Friedman, Daniel; Singh, Nirvikar

Working Paper

Negative Reciprocity: The Coevolution of Memes and Genes

Working Paper, No. 560

Provided in Cooperation with:
University of California Santa Cruz, Economics Department

Suggested Citation: Friedman, Daniel; Singh, Nirvikar (2003) : Negative Reciprocity: The Coevolution of Memes and Genes, Working Paper, No. 560, University of California, Economics Department, Santa Cruz, CA

This Version is available at:
https://hdl.handle.net/10419/83840

Negative Reciprocity: The Coevolution of Memes and Genes

By Daniel Friedman and Nirvikar Singh
Economics Department
University of California, Santa Cruz
Revised December 2003

Send correspondence to:
Professor Nirvikar Singh,
Department of Economics, Social Sciences 1,
University of California at Santa Cruz, Santa Cruz CA 95064, USA
Email: boxjenk@ucsc.edu
Phone: 831.459.4093
Fax: 831.459.5900

Running Title: Negative Reciprocity and Coevolution

Abstract

A preference for negative reciprocity is an important part of the human emotional repertoire. We model its role in sustaining cooperative behavior but highlight an intrinsic free-rider problem: the fitness benefits of negative reciprocity are dispersed throughout the entire group, but the fitness costs are borne personally. Evolutionary forces tend to unravel people's willingness to bear the personal cost of punishing culprits. In our model, the countervailing force that sustains negative reciprocity is a meme consisting of a group norm together with low-powered (and low-cost) group enforcement of the norm. The main result is that such memes coevolve with personal tastes and capacities so as to produce the optimal level of negative reciprocity.

Key words

Altruism, reciprocity, negative reciprocity, coevolution

Introduction

Negative reciprocity is the human act of harming those who wrong us. Typically it is accompanied by powerful emotions of anger that urge us to harm the culprit even at some personal cost to ourselves. For example, Ahab's fictional pursuit of Moby Dick, the great white whale, cost him his life. Many non-fictional people in the Middle East, the Balkans and elsewhere have lost their lives and ruined their countries in pursuing negative reciprocity. We all have personal experience with negative reciprocity, if only in the realm of office politics.

Negative reciprocity is not always a bad thing. It complements positive reciprocity, the desire to help others who have helped us. The folk theorem of game theory, which applies to repeated interactions, explains positive reciprocity as an individually rational (indeed, a subgame perfect) way to support efficient exchange of favors, as long as the discount factor exceeds the ratio of personal cost to social benefit. As explained below, negative reciprocity can further increase social value in two ways. It can support efficient exchange even when the discount factor is low, as for example when repeat interaction is sporadic. Moreover, it can deter opportunistic behavior that would otherwise undermine positive reciprocity.

As far as existence is concerned, it is beside the point whether negative reciprocity is helpful or harmful to society. The crucial theoretical issue from an evolutionary perspective is whether vengeful traits convey a selective advantage. Apparently the answer is “no,” because negative reciprocity results in a net fitness loss. We will show, in a stylized analysis that captures the essence of cooperation dilemmas, that negative reciprocity is weakly dominated by (i.e., never yields a higher payoff than) otherwise similar behavior that shirks on the personal cost. Therefore it is a theoretical puzzle how negative reciprocity ever established itself in the repertoire of human motives, and how it sustains itself. Until the puzzle is solved, theory will offer no guidance on how negative reciprocity might be regulated to increase its social value and to reduce its devastation.

In this paper we offer an evolutionary account of negative reciprocity in humans. Our definition restricts reciprocity, positive or negative, to social creatures that have the capacities to identify and recall the earlier behavior of specific individuals, and to reward or punish them contingent on earlier behavior. The account we offer also requires cultural transmission of codes of behavior. We do not explore the extent to which our model might apply to non-human species with these capacities.

Our account draws on the perspectives of both selfish genes and cultural memes. Dawkins (1982) defines a meme as “the unit of information that is conveyed from one brain to another during cultural transmission;” examples from Durham (1991, pp. 188-90) range from marriage customs to words for colors. Our concern is with memes that pertain to the group rather than to an individual, such as the routines and norms within a business corporation (Nelson and Winter, 1982). For general discussions of memes and

social transmission mechanisms, see Dawkins (1976), Blackmore (1999, 2000) and comments on the latter by Boyd and Richerson, Lee Alan Dugatkin, and Henry Plotkin.

Our account of negative reciprocity starts with a standard normal form game that captures, simply and directly, the idea of a personal cost incurred to reap social gains. The game illustrates how a preference for negative reciprocity realigns incentives and supports a socially efficient equilibrium, but demonstrates that negative reciprocity is itself evolutionarily problematic.

After noting several earlier treatments of the problem in Section 2, we propose an evolutionary model with individual learning and evolution as well as meme selection for groups of individuals. In section 3 we argue that groups of individuals can use low-cost sanctions (or simply status changes) to enforce a particular norm on the proper degree of negative reciprocity. Section 4 assembles the elements of a simple model, and Section 5 derives the main results. Actual behavior typically will fall short of the norm, but selection across groups will adjust the norm so that actual behavior maximizes the fitness of group members, and the free rider problem is overcome. Following a concluding discussion, an Appendix shows that the main conclusions survive the relaxation of many simplifying assumptions.

1. The Underlying Game

We begin by demonstrating how a preference for negative reciprocity can convert a standard prisoner's dilemma problem to a simple coordination problem with a Pareto efficient equilibrium. The idea is that, given a motive for negative reciprocity, cooperative behavior is no longer is dominated and can become part of a Nash Equilibrium (NE), even when there is no repeat interaction. Our subsequent analysis builds on this game, which captures in simple terms the conflict between social efficiency and individual self-interest.

The basic underlying game is a symmetric 2-player prisoner's dilemma with a cooperator payoff of 1, a temptation payoff of 2, a sucker payoff of -1, and an all-defect payoff of 0. In other words, the benefits of full cooperation of 2 are evenly split and the benefit of one-sided cooperation of 1 is very unevenly split at (2, -1), relative to the no cooperation payoff, which is normalized to (0, 0).

[Table 1 about here]

Payoffs so far are material, and describe both fitness and utility. The social dilemma is that there is a personal cost (or personal fitness reduction) of one unit to choosing the cooperative strategy, but it produces a social gain (or increase in the fitness sum), also of one unit. The game has a unique Nash equilibrium in which each player chooses the dominant strategy D and achieves fitness 0. The choices of specific payoffs are intended only to simplify the algebra and exposition; essentially the same results hold

for other fitness payoffs satisfying the usual Prisoner's dilemma inequalities: temptation > cooperation > all defect > sucker; and temptation + sucker < 2 × cooperation.

To this underlying game we add a punishment technology and a punishment motive with parameter v, which (as we shall soon see) is the incurred cost. We hypothesize that a player can inflict harm (fitness loss) h on the other player at personal fitness cost ch. The marginal cost c is a constant parameter between 0 and 1 that captures the technological opportunities for punishing others. We also hypothesize that inflicting harm h yields the player a utility bonus of $v \ln h$ (but no fitness bonus) when he is the victim of the sucker payoff and no bonus in other circumstances. Thus the motive is not spite (e.g., Hirshleifer, 1987; Levine, 1998), but rather is revenge for damage personally experienced, and so the action taken by the victim involves negative reciprocity. The motivational parameter v is subject to evolutionary forces and is intended to capture an individual's temperament, e.g., his susceptibility to anger (Hirshleifer, 1987; Frank, 1988).

The objective function for the victim of a sucker payoff with motivational parameter v is therefore $v \ln h - ch - 1$. The utility-maximizing degree of negative reciprocity h^* to inflict on a culprit (i.e., the defector and the beneficiary of the temptation payoff) is the unique solution of the first order condition $0 = v/h - c$, so $h^* = v/c$ is the inflicted damage. Hence $ch^* = v$, and the motivational parameter also becomes the incurred cost. Utility in this case is $v \ln v/c - v - 1$, while fitness is just $-v-1$. The game now has the same fitness payoffs as before on the main diagonal, but the sucker payoff is reduced by the cost of negative reciprocity, and the temptation payoff is reduced by the amount of harm inflicted, as in Table 2.

[Table 2 about here]

For $v > c$, the transformed game no longer has D as a dominant strategy. When population fraction s plays C, the expected fitness of C is $W(C)=1s - (1+v)(1-s)$ and the expected fitness of D is $W(D)=(2-v/c)s$. The two expressions are equal at $s^* = (1+1/v)/(1+1/c)$. For $s < s^*$ the expected utility is higher for D and we can expect cooperation to disappear as play converges to the inefficient (fitness 0) all-D equilibrium, as in the basic game. But for $s > s^*$ the expected utility is higher for C and we can expect negative reciprocity to drive out defection, resulting in the Pareto efficient all-C equilibrium. Thus for $v > c$ we have a coordination game which has two locally stable pure Nash equilibria and an unstable mixed Nash equilibrium at $s^* < 1$, as illustrated in Figure 1. These statements are true under any plausible evolutionary dynamics, in particular, compatible or monotone dynamics (Friedman, 1991; Weibull, 1995).

[Figure 1 about here]

Note that efficient all-C behavior can also be sustained as a repeated game Nash equilibrium even in the original (v=0) version if culprits can be detected and identified, and if all players have a discount factors that exceed 0.5. One uses standard tit-for-tat or similar punishment strategies. But it may well be the case that repeat meetings are

infrequent or culprits are hard to track, so the discount factor is too small to sustain the efficient outcome. Thus the anticipation of negative reciprocity can support efficient social outcomes that cannot be sustained by standard repeated game strategies.

2. The Viability Problem

There is a gap in the argument so far. The motivational parameter v itself is subject to evolutionary forces, perhaps slower forces than those determining the prevalence s of cooperation, but real forces nonetheless. Recall that the expected fitness of a cooperator is $W(C|\, s,\, v) = 2s - 1 - v(1-s)$, which is a strictly decreasing function of v for any fixed $s < 1$. Only when there are no culprits left to punish at $s = 1$ is the expected fitness independent of v. Assuming that players occasionally encounter culprits (an assumption we shall develop later), player v' is fitter than player v whenever $0 < v' < v$. Therefore the parameter v will be driven towards 0 under any plausible evolutionary dynamics. We have a variant of the classic free rider or chiseling problem, and it seems that negative reciprocity is not viable.

Existing literature offers several possible avenues for escaping the viability problem. Prominent among them is inclusive fitness (Haldane, 1955; Hamilton, 1964). The viability problem is attenuated for social creatures that interact with close genetic relatives, such as slime molds (index of relatedness $r = 1- \varepsilon$) or ants and bees (r up to 0.75). But we are interested in humans, who typically interact with others who are not necessarily closely related (say on average $r = 0$ to .25). Hence for our purposes this avenue is unpromising.

Friedman and Singh (1999, 2004) discuss a variety of other proposed avenues. Some—weakened notions of evolutionary stability, and mutation constraints that preclude intermediate levels of the trait, or that chain the trait to some adaptive trait—play no role in the subsequent analysis. Other proposed avenues, however, relate to our proposed solution. First, perhaps individuals with higher values of v encounter D play less frequently (e.g., Frank, 1987). Harrington (1989) points out the importance of observability; we shall focus on observability at the group level rather than at the individual level. Second, the personal cost of negative reciprocity – c in our model – might be zero, or even negative if looting is possible or in some forms of repeated play, e.g., Rosenthal (1996), Guttman (2003). We shall focus on once-off encounters outside the group, where c is positive, but we also consider low cost technologies for disciplining members within a group.

Third, one can impose some sort of group selection. The idea goes back at least to Darwin (1871): “A tribe including many members who...were always ready to aid one another, and to sacrifice themselves for the common good would be victorious over most other tribes; and this would be natural selection.” The idea has proved very controversial (e.g., Wynne-Edwards, 1962; Trivers, 1985; Alexander, 1987; Sober and Wilson, 1998). Our focus on group traits is related to recent work on cultural group selection (e.g., Boyd,

et al., 2003; Gintis et al., 2003). Finally, one can consider higher order punishments (punish those who don't punish D players, and punish those who fail to do so, etc.; e.g., Henrich and Boyd, 2001) or third party punishers (e.g., Sugden, 1986; Nowak and Sigmund, 1998; but see also Leimar and Hammerstein, 2000). Neither punishment in itself solves the viability problem for encounters outside the group, but both punishments reinforce our view of enforcement within the group.

3. Group Structure

How do humans overcome the viability problem? Our core idea is that groups discipline their members. During the vast majority of its evolutionary history, *Homo sapiens*, like other social primates, presumably lived in small groups of individuals who interacted with other individuals within the group on a daily basis. Within the group, everyone knows everyone else, and several devices are available to enforce the all-C equilibrium. Tit for tat and related repeated game strategies work well because repeat interaction is reliable and frequent (e.g., Sethi and Somanathan, 2003); third party and higher order punishment strategies become feasible; and reputations for vengeful behavior can be established with one's fellow group members. While these devices for disciplining behavior are not perfect, they do suggest that D behavior will be relatively rare within well-functioning groups.

How about interactions with individuals in other groups? Depending on the setting, a member of a given group may encounter a specific member of another group only very sporadically but, aggregating across all other groups and their members, such encounters could lead to significant fitness differences (Black-Michaud, 1975; O'Kelley and Carney, 1986; Fehr and Henrich, 2003). An individual who somehow could induce strangers to play C would do much better than one who (correctly or incorrectly) anticipates D play. Unfortunately, an individual in a cross-group encounter cannot reliably signal her true v because outward signs can be mimicked at low cost, and neither (due to the large numbers of sporadic personal encounters) can she easily establish a personal reputation for her true v. It is much more plausible that her group can establish a reputation, and that reputation would determine the outcome of the interaction. For example, if one of the authors met a stranger on a train in India, the stranger might try to ascertain the author's family village and his last name, as ways of assigning him to a group with a particular reputation. The questioner is likely to find such information more useful than personal details, which are easier to disguise.

Our concern here is with the social norms maintained by the group, the enforcement of norms, and their evolution. All known groups of humans maintain social norms that prescribe appropriate behavior towards fellow group members, and typically prescribe different appropriate behavior towards individuals outside the group (Sober and Wilson, 1998). For example, Nisbett and Cohen's (1996) "culture of honor" prescribes that a person responds with violence or the threat of violence to any insult or perceived affront. Nisbett and Cohen study only the American South, but their findings align well

with the anthropological literature on many other groups in the Mediterranean (Black-Michaud, 1975; Gilmore, 1990; Peristiany, 1965), Africa (Galaty & Bonte, 1991), North America (Lowie, 1954; Farb, 1978) and India (Pettigrew, 1975). Pettigrew describes the culture of honor for North India's Jats (herders, originally from Central Asia, who have become settled farmers over time) as follows:

> Relationships of extreme friendship and hostility between families were actively involved with the philosophy of life embodied in the concept of *izzat* -- the complex of values regarding what was honourable. … That aspect of *izzat* according to which the relationships between families were supposed to be ordered emphasized the principle of equivalence in all things, i.e., not only equality in giving but also equality in negative reciprocity. *Izzat* was in fact the principle of reciprocity of gifts, plus the rule of an eye for an eye and a tooth for a tooth … *Izzat* enjoined aid to those who had helped one. It also enjoined that revenge be exacted for personal insults and damage to person or property. (p. 58)

How might a group enforce a social norm like *izzat*? The vengeance technology already introduced could, of course, be used to punish norm violators within the group. But groups have at least two other, lower cost punishment technologies not available to individuals. First, members may choose to interact less frequently with norm violators, i.e., partial shunning. Norm violation may lead group members to regard the violator as less reliable, and therefore they will often prefer (and believe it to be in their material interest) to choose an alternative partner. Shunning reduces the overall fitness in the group because some opportunities for mutual gains are not fully realized. But the cost falls mainly on the violator, because the shunner can find the next best alternative partner.

Second, and even lower cost, the group may lower the status of a norm violator. Of course, status generally depends on individual traits of all sorts, including age, sex, height, strength, birth order and parental status. But it is reasonable to postulate that, other things equal, an individual will have higher status when his behavior better upholds the group's norms (again see Nisbett & Cohen 1996). Status matters because it affects resource allocation. The group allocates many resources; depending on the context, these might include marriage partners, home sites, and access to fishing holes or plots of land. Status is a device for selecting among the numerous allocation equilibria: the higher status individuals get the first choice on available home sites, desirable marriage partners tend to prefer higher status suitors, etc. (e.g., see MacDonald, 1994, on Jewish society in 13th century Spain, or Nisbett and Cohen, 1996, on the American South, past and present). The model introduced below uses a single parameter, a, to measure the sensitivity of fitness to status combined with the sensitivity of status to behavior.

Enforcement could affect the fitness of nondeviators as well as deviators. Indeed, since status is relative, a decrease in one individual's status will increase the status of some other group members and hence increase their fitness. Catanzaro (1992) makes precisely this point regarding the Sicilian Mafia: "... the men who usurped honor did so at the expense of others who stood to lose it to the same degree... Ultimately, honor has been described as a system of stratification [by Davis, 1980] ..." (pp. 46-47).

The combination of a group's relevant social norms and their enforcement devices is referred to below as the group's meme. The meme pertains to the group rather than to its individual members. For example, the membership of a street gang might turn over two or three times during a decade but its meme (e.g., its dress style, graffiti logos, and combat codes of conduct) could remain constant. Conversely, the group's meme could evolve with constant membership via mechanisms ranging from imitating more successful groups to conquest.

How do group memes evolve? We will assume that a given meme becomes more prevalent when it brings higher average fitness to its group members than do alternative memes. Such monotone dynamics are consistent with many specific mechanisms of meme preservation and transmission, which can include various kinds of communication and reinforcement behavior within the group as described in Nisbett and Cohen (1996 pp. 2, 86, 93), Weingart et al (1997), Durham (1991) and Boyd and Richerson (1990). We do make no sweeping claim (as do sociobiologists such as Wilson, 1980) that genes always hold memes on a "short leash" that allows only minor short-run deviations from genetic fitness. Our assumption is simply that the short leash is a reasonable approximation in the present case, group norms concerning negative reciprocity.

4. Elements of the Model

We now specify elements of a model in which group memes for negative reciprocity co-evolve with individual characteristics. A complete specification of a group's meme would include prescriptions for proper behavior towards culprits and cooperators within the group, and possibly different behavior towards culprits and cooperators outside the group, together with enforcement devices. We have already noted that the group has many available devices for ensuring good levels of cooperation within the group, and cooperation outside the group is not at issue. Our focus is the prescription for outgroup culprits and the enforcement of the prescription.

Hence we summarize the relevant memes using two parameters: v^n for the group's normative level of negative reciprocity outside the group, and a for the rigor with which the group enforces that norm. For example, *Izzat* applied to the basic game calls for $h = 2$, since the culprit causes a loss of 2 (relative to the cooperative outcome of 1) and therefore rather strict enforcement of the norm $v^n = 2c$ is enjoined.

Enforcement is modeled by a loss function $\rho(x)$, where $x = v^n - v$ is the deviation of an individual's vengeful behavior from the group norm. The group imposes an expected fitness loss ρ on a deviator by lowering that individual's status or reputation within the group. The idea is that the deviation sometimes will be observed by another member of the group and gossip will spread the news. The simplest possible quadratic specification is $\rho(x; a) = x^2/(2a)$, where enforcement is more rigorous the smaller is the parameter $a>0$. Recall that norm enforcement may also affect the fitness of nondeviators. Let R denote the fitness increment (zero or negative) an individual receives due to the

deviations of other group members from the normative level v^n. In the special case of enforcement by changes in relative status, R will exactly offset the loss associated with the enforcement function, ρ.

The other side of the co-evolution model specifies the individual traits. Each individual is characterized by two parameters: his actual negative reciprocity level v, and the maximum possible value v^{max} that any meme could induce. The *capacity* for feeling anger and expressing it by damaging others as summarized in v^{max} may well be genetically transmitted, but the actual v of an individual probably is best regarded as learned from personal experience.

A few remarks are in order about fitness, monotone dynamics and time scales. We shall assume that individual levels of v adjust rapidly within $[0, v^{max}]$; the idea is that people learn and accommodate themselves to the group's meme within a relatively short period, possibly only weeks or months. Memes also adjust, but in the medium run of years to decades. By definition, v^{max} is innate, but it, too, can adjust in the long run, over several generations. Thus for simplicity we assume that, at any given time scale, only a single (scalar) variable is adapting. With the assumption of monotone dynamics, the direction of change is immediate from the definition of fitness: values of v that bring higher fitness become more prevalent in the population at the expense of values that bring lower fitness.

The last element of our model incorporates the idea that external reputation is carried by the group as a whole, and defines the frequency f with which an individual encounters culprits. Consider a group of individuals with average negative reciprocity level $\bar{v} > c$. Outsiders on average correctly perceive an individual member's group affiliation and know the group's reputation, an unbiased estimate of $\bar{v}$, but have no other credible information regarding any specific group member. It is intuitive that a group with a reputation for higher levels of negative reciprocity will deter more outsiders from choosing D and thus its members will experience lower f. The Appendix confirms this intuition, and derives a smooth decreasing encounter function $f(\bar{v})$. Here we take the function f as exogenous and note that it will be shifted by changes in the group's environment, including the composition of neighboring groups: this is therefore a partial equilibrium approach. A convenient parameterization is $f(\bar{v}) = \exp(-\bar{v}/b)$.

The next section derives the uniform level v^o that is optimal for the group given the encounter function $f(\bar{v})$. Derivation of v^o is conceptually and technically straightforward, but its relevance is not immediately obvious, due to the basic viability problem. We will show that v^n mediates a close connection of v^o to the individual optimum and hence to the group average $\bar{v}$. The Appendix begins by listing the definitions of the key variables.

5. Results

Here we work with the simple parameterizations of the fitness loss function $\rho(x)$ and the encounter function $f(\bar{v})$ introduced in the previous section, leaving generalizations to the Appendix. Recall that a proportion $f(\bar{v})$ of encounters with outsiders are defections, yielding direct payoff –1 together with losses v due to costly negative reciprocity and ρ due to deviating from the group norm. Encounters with cooperators (proportion 1- $f(\bar{v})$) yield fitness payoff 1, so the individual's expected fitness is

$$W(v \mid \bar{v}, v^n) = 1(1-f(\bar{v})) - 1(1+v+\rho(v^n - v))f(\bar{v}) + R = 1 - f(\bar{v})(2+v+\rho(v^n - v)) + R,$$

where R is the base-level fitness including the (positive) effect on one's status from other group members' deviations from the norm v^n. The expression above does not account for the possibility that the individual will ever play D, but (as shown in the Appendix) shows this omission is harmless. The intuition is that the vengeance parameter affects own fitness when an individual cooperates but not when he defects, because defectors are never suckers. (More formally, terms that capture the own-effects of playing D are independent of v, and hence have no impact in our derivations.) Also, recall from the previous section that in the pure status case, R cancels the mean contribution of ρ. Hence in this case the group's average fitness is simply

$$W^g(\bar{v}) = 1(1-f(\bar{v})) - (1+\bar{v})f(\bar{v}) = 1 - f(\bar{v})(2+\bar{v}).$$

The first result shows that short-run learning dynamics will drive v and hence $\bar{v}$ toward some individually optimal level v^*. These short-run learning dynamics are assumed to operate at a time scale where v^n and a are constant: indeed, this defines the concept of the short run.

Proposition 1. In short run equilibrium, $v = \bar{v} = v^* = [v^n - a]$, truncated to the interval $[0, v^{max}]$, maximizing individual fitness for the given meme v^n and a.

The argument proceeds as follows. Recall that a v–cooperator encountering a defector will receive fitness loss $[1+v+\rho(v^n - v)]$, the sucker payoff plus the cost of imposing negative reciprocity plus the social loss from departing from the norm. The same individual will receive a fitness gain of 1 in encounters with cooperators. For given $\bar{v}$ and v^n, short run selection will drive v towards values that increase individual expected fitness $W(v \mid \bar{v}, v^n)$ or equivalently, that decrease the simpler expression $v + \rho(v^n - v)$. The first-order condition is $1 = \rho'(v^n - v) = (v^n - v)/a$, with solution $v^* = v^n - a$. It is easy to see that W is single peaked at v^*, so short run dynamics (under our monotonicity assumption) push the individual's parameter towards this optimum. The optimum will be attained as long as the value is within the allowable range; otherwise v^* is truncated below at 0 and above at v^{max}. Since learning dynamics are rapid, we obtain the desired conclusion that v^* is a good approximation of an individual v and an even better approximation of their average $\bar{v}$.

Of course, the individual optimum v^* does not necessarily maximize the group's fitness $W^g(\bar{v}) = 1 - f(\bar{v})(2+\bar{v})$. The group optimum v^o is the value that maximizes this expression on $(0, v^{max}]$. Inserting $f(v) = \exp(-v/b)$, the first order condition reduces to 2 +

$v = -f/f' = b$, so v^o is b - 2, truncated to $(0, v^{max}]$. While the solution here is particularly simple, the Appendix shows that similar conclusions hold quite generally.

What then is the relation between the group optimum v^o and the individual optimum v^*? Assume for the moment that both are interior, so $v^* = v^n - a$ and $v^o = b$-2. Our second result is that medium run meme selection aligns them as follows:

Proposition 2. Coevolution of memes and individual learning drives actual behavior v^* toward the group optimum v^o in the medium run, and interior equilibrium is achieved at $v^n = a + b - 2$.

This second result is easily established in the present setting. The group meme, embodied in the parameters a and v^n, is subject to selective pressures in the medium run, and W^g is again a single-peaked function. Any group whose memes bring $v^* = v^n - a$ closer to $v^o = b - 2$ has a selective advantage. Again, any monotone dynamics will work for this statement. So in the interior case considered, we get the expression claimed.

Our final result is a corollary of Proposition 2, taking into account the long run evolution of the individual's capacity v^{max}. If the constraint v or $v^* \leq v^{max}$ binds in the medium run, then there is a selective advantage to individuals with higher genetic capacity for negative reciprocity and for group memes that encourage its expression. (Durham, 1991, features several examples of such coevolution, such as lactose tolerance in herding communities.) Thus there is no truncation in the long run and the algebraic expressions can be rewritten as in the following result.

Proposition 3. Coevolution of memes and genes produces the socially optimal negative reciprocity level in long run evolutionary equilibrium, i.e., $v^o = v^*$, but the supporting meme, $v^n = v^o + a$, exaggerates the optimal level.

There can be shifts in the environment (as captured in the parameter b) and in the punishment technology (as captured in c). These shifts will affect the encounter function f and hence the group optimum v^o. Our results suggest that memes will adjust to these shifts under selective pressure in the medium run (and genes will adjust if necessary in the long run) so that individual behavior v^* will track the new group optimum. The coevolution of the meme (v^n and a) with the gene (v^{max}) allows actual behavior to track optimal behavior as the environment changes.

The Appendix shows that this conclusion holds under conditions far more general than the simple parametric model used here. The derivation starts with consistent estimates of the probabilities that two strangers will choose C or D given imperfect observation of each other's v parameters. It then identifies regions in the perceived characteristic space where the individual will choose C or D as in Figure 2 below. Here individual fitness is given by a sum of integrals over the choice regions. The encounter function f and the first order condition $1=\rho'(v^n - v)$ turn out to arise naturally in this setting.

[Figure 2 about here]

Two other technical questions are dealt with in Friedman and Singh (2004). First, how can $v^{max} > c$ get started from an initial value of $v^{max} = 0$? The key idea is that small values of v turn out to have selective advantage *within* the group because they are complementary with positive reciprocity. Second, how can high-$\bar{v}$ groups protect their reputation from faked membership by individuals who actually are members of low $\bar{v}$ groups? Our idea is that the high-$\bar{v}$ groups enjoin punishment of such individuals whenever they are detected. The same paper also contains an extended literature survey.

We close this section with some interpretive remarks. In the model everyone has the same vengeance parameter v and makes the same choices in equilibrium. Of course, many sources of variation are omitted from the model—members of a given group have different life experiences and different temperaments and they may resolve ambiguous situations differently—so in reality there will always be some behavioral heterogeneity; see Friedman and Singh (2003) for a model incorporating observational as well as behavioral errors (but no group structure). Even ignoring such heterogeneity, one might wonder about the status impact when everyone falls short of the group norm v^n by the same amount a. In equilibrium, of course, there is no net effect on status because the shortfall by others has impact R that exactly offsets the impact ρ of one's own shortfall. Actually, it seems to us a realistic and appealing feature of the model that actual behavior v falls short the group's vision of proper behavior v^n.

6. Discussion

Our argument can be summarized briefly. A capability for negative reciprocity is a significant part of the human emotional repertoire. We model its important role in sustaining cooperative behavior but highlight an intrinsic free-rider problem: the fitness benefits of negative reciprocity are dispersed throughout the entire group, but the fitness costs are borne personally. Evolutionary forces tend to unravel people's willingness to bear the personal cost of punishing culprits. In our model, the countervailing force that sustains negative reciprocity is a group norm together with low-powered (and low-cost) group enforcement of the norm. Such memes coevolve with personal tastes and capacities so as to produce the optimal level of negative reciprocity.

One could object to our account on several grounds. First, it is too simple. The underlying social dilemma was modeled as a specific prisoner's dilemma game. It is straightforward to adapt the model to other parameterizations of the prisoner's dilemma, but this evades the real point. In reality, the stakes and complexity of social interactions vary considerably, and actual memes are more complex and variable than in our model. Ours is the usual response: insight is clearest with an appropriate simple model, and for specific applications the model can be extended as necessary, to deal with specific complexities that are essential. A similar response can be made to the issue of tackling n-

person rather than dyadic social dilemmas: the essential logic of our analysis appears to extend to the more general case.

One could also object that the model is too complicated, especially if the main goal is to explain cooperation. Norms of cooperative behavior and their enforcement could be modeled directly. The same apparatus should suffice: preferences that offer a utility gain (but not a fitness gain) for positive reciprocity together with a social norm from which deviations lead to fitness loss. Negative reciprocity thus seems redundant. Our response is twofold. First, our primary goal is to explain negative reciprocity, not cooperation *per se*. Second, since culprits are rare and cooperators are ubiquitous in successful society, the fitness cost of a meme that relies entirely on positive reciprocation might be excessive. Our suggestion, therefore, is that social norms of negative reciprocity, in taking advantage of biological capacities in that direction, are able to reduce the burden on direct social norms of positive reciprocity in sustaining cooperative behavior. Thus, the existence of direct social norms of positive reciprocity does not make negative reciprocity redundant.

A third objection to our account is that it is too powerful: all sorts of behavior, including behavior that has never been seen and never will, could be described as coevolutionary equilibria. We concede this point, but have been unable to find a simpler account that convincingly explains the viability of preferences for negative reciprocity. Of course, one needs additional principles to get a reasonably sharp theory, and here we have relied on anthropological observations of phenomena such as "cultures of honor." There are indeed many ways to capture the potential gains to cooperation. Social insects, for example, rely on close genetic kinship. Likewise, bipedalism is not the only (or even necessarily the best) form of locomotion: it is worth studying because it is the one humans use. We claim nothing more (nor less) than this for our focus on negative reciprocity as a means of reaping the gains of cooperation.

It is natural to speculate how our model applies in different societies. Readers with specific knowledge may be in a position to assess the model's application to hunter-gatherer bands or to villagers. Here the parameter b would reflect directly the uncooperative tendencies of people from neighboring bands or villages, and c the opportunities to identify, track down and inflict harm on them. The parameter b might be related to the average vengefulness of these neighboring groups. To the extent that these groups are similar to the focal group, then, in a general equilibrium, $b = \psi(v)$, where ψ can still depend on environmental factors. In this case, the equilibrium value of v^o that was derived in Proposition 2 now reduces to the solution to $v^o = \psi(v^o) - 2$.

In more highly structured societies, some important acts of negative reciprocity are performed by designated specialists (e.g., courts and police), rather than solely by the aggrieved individual. This may lower the marginal cost c of negative reciprocity but not to zero, since it is still costly to lodge a complain, to testify, etc, and many situations (e.g., office politics) are not well suited for specialists. Our model therefore still applies to more complex societies, but it is incomplete in that it takes as given the institutional mechanisms that alter the technology parameter c.

What are the empirical implications and applications of our model? One can easily imagine laboratory experiments that would distinguish a taste for negative reciprocity from the egalitarian preferences hypothesized by recent writers. Fehr and Gächter (2000) collect many of the results so far, which generally confirm strong tastes for negative reciprocity. The comparative statics of the model are also clear in principle, and testable with anthropological data: norms of negative reciprocity and actual vengeful behavior should vary systematically with the hostility of the environment, the technology for harming culprits, and the technology for enforcing group norms. If the model is on the right track, there is reason to hope that extremely dysfunctional vengeful behavior might improve over time, as the relevant memes evolve**.**

Appendix

Notation

v^n	Group's normative negative reciprocity level
$\rho(x)$, $x = v^n - v$	Fitness loss ρ for deviation x imposed on deviator by group
a	Tolerance parameter when $\rho(x; a) = x^2/(2a)$
v^{max}	Maximum possible taste for negative reciprocity
$v \in [0, v^{max}]$	Actual negative reciprocity cost an individual prefers
$\bar{v} \in [0, v^{max}]$	Group average of v
$f(\bar{v})$	Frequency with which an individual encounters culprits
b	Environmental hostility parameter when $f(\bar{v}) = \exp(-\bar{v}/b)$.

Alternative Loss Functions

Consider the case $\rho = \exp(k\,|v^n - v|) - 1$, where k is a positive parameter that measures the severity of the enforcement of the norm. The kink in ρ at 0 implies a first order loss for first order small deviations. The first order condition $\rho'(v^n - v) = 1$ is now $k\exp[k(v^n - v)] = 1$, with solution $v^* = v^n + \ln k/k$. If $k \le 1$ then $v^* \le v^n$ and the solution is still of the form $v = v^n - a$, and the previous analysis therefore carries over to this case. If $k > 1$, we have a corner solution, given by $v^* = v^n$, which is a limiting case of $v^n - a$ as a approaches 0. In the medium run equilibrium in this case, $v^n = v^o$, that is, the memes that support this group-optimal equilibrium include the actual optimum value v^o. Thus the analysis proceeds as in the main text, with a treated as 0.

Asymmetry can be introduced by setting $\rho = 0$ for $v > v^n$, or by using different values of k for positive and negative deviations. Since $v^* \le v^n$ is the relevant range for solutions, such asymmetries will have no effect on the subsequent analysis.

Alternative Assumptions about Status

Recall the expression for individual fitness $W(v \mid \bar{v}, v^n) = 1 - f(\bar{v})(2+v+\rho(v^n - v)) + R$. Suppose now that status is not completely relative, so that R only partially

cancels out $\rho(v^n - v)$. We can model this by introducing a parameter $t \in [0, 1]$ that measures the net loss of average fitness due to deviations from the norm. Group average fitness becomes $W^g(\bar{v}) = 1 - f(\bar{v})(2 + \bar{v} + t\rho(v^n - \bar{v}))$. With f and ρ as specified in the main text, the first order condition for the medium run equilibrium is now $[1 - t(v^n - v)/a] \exp(-v/b) = -[2 + v + t(v^n - v)^2/2a] (-1/b) \exp(-v/b)$. Canceling the exponential terms, multiplying through by b, and substituting $v^n - v$ with a, yields $b(1 - t) = (2 + v^n - a + at/2)$, or $v^n = a(1 - t/2) + b(1 - t) - 2$.

If $t = 0$, we have the case analyzed in the text. At the other extreme, $t = 1$, only absolute status matters. In that case, $v^n = a/2 - 2$, independent of the parameter b. In general, greater weight on absolute rather than relative status (i.e., a higher t) decreases the equilibrium norm v^n, since the derivative $dv^n/dt = -a/2 - b$ is negative. The comparative statics for v^n with respect to a and b are qualitatively the same for all values of t in the unit interval, i.e., v^n increases as either a or b increases. In words, if enforcement is less stringent (higher a) or the environment is more hostile (higher b), then the norm of negative reciprocity in the medium run equilibrium will be higher.

Probabilities of Cooperation

To derive key constructs from more general assumptions, we first solve the decision problem faced by an individual encountering a new partner, or "stranger." The encounter function f and the characterization of the individual optimum will emerge endogenously. Let $i = 1$ index the given individual and $i = 2$ index the stranger. Their true degrees of vengefulness (v^1, v^2) are imperfectly perceived by the other person; 1's perception of 2's v is $\hat{v}^2 = v^2 + e^2$, and similarly (replacing 2 by 1) for 2's perception of 1. It is common knowledge that the perception errors (e^1, e^2) have mean zero and joint cumulative distribution function $G(e^1, e^2)$.

The expected payoffs to cooperation $W^i(C|\ldots)$ and to defection $W^i(D|\ldots)$ can be expressed in terms of i's perceptions of $j = 3 - i$ and i's own characteristics as follows. Let $p^i \in I = [0,1]$ be j's estimate of the probability that i will play C; for the moment it is arbitrary, but we shall derive it shortly. Let $\alpha^i = v^i + \rho(v^{n(i)} - v^i)$ be the full cost of negative reciprocity to i, taking into account the loss ρ that his group imposes when he deviates from the norm $v^{n(i)}$. Let $\tilde{e}^i$ denote the induced estimation error of α^i.

Then $W^i(C) = (1)p^j + (-1 - \alpha^i)(1 - p^j) = -(1 + a^i) + p^j(2 + \alpha^i)$, and $W^i(D) = (2 - v^j/c)p^j + (0)(1 - p^j) = p^j(2 - v^j/c)$. Each person i chooses C when the perceived advantage $A^i(p^j, v^j, \alpha^i) = W^i(C) - W^i(D)$ is positive and chooses D when A^i is negative.

Now we need some second-order reasoning. Write j's perception of i's perceived advantage as $A^i(p^j, v^j + e^j, \alpha^i + \tilde{e}^i)$. The error $\tilde{e}^i$ reflects the fact that j knows i's

negative reciprocity cost α^i imperfectly, and the error e^j is included because *j* realizes that *i* knows *j*'s own *v* imperfectly. (The error e^j was dropped out of the $W^i(D)$ expression above because it has mean zero, but now we need to keep track of it because covariances can be relevant.) The probability p^j is still arbitrary, but now we have the machinery in place to enforce consistency.

The construction of consistent (i.e., Bayesian Nash equilibrium) probability estimates uses best response B to map (p^1, p^2) into an updated choice (q^1, q^2), and looks for a fixed point. The idea is that the tentative choice probabilities plugged into the decision function *A* imply new choice probabilities, and the probabilities are internally consistent at a fixed point. Formally, the first component of $B(p^1, p^2)$ is $q^1 = m\left[A^1\left(p^2, v^2 + e^2, \alpha^1 - \tilde{e}^1\right) \mid G\left(e^1, e^2\right)\right]$, where the expression $m[a(x) \mid F(x)]$ denotes the measure (i.e., the probability mass) of the set of *x*'s such that $a(x) \geq 0$, given that *x* has distribution function *F*. The second component of *B* is $q^2 = m\left[A^2\left(p^1, v^1 + e^1, \alpha^2 + \tilde{e}^2\right) \mid G\left(e^1, e^2\right)\right]$.

One can show that the mapping $B:(p^1, p^2) \mapsto (q^1, q^2)$ of the positive unit square I^2 into itself satisfies the assumptions of the Brouwer theorem and therefore has a fixed point. This conclusion holds for any particular choice of (v^1, v^2); indeed, the mapping *B* depends smoothly on (v^1, v^2) if *G* has a density function. Therefore one can assign (not necessarily uniquely) fixed-point probability estimates (p^1, p^2) as a function of (v^1, v^2). Thus we have the mapping we sought, call it $P:\left[0, v^{\max}\right]^2 \to I^2, \left(v^1, v^2\right) \mapsto \left(p^1, p^2\right)$. One can verify (although it is not necessary for our purposes) that *P* is the assessment component of a Bayesian Nash equilibrium.

In practice, a nice way to implement *P* is to begin with initial estimates $p^1 = p^2 = 0.5$ and to iterate using the *B* map (for the actual values of the *v*'s) until convergence. The intuition is not that people actually do the iteration or the calculation, but rather that a stable convention emerges on how likely you (as member of a group with a particular value of *v*) are to encounter *C* play from a stranger with given apparent *v*.

The Individual Optimum and the Encounter Function

The next task is to derive general expressions for fitness functions and to characterize the individual optimum. We focus on a particular individual (*i* = 1 in the last subsection) whose negative reciprocity parameter *v* is to be shaped by the learning process. Others' perceptions of him have mean $\bar{v}$ and remain constant during this process; the interpretation in the text was that the others perceive his group affiliation but have no other credible information about him.

The individual faces an environment defined by a distribution function *F(u)* for strangers' negative reciprocity parameters $v^2 = u$. The distribution *F(u),* together with the mapping *P* derived above, induces a distribution function $H(p,u \mid \bar{v})$, where *p* denotes the first component p^1 of $P(\bar{v},u)$. The distribution *H* summarizes the fitness-relevant data for the individual: the probability *p* that the stranger will play C and her (correlated) negative reciprocity parameter *u*. Monotonicity properties of the mapping *P* imply an ordering by $\bar{v}$ of the distributions *H* via first-order stochastic dominance.

Consider the possible values of (*p, u*) in the rectangle $I \times \left[0, v^{\max}\right]$, as in Figure 2 of the text. Simplifying the notation of the previous subsection, the individual's decision function is $A^1\left(p,u,\alpha^1(v)\right) = A(p,u,\alpha) = -(1+\alpha) + (u/c + \alpha)p$. The locus $A(p,u,\alpha) = 0$, which is the graph of the relation $p = \dfrac{1+\alpha}{u/c+\alpha}$, separates the rectangle into two regions, denoted [C] and [D] to indicate the individual's choice. The measure (or probability mass, using the distribution *H)* of these regions gives the overall probabilities of C and D play by an individual whose imperfectly perceived negative reciprocity parameter is $\bar{v}$.

The individual's fitness is the expectation (with respect to the distribution *H)* of the fitness payoff to C or D over the possible new partners. It is given by the Stieltjes integral

$$w(v \mid \bar{v}, H, \rho) = \int_{p=0}^{1} \int_{u=0}^{v^{\max}} \max\{W(C), W(D)\} H(dp, du \mid \bar{v}) = \tag{1}$$
$$\iint_{[C]} W(C) H(dp, du \mid \bar{v}) + \iint_{[D]} W(D) H(dp, du \mid \bar{v}).$$

The key calculation is the fitness gradient. Taking the derivative in (1) with respect to *v* we obtain

$$\frac{dw}{dv} = \iint_{[C]} \frac{dW(C)}{dv} H(dp, du \mid \bar{v}) + \iint_{[D]} \frac{dW(D)}{dv} H(dp, du \mid \bar{v}) \tag{2}$$
$$+ \oint_{[A=0]} \left(W(C) - W(D)\right) \cdot (dA/dv) H(dp, du/\bar{v}).$$

The last term in (2) is a line integral over the locus $A = 0$. It comes from the relevant generalization of the fundamental theorem of calculus (or a special case of Stokes' Theorem) because the locus moves when *v* changes. Conveniently, it is zero because *W(C) = W(D)* precisely on the locus $A = 0$ where C and D are equally fit.

Recall that *W(D)* = $p(2 - u)$ depends on the stranger's negative reciprocity parameter *u* but is independent of the individual's own value of *v*, so the middle term in (2) also vanishes. That leaves only the first term, whose integrand is the derivative of $W(C) = -(1 + \alpha(v)) + p(2 + a(v))$ with respect to *v*. Hence

$$\frac{dw}{dv} = -(d\alpha/dv) \iint_{[C]} (1-p) H(dp, du \mid \bar{v}) = \left[\rho'\left(v^n - v\right) - 1\right] f(\bar{v}), \tag{3}$$

where the encounter function used in the text is now seen to be precisely the probability $f(\bar{v}) = \iint_{[C]} (1-p) H(dp, du \mid \bar{v})$ that the individual is the victim of the sucker payoff. This

probability is independent of *v*, so the shape of the payoff function *w* depends only on the group's enforcement function ρ.

It now is clear that the simple argument in the text applies directly since it was based on the same first order condition $\rho'(v^n - v) = 1$ that emerges here. We conclude as in the main text that individuals will adapt monotonically towards a point *v** somewhat below the group norm v^n, with the size of the gap depending on the rigor with which the norm is enforced.

Presumably there is some family of joint distributions *H* that gives rise to the exponential family $f(\bar{v})$ used in the text, but its description remains an open question. A deeper open question is to characterize the distribution *H* from parameters of a general equilibrium model whose state variable is the distribution of memes across all groups. Analytical work with such models involves nonlinear partial differential equations and is well beyond the scope of the present paper. Numerical simulations as in Boyd *et al* (2003) and numerous other studies could also provide some insight.

Acknowledgements

The first author is grateful to CES and the University of Munich for hospitality while writing the first fragments in May 1997. We have benefited greatly from the comments of Ted Bergstrom, Sam Bowles, Robert Boyd, Herb Gintis, Jack Hirshleifer, Peter Richerson, Donald Wittman, and seminar audiences at JAFEE2000, Indiana, Purdue, UCLA, and UCSC. Two anonymous referees and the editors of this journal helped improve the final version. Remaining shortcomings are our responsibility.

References

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

Akerlof, G. & Kranton, R. (2000). Economics and identity. *Quarterly Journal of Economics, 115*, 715-753.

Black-Michaud, J. (1975). *Cohesive Force: Feud in the Mediterranean*. Oxford: Basil Blackwell.

Blackmore, S. (2000). The power of memes. *Scientific American*, October, 64-73.

Blackmore, S. (1999). *The Meme Machine*. Oxford: Oxford University Press.

Bowles, S. (1998). Cultural group selection and human social structure: the effects of segmentation, egalitarianism and conformism. University of Massachusetts, Amherst working paper.

Bowles, S. & Gintis, H. (1998). The evolution of strong reciprocity. University of Massachusetts, Amherst working paper.

Boyd, R, Gintis, H., Bowles S. & Richerson P. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Science, 100:6*, 3531-3535.

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

Boyd R. & Richerson P. J. (1990). Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology*, *145,* 331-342.

Boyd R. & Richerson P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*, 171-195.

Boyd R. & Richerson P. J. (1998). The evolution of human ultra-sociality. In: Eibl-Eibisfeldt I., Salter F. K. (Eds.) *Indoctrinability, Ideology and Warfare: Evolutionary Perspectives*. Berghahn Books, New York

Catanzaro, R. (1992). *Men of Respect: A Social History of the Sicilian Mafia.* New York: The Free Press.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*, 2 vols. New York: Appleton.

Davis, J. (1980). *Antropologia della Societa Mediterranee: Un'analisi Comparata.* Turin: Rosenberg & Sellier.

Dawkins, R. (1976). *The Selfish Gene.* New York: Oxford University Press.

Dawkins, R. (1982). *The Extended Phenotype: The Gene as the Unit of Selection*. San Francisco: Freeman.

Dupré, J. (1987). *The Latest on the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press.

Durham, W. H. (1991). *Coevolution: Genes, Culture, and Human Diversity*. Stanford, CA: Stanford University Press.

Eshel, I. (1983). Evolutionary and continuous stability. *Journal of Theoretical Biology*, *103*, 99-111.

Farb, P. (1978). *Man's Rise to Civilization: The Cultural Ascent of the Indians of North America*, New York: Penguin.

Fehr, E. & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, *14, 3*, 159-182.

Fehr, E. and Henrich, J. (2003), Is Strong Reciprocity a Maladaptation? In (Ed.) Hammerstein, P., *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press, forthcoming.

Frank, R. (1987). If *Homo Economicus* could choose his own utility function, would he want one with a conscience? *American Economic Review, 77*, 593-604.

Frank, R. (1988). *Passions within Reason*: *The Strategic Role of the Emotions*. New York: WW Norton.

Friedman, D. (1991). Evolutionary games in economics. *Econometrica*, *59*, 637-666.

Friedman, D. & Singh, N. (1999). On the viability of vengeance. UC Santa Cruz Working Paper, http://econ.ucsc.edu/faculty/workpapers.html.

Friedman, D. & Singh, N. (2003). Equilibrium vengeance. UC Santa Cruz Working Paper, http://leeps.ucsc.edu/leeps/projects/misc/EqVenge/EqVenge.

Friedman, D. & Singh, N. (2004). Vengeance evolves in small groups. In Huck, S. (Ed.), *Festschrift in Honor of Werner Güth* (forthcoming).

Fudenberg, D. & Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.

Galaty, J.G. & Bonte, P., eds. (1991). *Herders, Warriors and Traders: Pastoralism in Africa*. Boulder, CO: Westview Press.

Gilmore, D.D. (1991). *Manhood in the Making: Cultural Concepts of Masculinity*. New Haven: Yale University Press.

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

Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*, 153-172.

Guttman, J.M. (2003). Repeated interaction and the evolution of preferences for reciprocity. *The Economic Journal*, *113*, 631-656.

Haldane, J.B.S. (1955). Population genetics. *New Biology*, *18*, 34-51.

Hamilton, W.D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*, 1-52.

Harrington, J.E. (1989). If homo economicus could choose his own utility function, would he want one with a conscience?: Comment. *American Economic Review*, *79*, 588-593.

Heckathorn, D. (1996). The dynamics and dilemmas of collective action. *American Sociological Review*, *61*, 250-277.

Henrich, J. and 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.

Hirshleifer, J. (1987). On the emotions as guarantors or threats and promises. In: J. Dupré (ed.) *The Latest on the Best: Essays in Evolution and Optimality.* Cambridge, MA: MIT Press.

Kaufman, S. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford U Press.

Leimar, O. and Hammerstein, P. (2000). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London B, 268*, 745-753.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics, 1*, 593-622.

Lowie, R. H. (1954). *Indians of the Plain*. New York: McGraw-Hill.

MacDonald K. B. (1994). *A People That Shall Dwell Alone: Judaism as a Group Evolutionary Strategy*. Westport, CT: Praeger.

Nelson, R.R. and Winter, S.G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press of Harvard University Press.

Nisbett, R.E. & Cohen, D. (1996). *Culture of Honor: the Psychology of Violence in the South*, Boulder, CO: Westview Press.

Nowak, M. A. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573-577.

O'Kelley, C.G. & Carney, L.S. (1986), *Women and Men in Society*. New York: D. Van Nostrand Co.

Peristiany, J.G. ed., (1965). *Honor and Shame: The Values of Mediterranean Society*. London: Weidenfeld and Nicolson.

Pettigrew, J. (1975). *Robber Noblemen: A Study of the Political System of the Sikh Jats*. London: Routledge & Kegan Paul.

Price, G. R. (1970). Selection and covariance. *Nature, 227(5257, August 1),* 520-521.

Rosenthal, R. W. (1996). Trust and social efficiencies. Boston University manuscript.

Sethi, R. & Somanathan, E. (2003). Understanding reciprocity. *Journal of Economic Behavior and Organization*, *50*, 1-27.

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

Sugden, R. (1986). *The Economics of Rights, Co-operation and Welfare*. New York: B. Blackwell.

Trivers, R. (1985). *Social Evolution*. Menlo Park CA: Benjamin/Cummings.

Weibull, J.W. (1995). *Evolutionary Game Theory.* Cambridge, MA: MIT Press.

Weingart, P., Boyd, R., Durham, W. H. & Richerson, P. J. (1997). Units of culture, types of transmission. In Weingart, P., Mitchell, S. D., Richerson, P. J., & Maasen, S. (Eds.) *Human By Nature: Between Biology and the Social Sciences*. Lawrence Erlbaum, Mahwah, NJ. .

Wilson, E. O. (1980). *Sociobiology*. Cambridge MA: Harvard University Press.

Wynne-Edwards, V.C. (1962). *Animal Dispersion in Relation to Social Behavior.* Edinburgh: Oliver and Boyd.

Table 1: Fitness with No Negative reciprocity

(v =0)	C	D
C	1 , 1	-1, 2
D	2 , -1	0, 0

Table 2: Fitness with Negative reciprocity

(v >0)	C	D
C	1 , 1	-1-v, 2 - v/c
D	2 - v/c , -1-v	0, 0

Figure 1: The Advantage of Cooperating.

The fitness advantage A(s)=W(C)-W(D) is graphed as a function of the population fraction s playing C for two values of the negative reciprocity parameter *v*. The graph of A rotates counterclockwise as *v* increases.

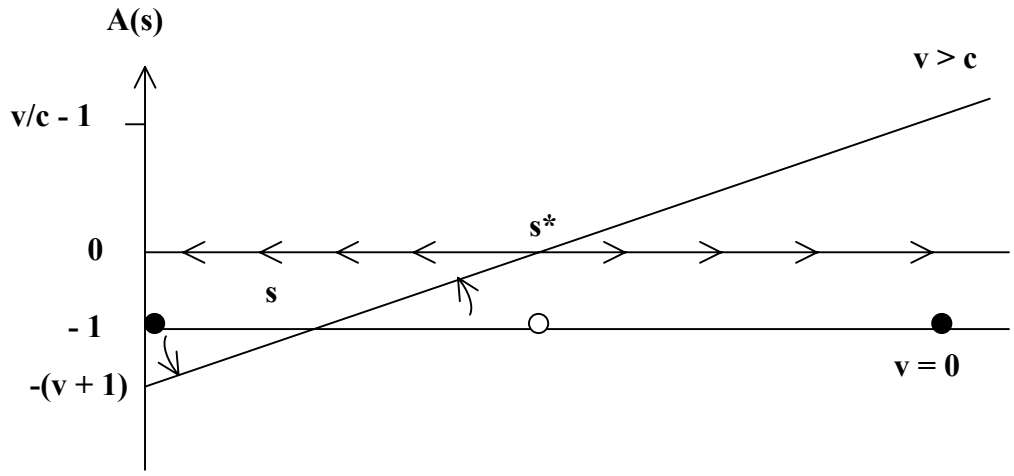

Figure 2: The Decision Rule.

The appropriate choice of C or D is given by the sign of the advantage function A(p,u), where p is the probability that the partner will choose C and u is an unbiased estimate of her negative reciprocity parameter. The A=0 locus shifts up with increases in the decision maker's direct (v) or full (α) negative reciprocity cost.