseleers, & Ratnieks, 2001). The evidence for human foragers indicates that demes are many orders of magnitude greater in size than the social group, so spiteful behavior will be selected against.

Analytical models show that spiteful behavior can evolve if the members of the social group compete in a perfect zero-sum situation, so that one individual's increase in fitness necessarily means that the reproductive output of all other members of the social group will decrease so as to compensate exactly (Boyd, 1982). This also is unlikely, given our knowledge of the hunter–gatherer subsistence and demography. In such societies, resources are not localized, groups are mobile, and there is frequent movement of individuals between bands, so the zero-sum conditions are widely violated.

It is also possible that a subset of the social group compete in an approximately zero-sum sense for some limiting resource. Men, for instance, may compete for available women, or compete for relative status that can be converted into fitness. In this case, selection might favor costly behaviors that reduced the fitness of competitors. However, the "cognitive circuits" should then be structured so that individuals only invest in reducing the fitness of specific others with whom they have a particularly competitive zero-sum relationship. For example, men should care about other men of similar ages and who are close to them in the status hierarchy, and not about much younger or older men, of women. This situation does not apply to the case of a public goods game with egalitarian sharing of the rewards of cooperation.

8.2. Strong reciprocity as maladaptive

Some behavioral scientists have suggested that the behavior we have described in this article is maladaptive and not relevant to real-life social interactions. The human brain, they note, is not a general purpose information processor, but rather a set of interacting modular systems adapted to solving the particular problems faced by our species in its evolutionary history. Since the anonymous, nonrepeated interactions characteristic of experimental games were not a significant part of our evolutionary history, we could not expect subjects in experimental games to behave in a fitness-maximizing manner. Rather, we would expect subjects to confuse the experimental environment in more evolutionarily familiar terms as a nonanonymous, repeated interaction, and to maximize fitness with respect to this reinterpreted environment.

This critique, even if correct, would not lessen the importance of strong reciprocity in contemporary societies, to the extent that modern life leads individuals to face the frequent anonymous, nonrepeated interactions that are characteristic of modern societies with advanced trade, communication, and transportation technologies. Thus, even if strong reciprocity were a maladaptation, it could nevertheless be an important factor in explaining human cooperation today.

But we do not believe that this critique is correct. In fact, humans are well capable of distinguishing individuals with whom they are likely to have many future interactions, from others, with whom future interactions are less likely. Indeed, human subjects cooperate much more if they expect frequent future interactions than if future interactions are rare (Gächter & Falk, 2002; Keser & van Winden, 2000).

Humans with fine-tuned behavioral repertoires depending whether they face kin or nonkin, repeated or one-time interactors, and whether they can or cannot gain an individual reputation probably had an evolutionary advantage in our ancestral environment. The likely reason for this advantage is that humans faced many interactions where the probability of future interactions was sufficiently low to make defection worthwhile (Gintis, 2000b; Manson & Wrangham, 1991).

9. Conclusion

Much more experimental and theoretical work must be done to understand the major outlines of human prosocial behavior. We suspect, on the basis of the many studies completed over the past several years, that the new knowledge obtained will give us a picture of prosociality (and its obverse, antisociality) that is fundamentally incompatible with the economist's model of the self-interested actor and the biologists' model of the self-regarding reciprocal altruist.

Contemporary behavioral theory is the legacy of several major contributions (Cosmides & Tooby, 1992; Dawkins, 1989; Hamilton, 1964; Maynard Smith, 1982; Trivers, 1971; Williams, 1966; Wilson, 1975), all of which assumed the relations between nonkin could be modeled using self-interested actors. It is not surprising, then, that the most successful research in behavioral theory has been in the area of the family, kinship, and sexual relations, while the attempts to deal with the more complex interactions characteristics of social group behavior have been less persuasive. To address this situation, we believe that more attention should be paid to (a) the origin and nature of social emotions (including guilt, shame, empathy, ethnic identity, and ethnic hatred), (b) to the coevolution of genes and culture in human social history, (c) the role of group structure and group conflict in human evolution, and (d) integrating sociobiological insights into mainstream social sciences.

Acknowledgments

We would like to thank Martin Daly, Steve Frank, and Margo Wilson for helpful comments, and the Santa Fe Institute and John D. and Catherine T. MacArthur Foundation for financial support.

References

Acheson, J. (1988). *The lobster gangs of Maine*. Hanover, NH: New England Universities Press.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, *97*, 543–569.

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine.

Andreoni, J. (1995). Cooperation in public goods experiments: kindness or confusion. *American Economic Review*, *85*, 891–904.

Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, *36*, 818–860.

Bewley, T. F. (2000). *Why wages don't fall during a recession*. Cambridge: Harvard University Press.

Bingham, P. M. (1999). Human uniqueness: a general theory. *Quarterly Review of Biology*, *74*, 133–169.

Blau, P. (1964). *Exchange and power in social life*. New York: Wiley.

Blount, S. (1995). When social outcomes aren't fair. The effect of causal attributions on preferences. *Organizational Behavior & Human Decision Processes*, *63*, 131–144.

Boehm, C. (1982). The evolutionary development of morality as an effect of dominance behavior and conflict interference. *Journal of Social and Biological Structures*, *5*, 413–421.

Bowles, S. (2001). Individual interactions, group conflicts, and the evolution of preferences. In: S. N. Durlauf, & H. P. Young (Eds.), *Social dynamics* (pp. 155–190). Cambridge, MA: MIT Press.

Bowles, S., Choi, J., & Hopfensitz, A. (in press). *The co-evolution of individual behaviors and social institutions*. Santa Fe, NM: Santa Fe Institute.

Bowles, S., & Gintis, H. (2002). Homo reciprocans. *Nature*, *415*, 125–128.

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

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2002). *Altruistic punishment in large groups evolves by interdemic group selection.* Unpublished data.

Boyd, R., & Richerson, P. (in press). *The nature of cultures.* Department of Anthropology, UCLA, Los Angeles, CA.

Boyd, R., & Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, *132*, 337–356.

Camerer, C., & Thaler, R. (1995). Ultimatums, dictators, and manners. *Journal of Economic Perspectives*, *9*, 209–219.

Choi, J.-K. (2002). *Three essays on the evolution of cooperation*. Amherst, MA: University of Massachusetts.

Cosmides, L., & Tooby, J. (1992). The psychological foundations of culture. In: J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

Darlington, P. J. (1975). Group selection, altruism, reinforcement and throwing in human evolution. *Proceedings of the National Academy of Sciences, U. S. A.*, *72*, 3748–3752.

Dawes, R. M., Orbell, J. M., & Van de Kragt, J. C. (1986). Organizing groups for collective action. *American Political Science Review*, *80*, 1171–1185.

Dawes, R. M., & Thaler, R. (1988). Cooperation. *Journal of Economic Perspectives*, *2*, 187–197.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.

Dawkins, R. (1989). *The selfish gene* (2nd ed.). Oxford: Oxford University Press.

Eibl-Eibesfeldt, I. (1982). Warfare, man's indoctrinability and group selection. *Journal of Comparative Ethnology*, *60*, 177–198.

Falk, A., Fehr, E., & Fischbacher, U. (2001). *Driving forces of informal sanctions.* Working Paper No. 59, Institute for Empirical Research in Economics.

Falk, A., Fehr, E, & Fischbacher, U. (2002). *Testing theories of fairness and reciprocity—intentions matter.* Zürich: University of Zürich.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment. *American Economic Review*, *90*, 980–994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: experimental evidence. *Econometrica*, *65*, 833–860.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? *Quarterly Journal of Economics*, *108*, 437–459.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, *42*, 1–34.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

Feldman, M. W., Cavalli-Sforza, L. L., & Peck, J. R. (1985). Gene–culture coevolution: models for the evolution of altruism with cultural transmission. *Proceedings of the National Academy of Sciences, U. S. A.*, *82*, 5814–5818.

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

Foster, K. R., Wenseleers, T., & Ratnieks, F. I. W. (2001). Spite: Hamilton's unproven theory. *Annales Zoologici Fennici*, *38*, 229–238.

Fudenberg, D., & Maskin, F. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*, 533–554.

Gächter, S., & Falk, A. (2002). Reputation or reciprocity? Consequences for labour relations. *Scandinavian Journal of Economics*, *104*, 1–25.

Gächter, S., & Fehr, E. (1999). Collective action as a social exchange. *Journal of Economic Behavior and Organization*, *39*, 341–369.

Ghiselin, M. T. (1974). *The economy of nature and the evolution of sex*. Berkeley, CA: University of California Press.

Gintis, H. (2000a). *Game theory evolving*. Princeton, NJ: Princeton University Press.

Gintis, H. (2000b). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 169–179.

Gintis, H. (in press-a). The hitchhiker's guide to altruism: genes and culture, and the internalization of norms. *Journal of Theoretical Biology.*

Gintis, H. (in press-b). The puzzle of human prosociality. *Rationality and Society, 15.*

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. *Journal of Economic Psychology*, *11*, 417–449.

Hamilton, W. D. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology*, *37*, 1–52.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature*, *228*, 1218–1220.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*, 79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElrcath, R. (2001). Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review*, *91*, 73–78.

Homans, G. (1961). *Social behavior: its elementary forms*. New York: Harcourt Brace.

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

Joshi, N. V. (1987). Evolution of cooperation by reciprocation within structured demes. *Journal of Genetics*, *66*, 69–84.

Keser, C., & van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, *102*, 23–39.

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

Manson, J. H., & Wrangham, R. W. (1991). Intergroup aggression in chimpanzees. *Current Anthropology*, *32*, 369–390.

Maynard Smith, J. (1976). Group selection. *Quarterly Review of Biology*, *51*, 277–283.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.

Ostrom, E. (1990). *Governing the commons: the evolution of institutions for collective action*. Cambridge, UK: Cambridge University Press.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, *86*, 404–417.

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

Price, M., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution & Human Behavior*, *23*, 203–231.

Rogers, A. R. (1990). Group selection by selective emigration: the effects of migration and kin structure. *American Naturalist*, *135*, 398–413.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ijubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review*, *81*, 1068–1095.

Sato, K. (1987). Distribution and the cost of maintaining common property resources. *Journal of Experimental Social Psychology*, *23*, 19–31.

Sober, E., & Wilson, D. S. (1998). *Unto others: the evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Taylor, M. (1976). *Anarchy and cooperation*. London: Wiley.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

Williams, G. C. (1966). *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton, NJ: Princeton University Press.

Wilson, E. O. (1975). *Sociobiology: the new synthesis*. Cambridge, MA: Harvard University Press.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*, 110–116.

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

Yamagishi, T. (1988b). Seriousness of social dilemmas and the provision of a sanctioning system. *Social Psychology Quarterly*, *51*, 32–42.

Yamagishi, T. (1992). Group size and the provision of a sanctioning system in a social dilemma. In: W. Liebrand, D. M. Messick, & H. Wilke (Eds.), *Social dilemmas: theoretical issues and research findings* (pp. 267–287). Oxford: Pergamon.