The Puzzle of Prosociality*

Herbert Gintis

October 10, 2001

Abstract

How is cooperation among large numbers of unrelated individuals sustained? Cooperation generally requires *altruism*, where individuals take actions that are group-beneficial but personally costly. Why do selfish agents not drive out altruistic behavior? This is the *puzzle of prosociality*.

Altruism is supported by culture. Sociology treats culture as a set of norms that are transmitted by socialization institutions and internalized by individuals. Altruism, in this approach, is thus sustained by the internalization of norms. Biology treats culture as knowledge that is passed to children from parents (vertical transmission), from other prominent adults (oblique transmission), and from peers (horizontal transmission), such that individuals with higher payoffs have a higher level of biological fitness, leading norms to follow a dynamic of Darwinian selection. Altruism, in this approach, can be sustained only if group selection is feasible, which it rarely is. Economics uses evolutionary game theory to model culture as strategies deployed in social interaction that evolve according to a replicator dynamic, in which individuals shift from lower to higher payoff norms. In this approach, altruism cannot be sustained, but cooperation is possible with repeated interactions and a sufficiently low discount rate. This paper integrates these approaches and shows that altruism, as well as norms that reduce both individual and group payoffs, can be supported in a stable equilibrium.

1 Introduction

How is cooperation among large numbers of unrelated individuals sustained in human societies? Cooperation generally requires *altruism*, where individuals take

*Department of Economics, University of Massachusetts, Amherst, hgintis@mediaone.net, http://www-unix.oit.umass.edu/˜gintis. Presented at the Santa Fe Institute, October 17, 2001, based on research done in conjunction with the on-going SFI workshop on the co-evolution of behaviors and institutions. I would like to thank Samuel Bowles, Ernst Fehr, and Eric Alden Smith for helpful comments, and the John D. and Catherine T. MacArthur Foundation for financial support.

actions that are group-beneficial but personally costly. Why do not selfish agents drive out altruistic behavior? This is the *puzzle of prosociality*.

Diverse and incompatible answers to the puzzle of prosociality are harbored within the various academic disciplines that comprise the behavioral sciences (Coleman 1990, Heckathorn 1998a,b). This situation violates the basic scientific principle that different forms of scientific explanation should agree in areas where their objects of investigation overlap. This paper integrates some major theoretical tools that operate in the intersection of economics, sociology, and biology. This resolves the puzzle of prosociality in a manner compatible with the basic tenets of all three disciplines.

Altruism is supported by culture. Sociology treats culture as a set of norms that are transmitted by socialization institutions and internalized by individuals (Durkheim 1951, Parsons 1967, Grusec and Kuczynski 1997).[1] Altruism, in this approach, is sustained by the internalization of norms. Biology treats culture as knowledge that is passed to children from parents (vertical transmission), from prominent individuals and social practices (oblique transmission), and from peers (horizontal transmission), such that individuals with higher payoffs have a higher level of biological fitness, leading norms to follow a dynamic of Darwinian selection (Cavalli-Sforza and Feldman 1981, Lumsden and Wilson 1981, Boyd and Richerson 1985). Altruism, in this approach, can be sustained by group selection, which is generally infeasible (Williams 1966, Maynard Smith 1976, Boorman and Levitt 1980). In the biological model, cooperation among unrelated individuals is sustained through "reciprocal altruism," where selfish individuals reciprocate acts of generosity (Trivers 1971, Alexander 1987, Nowak and Sigmund 1998). Economics uses evolutionary game theory to model culture as a set of strategies deployed in social interaction that evolve according to a replicator dynamic, in which individuals shift from lower to higher payoff strategies (Weibull 1995, Gintis 2000a). In this approach, altruism cannot be sustained, but as in the biological approach, cooperation is possible with repeated interactions and a sufficiently low discount rate (Axelrod and Hamilton 1981, Fudenberg and Maskin 1986, Bowles and Gintis 1998). This paper integrates these approaches and shows that altruism, as well as norms that reduce both individual and group payoffs, can be supported in a stable equilibrium without group selection.

Biologists and economists have rejected the "oversocialized" concept of the individual social actor in socialization theory. The notion that an individual's norms are accepted without regard to their payoff-relevance is at odds with the fact that people often reject and transform social rules (Wrong 1961, Gintis 1975, Wrong

[1] For simplicity, we use the term 'norm' to include values, beliefs, standard practices, and other objects of cultural transmission.

1993). Whereas biologists and economists have simply rejected the socialization approach, we merely correct it by adding a payoff-sensitive form of horizontal cultural transmission widely used in evolutionary game theory—the *replicator dynamic* (Taylor and Jonker 1978, Nowak and Sigmund 1998). The replicator dynamic models individuals as, with some positive probability in each period, moving towards higher-payoff norms. Agents thus treat culture instrumentally—as a set of social practices that may be adopted, abandoned, and transformed in organizing social interactions (Gintis 1980, Boyd and Richerson 1988, Skyrms 1996, Sperber 1996, Epstein and Axtell 1997, Young 1998, Binmore 1998, Staddon 2001). Finally, we incorporate the fact, central to biology, that agents who use low-payoff practices may have biological fitness handicaps, leading them to be replaced by agents using practices that afford higher payoffs (Nowak, Page and Sigmund 2000, Gintis 2000a, Gintis 2000b).

For analytical specificity, in this paper we study the dynamics of a two-norm cultural system in which one norm has a fitness advantage over the other, but the disadvantaged norm is transmitted vertically and obliquely.[2] We allow four types of cultural change. First, families pass on their norms to their offspring (vertical transmission). Second, families who use lower payoff norms have fewer offspring (Darwinian selection). Third, offspring may adopt the norms of influential non-parental elders and promulgated by respected social institutions (oblique transmission). Finally, individuals may change their norms to conform to the norms of other individuals who have higher payoffs (replicator dynamics).[3]

The following are examples norms that are disadvantaged in the above sense, and to which our analysis applies:

a. **First Order Altruism.** Personally costly behavior that benefits others in the group at one's own expense.

b. **Second Order Altruism.** Punishing individuals who violate a social norm at a cost to oneself.

c. **Ritualistic Practices.** Engaging in fitness-reducing rituals and practices when fitness-neutral or fitness-enhancing alternatives are available.

d. **Harmful Beliefs.** Reacting to illness, death, crop failure, and other payoff-reducing events by adopting defective explanations and ineffective remedies when fitness-neutral or fitness-enhancing alternatives are available.

[2] Modeling cultural transmission in this manner was pioneered by Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985).

[3] A more complete analysis would model the interaction of genes and culture (Lumsden and Wilson 1981, Feldman, Cavalli-Sforza and Peck 1985, Durham 1991, Feldman and Zhivotovsky 1992, Gintis 2001). We avoid this complication in the body of this paper, but will return to it in the Conclusion.

Note that the first two disadvantaged norms are prosocial in that they enhance the fitness of other group members, whereas the last two are not. Thus the maintenance of disadvantaged norms can be either fitness enhancing or fitness reducing for the society as a whole, depending on the norm in question. In particular, our model can explain the persistence of altruism in equilibrium, without resort to group selection.

From this point onward I shall call the disadvantaged norm 'altruistic,' and the advantaged norm 'selfish.' The analysis applies equally, however, to ritualistic practices, harmful beliefs, and other norms that reduce *both* individual and group welfare. Moreover, I shall assume that the altruistic norm impacts all group members equally, so its magnitude does not affect neither behavior nor relative fitness, and therefore this effect can be dropped from the model.[4]

Our model yields several general conclusions.

- If oblique transmission is absent or favors the selfish norm, the selfish norm always drives out the altruistic norm. This implies that *extra-familial socialization institutions are necessary to support altruism.*.
- When there is oblique transmission of the altruistic norm, a positive frequency of this norm can persist in equilibrium. Depending on the specific assumptions of the model and the specific value of parameters, there can either be two stable "homogeneous" equilibria involving very high frequencies of either the selfish or altruistic norm, or a single stable "heterogeneous" equilibrium involving a moderate frequency of both norms.
- When there are two stable homogeneous equilibria, cultural change induced, say, by an external shock to the system, can dramatically shift the system from the ubiquity of one norm to the ubiquity of the other. Such cultural change is often observed in human societies (Moore, Jr. 1978, Skocpol 1979, Button 1989, Chong 1991, Cox 1993, Barry 1995, Mackie 1996).
- A very high level of cooperation can be sustained in equilibrium by the persistence of a minority of agents who adopt the altruistic norm of *strong reciprocity*: cooperating unconditionally and punishing defectors at a cost to themselves (Gintis 2000b).

.

[4] In modeling the evolution and diffusion of altruistic norms across social groups, we would, of course, explicitly include the group benefits of altruistic norms. Since such an analysis is fairly straightforward, we do not include it here.

2 A Model of Cultural Evolution

Consider a group in which members can either adopt or fail to adopt the altruistic norm A. We shall describe the absence of norm A is norm B, so for instance if norm A is "help strangers in need," then norm B as "do not help strangers in need." Selfish norm B is individually superior in the sense that B-users have fitness 1, as compared with norm A, whose users have fitness $1 - s$, where $0 < s < 1$. On the other hand, the altruistic norm A is more complex and difficult to transmit, since it specifies an action, whereas B specifies non-action in the same situation.

We assume in each period that A-users and B-users pair off randomly and have offspring in proportion to their fitness, after which they die. Families pass on their norms to their offspring, so offspring of AA parents are A-users, offspring of BB parents are B-users, and half of the offspring of AB-families are A-users.[5] This is *vertical transmission*. We also assume that the selfish offspring of AB- and BB-families (B-users) are susceptible to influence by salient A-users in the community and community institutions promoting norm A, a fraction of such offspring becoming A-types. This is *oblique transmission*.

At the beginning of a period, if the fraction of A-users is α, the fraction of AA-families is α^2, who will have a fraction $\alpha^2(1-s)^2\beta$ offspring, all of whom are A-users, where we choose β so that population remains constant from generation to generation. There will also be a fraction $\alpha(1-\alpha)$ AB-families, who will have $\alpha(1-\alpha)(1-s)\beta$ offspring, half of whom are A-users. Finally there will be a fraction $(1-\alpha)^2$ of BB-families who will have a fraction $2(1-\alpha)^2\beta$ of offspring. Adding up the number of offspring, we see that we must have $\beta = 1/(1-s\alpha)^2$ to maintain a constant population size. Thus the frequencies of offspring from AA, AB, and BB families are given by (Cavalli-Sforza and Feldman 1981)

$$f_{AA} = \frac{\alpha^2(1-s)^2}{(1-\alpha s)^2}, \quad f_{AB} = \frac{2\alpha(1-\alpha)(1-s)}{(1-\alpha s)^2}, \quad f_{BB} = \frac{(1-\alpha)^2}{(1-\alpha s)^2}. \quad (1)$$

Second, a fraction γ of offspring of AB-families who are B-users, and a fraction $\nu \leq \gamma$ of offspring of BB-families, switch to being A-users under the influence of oblique transmission. It is easy to check that the change in the fraction of A-users in the next generation is given by

$$\dot{\alpha} = f(\alpha) = \frac{1-\alpha}{(1-s\alpha)^2}(s^2\alpha^2 - s\alpha(1+\gamma) + \alpha\gamma + (1-\alpha)\nu). \quad (2)$$

[5]This form of transmission is *Medelian*, in that it entails unbiased segregation and recombination of phenotypes. Given our assumption that norm A is more difficult to transmit than norm B, it might be more appropriate to bias transmission in favor of B (e.g., by having some fraction of AA parents give rise to B offspring). As will become clear, biasing parental transmission in favor of B will only strengthen our conclusions.

Third, in every time period, each group member i with probability $\delta_1 > 0$ learns the fitness and the type of a randomly chosen other member j, and changes to j's type if j's fitness is higher. However, information concerning the difference in fitnesses of the two strategies is imperfect, so the larger the difference in the payoffs, the more likely the agent is to perceive it, and change. Specifically, we assume the probability p that an agent using A will shift to B is proportional to the fitness difference of the two types, so $p = \delta_2 s$ for some proportionality constant $\delta_2 > 0$.[6]

The expected fraction α' of the population using A after the above shifts is then given by

$$\alpha' = \alpha - \delta_1\delta_2\alpha(1-\alpha)s,$$

which, expressed in differential equation form, and defining $r(\alpha) = -\alpha(1-\alpha)s$

$$\dot{\alpha} = r(\alpha)/\delta_1\delta_2 \tag{3}$$

This is a special case of the *replicator dynamic* in cultural evolution.[7] We now combine these two sources of change in the fraction of A-users, giving

$$\dot{\alpha} = f(\alpha) + r(\alpha)/\delta_1\delta_2.$$

For notational convenience, we shall multiply the right hand side by $\sigma = \delta_1\delta_2$, which merely redefines the unit time period. This gives

$$\dot{\alpha} = h(\alpha) = \sigma f(\alpha) + r(\alpha) \tag{4}$$

where σ now represents the relative speed of the socialization and biologically adaptive processes, given by $f(\alpha)$, in comparison with the social change replicator dynamic, given by $r(\alpha)$. This becomes, in reduced form,

$$\dot{\alpha} = h(\alpha) = \frac{\alpha(1-\alpha)}{(1-s\alpha)^2}\left(\sigma\left(\gamma + \frac{1-\alpha}{\alpha}\nu\right) - s(s^2\alpha^2 - s\alpha(\sigma+2) + 1 + \sigma(1+\gamma))\right). \tag{5}$$

We call the situation $\dot{\alpha} = 0, \alpha \in [0, 1]$ an *equilibrium* of the dynamical system. An equilibrium with $\alpha = 1$ is called an *altruistic equilibrium*, and an equilibrium with $\alpha = 0$ is called a *selfish equilibrium*. We then have the following theorem.

[6]The assumption that agents only switch from lower to higher payoff norms—in our case, from altruistic norm A to selfish norm B—is quite strong, but weakening this assumption will only strengthen our conclusions, as we shall see.

[7]For a more general derivation, see Gintis (2000a), Ch. 9.

Theorem 1. *Assume $\gamma \geq 0$ is given.*

a. If $\gamma = 0$, there is a globally stable selfish equilibrium.

b. If

$$0 < s < s_{min} = \frac{\gamma\sigma}{1+\sigma(1+\gamma)}, \tag{6}$$

there is a globally stable altruistic equilibrium for all $\nu \in [0, \gamma]$.

c. If

$$s_{min} < s < s_{max} = \frac{1}{2}\left\{1+\sigma-\sqrt{(1+\sigma)^2-4\gamma\sigma}\right\}, \tag{7}$$

there are functions $\alpha_*(\nu)$ and $\alpha^*(\nu)$, and constants $\hat{\nu}$ and $\hat{\alpha}$ such that $0 < \hat{\nu}, \hat{\alpha} < 1$ with the following properties:

i. For $0 < \nu < \hat{\nu}$, $\alpha_*(\nu)$ is an increasing function and $\alpha^*(\nu)$ is a decreasing function, and $0 < \alpha_*(\nu) < \alpha^*(\nu) < \hat{\alpha}$. The stable equilibria of the system are $\alpha = \alpha_*(\nu)$ and $\alpha = 1$, while $\alpha^*(\nu)$ is an unstable equilibrium. The interval $[\alpha_*(\nu), \alpha^*(\nu))$ is the basin of attraction of $\alpha = \alpha_*(\nu)$, and interval $(\alpha^*(\nu), 1]$ is the basin of attraction of $\alpha = 1$.

ii. For $\gamma \geq \nu > \hat{\nu}$, $\alpha = 1$ is the only equilibrium.

d. If $s_{max} < s < 1$, there is a function $\alpha^*(\nu)$, increasing in ν, such that for all $\nu > 0$, $\alpha^*(\nu)$ is the only stable equilibrium of the system, and its basin of attraction is (0,1).

The proof of this theorem is straightforward and details are left to the reader. Briefly, there are four zeros of (5), of which two are $\alpha = 0$ and $\alpha = 1$. At most one of the other two equilibria can lie in $\alpha \in (0, 1)$. Investigating the signs of $h(\alpha)$ and $h'(\alpha)$ at $\alpha = 0, 1$ determines the conditions under which these equilibria are stable, as well as the nature of the third equilibrium, should it exist. The assertions concerning varying ν follow the fact that $h(\alpha)$ shifts upward as ν increases, except that $h(1) = 0$ for all ν.

Corollary 1.1. *In the absence of oblique transmission, the selfish equilibrium is globally stable. In the presence of oblique transmission, there are numbers s_{min} and s_{max} with $0 < s_{min} < s_{max} < 1$, such that, $0 < x < s_{max}$ the altruistic equilibrium is stable, and for $s_{max} < s < 1$ there is a heterogeneous equilibrium in which both altruistic and selfish types persist.*

Corollary 1.1 justifies our assumption that oblique transmission always favors the altruistic trait, since without this assumption altruism is impossible. Also, it is

easy to see from (6) and (7) that a weaker replicator dynamic (higher σ), or more strongly biased oblique transmission (higher γ) increase both $s_{\min}$ and $s_{\max}$, thus favoring the existence and stability of the altruistic equilibrium. This justifies our assumption that the replicator dynamic is highly fitness-sensitive, since weakening this assumption merely strengthens our conclusion that an altruism equilibrium exists if the fitness handicap of the A norm is not too great.

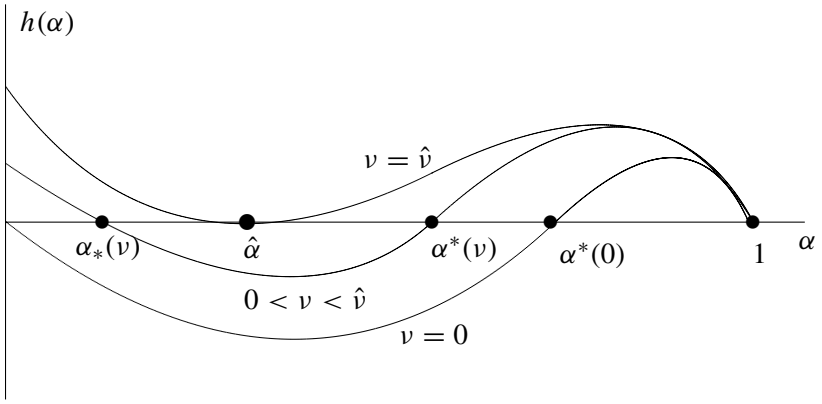

Figure 1: The Comparative Dynamics of Varying Oblique Transmission: The Case of Low Disadvantage ($s_{\min} < s < s_{\max}$).

The *low disadvantage* case of Theorem 1, in which $s_{\min} < s < s_{\max}$, is illustrated in Figure 1, and the *high disadvantage case*, in which $s_{\max} < s < 1$ is illustrated in Figure 2.

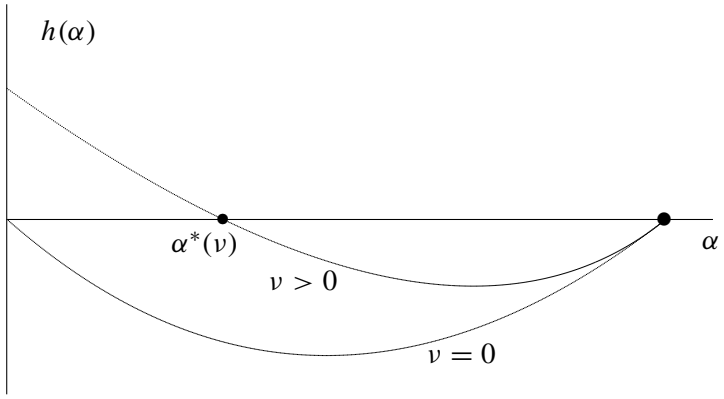

Figure 2: The Comparative Dynamics of Varying Oblique Transmission: The Case of High Disadvantage ($s_{\max} < s < 1$).

3 Cultural Dynamics when Payoffs are Frequency Dependent

We have assumed that the fitness deficit s of the altruistic norm is constant. However, if the payoffs to altruism and selfish behavior are frequency dependent, as would be the case if these norms represented strategies in a noncooperative game, then we will have in general the functional relationship $s = s(\alpha)$. While we could extend Theorem 1 broadly to this new situation, for simplicity we will deal with only partial results. We have

Theorem 2. *Consider a cultural system satisfying the conditions of Theorem 1, except that the fitness deficit of altruism is a differentiable function of the frequency of altruism, $s = s(\alpha)$. Let s_{min} and s_{max} be given by (6) and (7). Then if $s(1) < s_{max}$, $\alpha = 1$ is a stable equilibrium of the cultural dynamic. Also, if $s(0) > s_{min}$, there is a $\nu^* > 0$ and a continuous function $\alpha_*(\nu) \geq 0$ with $\alpha_*(0) = 0$, such that for all $\nu \in [0, \nu^*)$, $\alpha_*(\nu)$ is a stable equilibrium.*

To see that this is the case, note that $h(1) = 0$ and $h'(1) = (s(1)(1 + \sigma - s(1)) - \gamma\sigma)/(1 - s(1))$, which is negative for $s(1) < s_{\max}$. Thus $\alpha = 1$ is a stable equilibrium. Suppose $\nu = 0$. Then $h(0) = 0$ and $h'(0) = \gamma\sigma - s(0)(1+\sigma(1+\gamma)$, which is negative for $s(0) > s_{\min}$. By the implicit function theorem, there is a $\nu^* > 0$ and a function $\alpha_*(\nu)$ with $h(\alpha_*(\nu)) = 0$ for $0 \leq \nu < \nu^*$. Since $h(0) = \gamma\nu > 0$ and $h'(0) > 0$, $\alpha_*(\nu) > 0$ for sufficiently small ν, which proves the theorem.

We conclude that in the frequency dependent case, there is a range of parameter values for which the altruistic norm cannot be invaded by the selfish norm, although there is a second stable equilibrium in which the selfish norm does occur, and for small ν, is adopted by most of the population.

4 Cultural Dynamics when Payoffs do not Affect Fitness

To this point we have treated the payoffs to norms as biological fitness. If we assume payoffs represent the subjective utility of agents rather than their biological fitness, we must replace (1) with

$$f_{AA} = \alpha^2, \quad f_{AB} = 2\alpha(1-\alpha), \quad f_{BB} = (1-\alpha)^2. \tag{8}$$

In this case the equation of motion becomes

$$\dot{\alpha} = h(\alpha) = (1-\alpha)(\sigma(\alpha(\gamma - \nu) + \nu) - \alpha s). \tag{9}$$

Theorem 3. *Suppose the conditions of Theorem 1 hold, except now payoffs represent subjective utility rather than biological fitness. Then there is an altruistic*

equilibrium that is stable if $s < \gamma\sigma$. When $s > \gamma\sigma$, there is a stable equilibrium at $\alpha = \nu\sigma/(s - \sigma(\gamma - \nu)) \in (0, 1)$, with basin of attraction is $\{\alpha|0 \leq \alpha < 1\}$.

We omit the proof, which is straightforward.

If the fitness deficit s is frequency dependent, the behavior of the system is more interesting. For example, suppose in each period members of the population pair of randomly and play a prisoner's dilemma in which the (defect,defect) payoffs are (0,0), the (cooperate,cooperate) payoffs are (1,1), and the (cooperate,defect) payoffs are $(-b, a)$ where $a > 1$, $b > 0$ and $a - b < 2$. The selfish norm is thus to defect, and the altruistic norm is to cooperate. We assume the payoff to each agent is one plus the expected payoff to the game, and we normalize to make the payoff to the selfish strategy unity; i.e.,

$$s(\alpha) = 1 - \frac{2 - \alpha(1 + b)}{1 + a(1 - \alpha)}. \tag{10}$$

Note that $0 < \sigma(\alpha) < 1$ for $b < 1$. We then have the following theorem.

Theorem 4. *Suppose in each period members of the population pair of randomly and play a prisoner's dilemma, with payoffs as given in the previous paragraph. Then*

a. There is an altruistic equilibrium. This equilibrium is stable when $b > \gamma\sigma >$.

b. If there are no other equilibria, the altruistic equilibrium is globally stable.

c. If the altruistic equilibrium is stable but not globally stable, there are two additional equilibria, the smaller of which is stable and the larger of which is unstable, separating the basins of attraction of the smaller equilibrium and the altruistic equilibrium.

d. Suppose the altruistic equilibrium is unstable, which occurs when $b < \gamma\sigma$. Then

   i. If $\sigma(\gamma - 2\nu) > (a - 1)/(a + 1)$, there is an interior globally stable equilibrium.

   ii. If $\sigma(\gamma - 2\nu) < (a - 1)/(a + 1)$, then the selfish equilibrium is globally stable.

The proof of this theorem is straightforward and is left to the reader.

In short, with frequency dependent payoffs, the case where payoffs do not affect fitnesses exhibit the same array of alternative equilibria types as in the case where payoffs represent biological fitnesses.

5 Maintaining Cooperation Through Punishment

In this section we show that a variant of our model illustrates how second order punishment may persist in equilibrium, and can be powerfully conducive to promoting social cooperation even when its frequency in the population is low. We model cooperation as contributing in a public goods game, defection as not contributing, and altruism as punishing defectors at personal cost.

A group of n individuals plays a public goods game in which each member can either cooperate or defect. Defecting costs nothing, but adds nothing to the payoffs of the other members. Cooperating costs $c > 0$, but contributes an amount $b > c$ shared equally by the other members. In a one-shot encounter, the only Nash equilibrium is universal defection. By using either group selection (Gintis 2000b, Bowles and Hopfensitz 2000, Henrich and Boyd 2001, Bowles 2001, Bowles, Boyd, Gintis and Richerson 2001) or repeated interactions with a suitably low rate of discounting future benefits (Fudenberg and Maskin 1986, Hirshleifer and Rasmusen 1989, Nowak and Sigmund 1998), a high level of cooperation can be sustained in equilibrium. We here show cooperation can also be maintained in our framework without the need for group selection or repeated interactions.

Let A be a norm that induces its bearer to cooperate in the public goods game, and let B be a norm that induces its bearer to defect. Clearly the fitness deficit of the altruistic norm is $s = c$, so Theorem 1 implies that if

$$c < \frac{\gamma\sigma}{1+\sigma(1+\gamma)}, \tag{11}$$

complete cooperation in the public goods game is a stable equilibrium. If this inequality fails but

$$c \leq \left(1+\sigma-\sqrt{(1+\sigma)^2-4\gamma\sigma}\right)/2, \tag{12}$$

full cooperation remains a stable equilibrium, but there is another stable equilibrium with a positive, possibly large, level of defection. If (12) fails, full cooperation is no longer a stable equilibrium, but there is a stable equilibrium with a positive level of cooperation.

If the replicator dynamic is sufficiently strong compared to vertical and oblique transmission (i.e., if σ is small), (12) will fail and only a low level of cooperation can be sustained in a equilibrium.

This model includes no mechanism form punishing defectors, so a high level of cooperation occurs only when virtually all members carry the altruistic norm. However most social groups that rely on cooperation have forms of punishment of defectors that induce even selfish agents to cooperate.[8]

Suppose that by bearing a cost $w > 0$, an agent can inflict a punishment $c_p > 0$ on a defector. Suppose now B-type individuals are selfish, while A-type individuals cooperate unconditionally and punish defectors, provided the threat of punishment leads defectors to cooperate. If punishment cannot deter defectors, then A-type neither cooperate nor punish. The experimental literature supports the existence of such behavior for humans, as well as the ability of such agents to induce cooperation in public goods games (Fehr and Gächter 2000). I will call A-types *strong reciprocators* (Gintis 2000a,b).

Suppose that while defectors are always detected, a certain fraction $\beta > 0$ of cooperators appear to have defected although they have not. If α is the fraction of strong reciprocators, $n(1-\alpha)$ individuals defect, and $n\alpha\beta$ cooperate but are treated as defectors. The total number of 'violators' to be punished is then $n(1-\alpha(1-\beta))$. The total harm inflicted on real and perceived defectors is $n\alpha c_p$, so the harm per defector imposed by strong reciprocators is $\alpha c_p/(1-\alpha(1-\beta))$. The cost of cooperating in the one-shot game is now $c + \beta\alpha c_p/(1-\alpha(1-\beta))$, while the cost of defecting is $\alpha c_p/(1-\alpha(1-\beta))$. The net gain from defecting is $\alpha c_p(1-\beta)/(1-\alpha(1-\beta)) - c$, so full cooperation is a Nash equilibrium in the one-shot game if

$$\alpha \geq \alpha_{\min} = \frac{c}{(c_p + c)(1-\beta)}. \tag{13}$$

If $\alpha < \alpha_{\min}$, punishment will not deter defectors, so strong reciprocators will neither punish nor cooperate, and universal defection will obtain.

The cost of cooperation is now frequency-dependent, with

$$s(\alpha) = \begin{cases} 0 & \alpha < \alpha_{\min} \\ w(1-(1-\beta)^{n-1}) & \alpha \geq \alpha_{\min} \end{cases} \tag{14}$$

The dynamics of the system are now given by

$$\dot{\alpha} = h(\alpha), \tag{15}$$

but now for $\alpha < \alpha_{\min}$ we have

$$h(\alpha) = (1-\alpha)(\sigma(\alpha(\gamma - \nu) + \nu)), \tag{16}$$

while for $\alpha \geq \alpha_{\min}$, $h(\alpha)$ is given by (5) with $s = w(1-(1-\beta)^{n-1})$.

[8]For evidence in animal behavior, see Clutton-Brock and Parker (1995). For eusocial insects, see Gadagkar (1991) and Frank (1995). For cooperation among cells in multicellular organisms, see Maynard Smith and Szathmary (1997), Keller (1999), and Michod (1999). For human societies, see Weissing and Ostrom (1991), Ostrom, Walker and Gardner (1992), Boyd and Richerson (1992), Gintis (2000b), Fehr and Gächter (2000), Henrich and Boyd (2001), and Henrich, Boyd, Bowles, Camerer, Fehr, Gintis and McElreath (2001).

We cannot use Theorem 1 to analyze the behavior of this cultural system, since s is now frequency dependent. We give a direct argument, assuming $0 < \nu \leq \gamma < 1$, and $0 < \alpha_{\min} < 1$. When $s = 0$, clearly $h(\alpha) > 0$ for $\alpha \in [0, 1)$. Therefore the fraction of strong reciprocators increases for $\alpha < \alpha_{\min}$. The equilibrium $\alpha = 1$ is easily seen to be stable for $s = w(1 - (1 - \beta)^{n-1}) < s_{\max}$, where $s_{\max}$ is given by (7). If β or w is small, then we have a stable equilibrium with full cooperation and the whole population consisting of strong reciprocators.

In the case of very costly punishment, where $w(1 - (1 - \beta)^{n-1}) > s_{\max}$, the equilibrium $\alpha = 1$ is unstable, but we know from Theorem 1 that in this case, if $s = w(1 - (1 - \beta)^{n-1})$ for all $\alpha \in [0, 1]$, there is a globally stable equilibrium at some $\alpha^* \in (0, 1)$. If $\alpha^* > \alpha_{\min}$, then clearly $\alpha = \alpha^*$ is also a globally stable equilibrium of the current system. If the opposite inequality holds, then $\alpha = \alpha_{\min}$ is a globally stable equilibrium. In both cases the system achieves full cooperation with only a portion of the population (α^* or $\alpha_{\min}$) carrying the disadvantage norm (in this case, strong reciprocity). Indeed, the fraction of strong reciprocators may even be quite small, provided the cost of being punished, c_p, is large compared to c, the cost of cooperating, and vertical and oblique transmission is weak compared to the replicator dynamic (σ small). Laboratory experiments indicate that about half of subjects do punish defectors (Fehr and Gächter 2000) in modern societies, though the frequency is quite variable in simpler societies (Henrich et al. 2001).

6 Conclusion

Humans cooperate in situations of anonymity, and in situations where the probability of future interaction is very low. Altruistic behavior occurs even when interactions are not repeated, which implies contributors cannot in any way expect to be repaid in the future for their current sacrifices. For instance, victims of crime spend time and energy ensuring that the perpetrators are apprehended and receive harsh sentences, and jilted lovers retaliate a great personal cost. In addition, people participate in movements for democratic rights and civil liberties in authoritarian states, often only once in their lives, and often at great personal cost. Moreover, a variety of controlled experiments indicate that many people behave prosocially even when the social interaction is clearly one-shot and anonymous (Roth, Prasnikar, Okuno-Fujiwara and Zamir 1991, Fehr and Gächter 2000, Henrich et al. 2001).

This paper models altruism by assuming that both socialization (vertical and oblique transmission) and imitation of successful behavior (replicator dynamics) are operative. This model is attractive in that it is based on well-known and undeniable social forces (socialization and imitation), it is analytical tractable, and it is easy to extend to a variety of situations. For instance, in place of the public goods

game, virtually any game in which the Nash equilibrium based on selfish player is suboptimal for the group (e.g., a prisoner's dilemma or a trust game) can be substituted. In these cases $s = s(\alpha)$ will be frequency dependent. In addition, the assumption that s is a fitness cost can be replaced by the assumption that s is a utility cost not reflected in reproduction. Varying these and related assumptions does not change the conclusion that cooperation can be sustained in a equilibrium provided the power of socialization forces is sufficiently great in comparison with the replicator dynamic.

One plausible critique of the class of models developed in this paper is that a 'mutant' individual who does *not* internalize norms, and hence relies on the replicator dynamic alone in choice of norms, will be more fit than the altruistic internalizers. We deal with this and related problems in a full gene-culture coevolutionary model (Gintis 2001), in which we show that if there are two norms that are internalized, one that is fitness enhancing and the other that is altruistic, the altruistic norm can 'hitchhike' on the fitness-enhancing norm, so again we can have altruism in equilibrium. This approach to solving the problem of prosociality was first suggested by Herbert Simon (1990), one of the few behavioral scientists of the Twentieth Century who truly transcended disciplinary boundaries.

References

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

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

Barry, Adam D., *The Rise of a Gay and Lesbian Movement* (New York: Prentice Hall, 1995).

Binmore, Ken, *Game Theory and the Social Contract: Just Playing* (Cambridge, MA: MIT Press, 1998).

Boorman, Scott A. and Paul Levitt, *The Genetics of Altruism* (New York: Academic Press, 1980).

Bowles, Samuel, "Individual Interactions, Group Conflicts, and the Evolution of Preferences," in Steven N. Durlauf and H. Peyton Young (eds.) *Social Dynamics* (Cambridge, MA: MIT Press, 2001) pp. 155–190.

— and Astrid Hopfensitz, "The Co-evolution of Individual Behaviors and Social Institutions," 2000. Santa Fe Institute Working Paper #00-12-073.

— and Herbert Gintis, "The Moral Economy of Community: Structured Populations and the Evolution of Prosocial Norms," *Evolution & Human Behavior* 19,1 (January 1998):3–25.

— , Robert Boyd, Herbert Gintis, and Peter J. Richerson, "Inter-demic Group Selection Can Lead to the Evolution of Group Beneficial Punishment in Large Groups," 2001. Working Paper.

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

— and — , "An Evolutionary Model of Social Learning: the Effects of Spatial and Temporal Variation," in T. R. Zentall and G. Galef Jr. (eds.) *Social Learning: Psychological and Biological Perspectives* (Hillsdale NY: Erlbaum, 1988) pp. 29–48.

— and — , "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups," *Ethology and Sociobiology* 113 (1992):171–195.

Button, James W., *Blacks and Social Change: Impact of the Civil Rights Movement in Southern Communities* (Princeton, NJ: Princeton University Press, 1989).

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

Chong, Dennis, *Collective Action and the Civil Rights Movement* (Chicago, IL: University of Chicago Press, 1991).

Clutton-Brock, T. H. and G. A. Parker, "Punishment in Animal Societies," *Nature* 373 (1995):58–60.

Coleman, James S., *Foundations of Social Theory* (Cambridge, MA: Belknap, 1990).

Cox, Cece, *One Million Strong: the 1993 March on Washington for Lesbian, Gay, and Bi Equal Rights* (Boston: Alyson, 1993).

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

Durkheim, Emile, *Suicide, a Study in Sociology* (New York: Free Press, 1951). Translated by John A. Spaulding and George Simpson. Edited, with an Introduction by George Simpson.

Epstein, Joshua and Robert Axtell, *Sugarscape* (Cambridge, MA: MIT Press, 1997).

Fehr, Ernst and Simon Gächter, "Cooperation and Punishment," *American Economic Review* 90,4 (September 2000).

Feldman, Marcus W. and Lev A. Zhivotovsky, "Gene-Culture Coevolution: Toward a General Theory of Vertical Transmission," *Proceedings of the National Academy of Sciences* 89 (December 1992):11935–11938.

— , Luca L. Cavalli-Sforza, and Joel R. Peck, "Gene-Culture Coevolution: Models for the Evolution of Altruism with Cultural Transmission," *Proceedings of the National Academy of Sciences* 82 (1985):5814–5818.

Frank, Steven A, "Mutual Policing and Repression of Competition in the Evolution of Cooperative Groups," *Nature* 377 (October 1995):520–522.

Fudenberg, Drew and Eric Maskin, "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* 54,3 (May 1986):533–554.

Gadagkar, Raghavendra, "On Testing the Role of Genetic Asymmetries Created by Haplodiploidy in the Evolution of Eusociality in the Hymenoptera," *Journal of Genetics* 70,1 (April 1991):1–31.

Gintis, Herbert, "Welfare Economics and Individual Development: A Reply to Talcott Parsons," *Quarterly Journal of Economics* 89,2 (June 1975):291–302.

— , "Theory, Practice, and the Tools of Communicative Discourse," *Socialist Review* 50 (March–June 1980):189–232.

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

— , "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology* 206 (2000):169–179.

— , "The Hitchhiker's Guide to Altruism: Genes and Culture, and the Internalization of Norms," 2001. Santa Fe Institute Working Paper.

Grusec, Joan E. and Leon Kuczynski, *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory* (New York: John Wily & Sons, 1997).

Heckathorn, Douglas, "Dynamics and Dilemmas of Collective Action," *American Sociological Review* 61 (1996):250–278.

— , "Collective Action, Social Dilemmas, and Ideology," *American Sociological Review* 10,4 (1998):451–479.

Henrich, Joseph and Robert Boyd, "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas," *Journal of Theoretical Biology* 208 (2001):79–89.

— , — , Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath, "Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies," *American Economic Review* 91 (May 2001):73–78.

Hirshleifer, David and Eric Rasmusen, "Cooperation in a Repeated Prisoners' Dilemma with Ostracism," *Journal of Economic Behavior and Organization* 12 (1989):87–106.

Keller, Laurent, *Levels of Selection in Evolution* (Princeton, NJ: Princeton University Press, 1999).

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

Mackie, Gerry, "Ending Footbinding and Infibulation: A Convention Account," *American Sociological Review* 61 (December 1996):999–1017.

Maynard Smith, John, "Group Selection," *Quarterly Review of Biology* 51 (1976):277–283.

— and Eors Szathmary, *The Major Transitions in Evolution* (Oxford: Oxford University Press, 1997).

Michod, Richard E., *Darwinian Dynamics* (Princeton, NJ: Princeton University Press, 1999).

Moore, Jr., Barrington, *Injustice: The Social Bases of Obedience and Revolt* (White Plains: M. E. Sharpe, 1978).

Nowak, Martin A. and Karl Sigmund, "Evolution of Indirect Reciprocity by Image Scoring," *Nature* 393 (1998):573–577.

— , Karen M. Page, and Karl Sigmund, "Fairness Versus Reason in the Ultimatum Game," *Science* 289 (8 September 2000):1773–1775.

Ostrom, Elinor, James Walker, and Roy Gardner, "Covenants with and without a Sword: Self-Governance Is Possible," *American Political Science Review* 86,2 (June 1992):404–417.

Parsons, Talcott, *Sociological Theory and Modern Society* (New York: Free Press, 1967).

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

Simon, Herbert, "A Mechanism for Social Selection and Successful Altruism," *Science* 250 (1990):1665–1668.

Skocpol, Theda, *States and Social Revolutions* (Cambridge, UK: Cambridge University Press, 1979).

Skyrms, Brian, *Evolution of the Social Contract* (Cambridge: Cambridge University Press, 1996).

Sperber, Daniel, *Explaining Culture: A Naturalistic Approach* (New York: Blackwell, 1996).

Staddon, J. E. R., *Adaptive Dynamics: The Theoretical Analysis of Behavior* (Cambridge: MIT Press, 2001).

Taylor, P. and L. Jonker, "Evolutionarily Stable Strategies and Game Dynamics," *Mathematical Biosciences* 40 (1978):145–156.

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

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

Weissing, Franz and Elinor Ostrom, "Irrigation Institutions and the Games Irrigators Play: Rule Enforcement without Guards," in Reinhard Selten (ed.) *Game Equilibrium Models II: Methods Morals and Markets* (Berlin: Springer-Verlag, 1991) pp. 188–262.

Williams, G. C., *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton, NJ: Princeton University Press, 1966).

Wrong, Dennis H., "The Oversocialized Conception of Man in Modern Sociology," *American Sociological Review* 26 (April 1961):183–193.

Wrong, Dennis W., *The Oversocialized Conception of Man in Modern Sociology* (New York: Irvington, 1993).

Young, H. Peyton, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton, NJ: Princeton University Press, 1998).