SOLVING THE PUZZLE OF PROSOCIALITY

Herbert Gintis

ABSTRACT

Homo sapiens is the only species in which we observe extensive cooperation among large numbers of genetically unrelated individuals. Incompatible approaches to explaining cooperation among humans have been offered by sociologists, biologists, and economists. None is wholly successful. Each discipline, moreover, has ignored basic insights of the others. This article explains cooperation by combining central contributions of these disciplines, developing a model of cultural evolution in which we use (a) the sociological concept of the internalization of norms to explain cultural transmission; (b) the biological concepts of vertical and oblique transmission to model the interaction of cultural and biological adaptation; and (c) the economic concepts of rational action and the replicator dynamic to model the interaction between self-interested and altruistic behavior. The article closes with a bioeconomic explanation of the human capacity to internalize norms.

KEY WORDS • altruism • conditional cooperation • conditional punishment • cultural dynamics • cultural evolution • strong reciprocity

1. Introduction

Homo sapiens is the only species in which we observe extensive cooperation among large numbers of genetically unrelated individuals. Incompatible approaches to explaining cooperation among humans have been offered by sociologists, biologists, and economists.[1] None is wholly successful. Each discipline, moreover, has ignored basic insights of the others. This article explains cooperation by combining central contributions of these disciplines, developing a model of cultural evolution in which we use (a) the sociological concept of the internalization of norms to explain cultural transmission; (b) the biological concepts of vertical and oblique transmission to model the interaction of cultural and biological adaptation; and (c) the economic concepts of rational action and the replicator

Rationality and Society Copyright © 2003 Sage Publications (London, Thousand Oaks, CA and New Delhi), Vol. 15(2): 155–187. [1043–4631(200305)15:2; 155–187; 032122]

dynamic (imitating the successful) to model the interaction between self-interested and altruistic behavior. The article closes with a bio-economic explanation of the human capacity to internalize norms.

Individuals often do better by coordinating and sharing the benefits of their activities rather than each acting alone. The benefit accruing to the group from each individual's cooperation in such cases is greater than the cost to the individual, but nonetheless each individual would be better off not incurring the costs of co-operation and simply benefiting from the efforts of the other group members. If all participants follow this self-interested logic, however, cooperation will fail. When it is maintained, cooperation is *altruistic*, in the sense of being group-beneficial, but personally costly. Why are such altruistic behaviors not driven out by self-interested agents? This is the *puzzle of prosociality*.

The distinctive contribution of sociological theory to solving the puzzle of prosociality is the insight that altruism is possible if, and to the extent that, the norm of contributing to the group and punishing non-contributors is acted upon by a sufficiently large fraction of social participants. Indeed, a key tenet of socialization theory is that a society's values are passed from generation to generation through the *internalization of norms* (Benedict 1934; Durkheim 1951; Mead 1963; Parsons 1967; Grusec and Kuczynski 1997). In the language of rational action, internalized norms are accepted not as instruments towards and constraints upon achieving other ends, but rather as *arguments in the preference function that the individual maximizes*.[2] A variety of prosocial emotions then come into play, including prominently *shame*, *guilt*, and *empathy*, directly biasing individual choices in prosocial directions.

The human capacity to internalize norms, which consists in an older generation instilling the values and objectives of a younger generation through an extended series of personal interactions, relying on a complex interplay of affect and authority, is based on a distinctive psychological predisposition. This capacity is not recognized in biology and economics.

Economic theory takes preferences (and hence culture) as exogenously given, and its default condition is that individuals are *self-regarding*, acting to maximize their personal payoffs from participating in social situations. Economists summarily reject the suggestion that one could solve the puzzle of prosociality by positing that 'cooperating with others' is an argument in agents' objective functions. Yet this is precisely the position taken in this article.

The biological literature admits the endogeneity of culture (Cavalli-Sforza and Feldman 1981; Bonner 1984; Boyd and Richerson 1985), but treats culture as *conventional* (e.g. what side of the road to drive on, what vocalization is associated with what state of affairs, what rituals one performs when a relative dies, what foods one may and may not eat) or *informational* (e.g. how one best tills a certain type of land, how one best wards off disease), rather than *fundamentally normative* (how we should behave, in principle). To explain cultural transmission, biological models generally assume that recipients either directly *perceive* the utility of the conventions or information passed to them, or find it prudent to accept the opinions of informed others rather than engage in innumerable costly personal experiments in deciding which behavior has higher payoff (Conlisk 1988; Henrich and Boyd 2001). Since altruistic norms are generally neither conventional nor informational, biologists have no explanation for the transmission of altruistic norms.

The distinctive contribution of biology to solving the puzzle of prosociality is the development of evolutionary models of cultural transmission by analogy with population biology and epidemiology (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985), using quantitative mathematical constructs involving *vertical transmission* (from parents to children), *oblique transmission* (through socialization institutions, including secular and religious rituals, schools, and communications media), and *horizontal transmission* (from peer interactions).

The distinctive contribution of economics to understanding cooperation is the treatment of agents as maximizing a preference function by choosing from a set of available actions (rational action). Sociologists have widely rejected this model, probably because it conflicts fundamentally with socialization theory, which sociologists do, but economists do not, consider to be empirically well supported. Like economists, biologists have maintained that the theory of self-interested optimizing behavior, properly applied, can explain all important aspects of human behavior (Williams 1966; Dawkins 1976; Alexander 1987).

This article presents an evolutionary game-theoretic model of altruism in which vertical and oblique cultural transmission are justified in terms of the internalization of norms, thus allowing altruistic as well as conventional and informational norms to be transmitted. We also assume individuals optimize by abandoning norms when they find alternatives that provide higher payoffs (the

replicator dynamic). This corrects the 'oversocialized' concept of the individual that is found in some prominent versions of socialization theory (Wrong 1961; Gintis 1975).

For analytical specificity, this article studies the dynamics of a single altruistic norm that has a payoff disadvantage for those who adopt it, but is transmitted vertically by parents and obliquely through socialization institutions. We allow altruism to be either beneficial or harmful to the group, and we admit four types of cultural change.

- Individuals mate and have offspring. Families who use lower payoff strategies have fewer offspring (biologically adaptive dynamics).
- Families pass on their cultural traits, self-interested or altruistic, to their offspring (vertical transmission) through internalization.
- A fraction of self-interested offspring are induced to adopt altruistic norms by socialization institutions (oblique transmission).
- Some of the resulting population change their cultural values to conform to the behavior of other individuals who have higher payoffs (replicator dynamics).

This model can be extended trivially to the case of many cultural norms as long as their costs and benefits are additive. Where there are non-linear interactions among norms, a more subtle analysis is needed.

The following are examples of cultural forms that are altruistic in the above sense, and to which our analysis applies:

- **Altruistic Cooperation.** Personally costly behavior that benefits others in the group. This includes being trustworthy, being a fearless warrior, strenuously hunting game that will be shared equally among all group members, cooperating in team production, and not overexploiting a common pool resource. In a more modern setting, classical altruism includes willingness to vote and otherwise participate in the political life of the community, giving anonymously to charity, honestly paying taxes, acting on behalf of one's ethnic, racial, or religious group, and identifying with the goals of an organization of which one is a member.
- **Altruistic Punishment.** Punishing individuals who violate a social norm at a cost to oneself.

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

The first two of these are strongly prosocial in that they enhance the fitness of other group members (positive altruism), whereas the last two are antisocial (negative altruism). Thus the maintenance of altruistic norms can be either fitness enhancing or fitness reducing for the group as a whole. In fact, most altruistic norms promulgated in most successful societies are prosocial (Brown 1991), but negatively altruistic norms are not difficult to find (Edgerton 1992). In the remainder of this article, 'altruism' refers to 'positive altruism', unless explicitly stated otherwise.

Our model yields two general conclusions.

- In the absence of oblique transmission of the altruistic norm, altruism is driven out by self-interested behavior. When oblique transmission of altruism is present, a positive frequency of altruism can persist in cultural equilibrium.
- A high level of cooperation can be sustained in cultural equilibrium by the presence of a minority of agents who adopt the altruistic norm of what we call *strong reciprocity*: cooperating unconditionally and punishing defectors at a personal cost, the remaining agent being self-interested.

The first assertion states what might be called the 'The Fundamental Theorem of Sociology': *extra-familial socialization institutions are necessary to support altruistic forms of prosociality*. The second assertion expresses the insight that cooperation is robustly stable when antisocial behavior is punished by the individual, voluntary, and largely decentralized, behavior of group members.

2. The Prerequisites of a Model of Human Cooperation

The model developed below (a) does not assume that agents are self-interested; (b) does not derive cooperative behavior from repeated

interactions; and (c) does not depend on genetic group selection. The admission of non-self-interested behavior is based on empirical evidence, as presented in Section 3. We assume one-shot interaction because in many real-life situations people cooperate even when repetition is unlikely, as in such collective actions as mass demonstrations for civil liberties, democracy, and racial or gender equality, and because the experimental evidence demonstrates that people regularly cooperate even in anonymous, one-shot situations (Section 3).

Moreover, it has proven difficult to develop a plausible model of cooperation in large groups using repetition alone. Axelrod and Hamilton (1981) and others have shown that repetition alone is sufficient in two-person groups, provided agents are sufficiently future-oriented. However, their argument does not extend to larger groups (Boyd and Richerson 1988). One can generate Nash equilibria with high levels of cooperation in large groups if agents are sufficiently far-sighted (Fudenberg and Maskin 1986), but such equilibria have very poor dynamic stability properties and depend on agents being implausibly far-sighted and patient (Gintis 2000a).

Second, in the primitive conditions under which human sociality evolved, when a group was threatened with extinction or dispersal, say through war, pestilence, or famine, cooperation was most needed for survival. But since the probability that the group will dissolve increases sharply under such conditions, cooperation based on future reciprocation cannot be maintained. Thus, *precisely when a group is most in need of prosocial behavior, cooperation based on repeated interactions will collapse*.

Such critical periods were probably common in the evolutionary history of *Homo sapiens*.[3] Gintis (2000b) shows that a small number of *strong reciprocators*, who punish defectors *without regard for future reward*, can dramatically improve the survival chances of human groups. An argument using the tools of population biology (specifically, Price's equation) then shows that under these conditions strong reciprocators can invade a population of self-interested types, and can persist in equilibrium, thus stabilizing cooperation in large groups (Gintis 2000b).

By positing that individuals internalize norms, we derive altruistic behavior without requiring inter-group competition of the sort needed by group selection models.[4] This is desirable because biologists have shown that genetic group selection is very difficult to

sustain unless genetic relatedness among group members is high (Hamilton 1963; Boorman and Levitt 1980; Maynard Smith 1998).[5]

In a related article, Gintis (2003) shows that a gene-culture coevolutionary model can account for the evolutionary stability of genes for internalization.[6] Moreover, a cultural group selection argument is needed to explain why norms are generally prosocial: competition among social groups will strongly favor those whose cultural systems are dominated by prosocial norms (Parsons 1964; Boyd and Richerson 1990; Soltis et al. 1995; Gintis 2003).

3. Strong Reciprocity: The Behavioral Evidence

There are many civic-minded acts that are difficult to explain by self-interest. These include voting, giving anonymously to charity, participating in collective actions, and sacrificing oneself in battle. More mundanely, victims of crime often spend time and energy ensuring that the perpetrators are apprehended and receive harsh sentences, and jilted lovers retaliate at great personal cost. In modeling these behaviors, a suggestive body of evidence points to a schema that may be termed *strong reciprocity* (Gintis 2000a). A strong reciprocator comes to a new social situation with a predisposition to cooperate, is predisposed to respond to cooperative behavior on the part of others by maintaining or increasing his level of cooperation, and responds to free-riding behavior on the part of others by retaliating against the offenders, even at a cost to himself, and even when he cannot not reasonably expect future personal gains from such retaliation. The strong reciprocator is thus both a *conditionally altruistic cooperator* and a *conditionally altruistic punisher*. We call this 'strong' reciprocity to distinguish it from reciprocal altruism (Trivers 1971), indirect reciprocity (Alexander 1987; Nowak and Sigmund 1998), and other forms of reciprocity that require repeated interactions and can be explained using models of self-interested behavior (Axelrod and Hamilton 1981; Fudenberg and Maskin 1986).

The observations of reciprocal behavior in everyday life are insufficient grounds for accepting strong reciprocity as a model of human behavior, however, because (a) anonymity is rarely achieved (for instance, even ostensibly anonymous charity donations are known to, or could be discovered by, family members); (b) subjects could believe their actions were part of a set of repeated interactions;

and (c) we have no reliable measure of the frequency of what appear to be strongly reciprocal acts. These defects can be remedied by observing human behavior in an experimental game-theoretic setting. Several examples follow. In all cases, the subjects are anonymous, the payoffs are in real money, subjects are not misled by experimenters, and unless otherwise noted interactions are one-shot rather than repeated.

3.1. Strong Reciprocity in an Experimental Labor Market

The experimenters (Fehr et al. 1997) divided a group of 141 subjects (college students who had agreed to participate in order to earn money) into a set of 'employers' and a set of 'employees.' The rules of the game are as follows. If an employer hires an employee and pays wage $w \in [0, 100]$, his profit is $\pi = 100e - w$, where $e \in [0.1, 1]$ is the amount of 'effort' exerted by the employee. The payoff to the employee is then $u = w - c(e)$, where $c(e)$ is the 'cost of effort' function shown in Figure 1. All payoffs involve real money that the subjects are paid at the end of the experimental session.

The employer then offers a 'contract' specifying a wage w and a desired amount of effort e^*. A contract is made with the first employee who agrees to these terms. An employer can make a

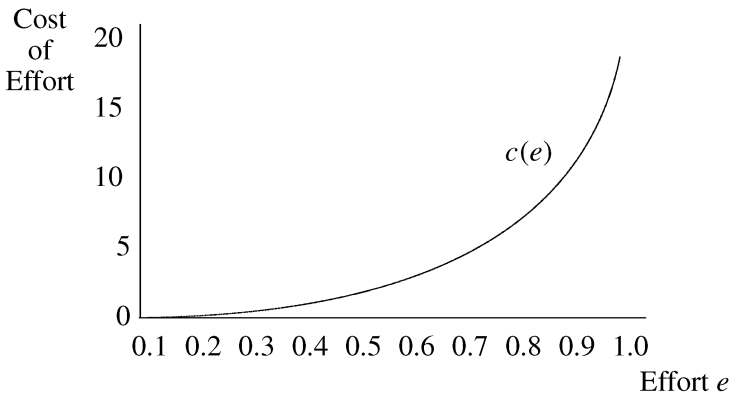

Figure 1. The Cost of Effort Schedule in Fehr et al. (1997)

contract (w, e^*) with at most one employee. The employee who agrees to these terms receives the wage w and supplies an effort level e, which *need not equal the contracted effort*, e^*. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0.1, 1]$, with impunity. Although subjects may play this game several times, each employer–employee interaction is a one-shot (non-repeated) event.

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

In fact, however, this self-interested outcome rarely occurred in this experiment.[7] The average net payoff to employees was $u = 35$, and the higher the employer's choice of demanded effort, the more *both* employers and employees earned. In effect, employers presumed the strong reciprocity predispositions of the employees, making more generous wage offers and receiving higher effort, as a means of increasing both their own and the employee's payoff, as depicted in Figure 2.

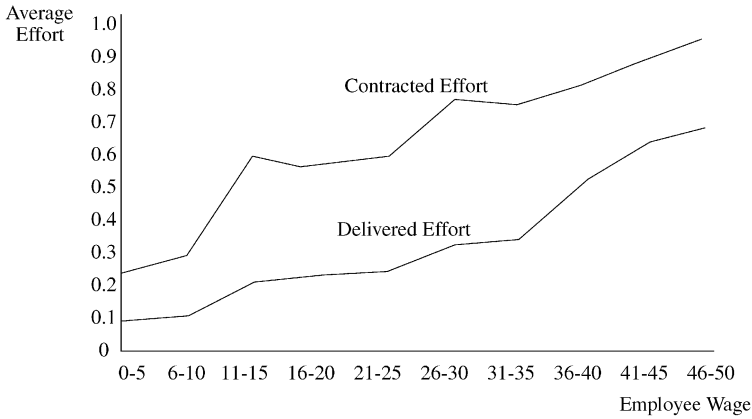

Figure 2. Relation of Contracted and Delivered Effort to Worker Wage (141 subjects). From Fehr et al. (1997)

Figure 2 also shows that, though most employees are strong reciprocators, at any wage rate there is still a significant gap between the amount of effort agreed upon and the amount actually delivered. This is not because there are a few 'bad apples' among the set of employees, because in fact only 26% of employees delivered the level of effort they promised! We conclude that, according to this experiment, at least, even strong reciprocators are inclined to compromise their morality to some extent.

The above evidence is compatible with the notion that the employers were purely self-interested, since their beneficent behavior *vis-à-vis* their employees was effective in increasing employer profits. To see if employers were also strong reciprocators, following this round of experiments, the authors extended the game by allowing the employers to respond reciprocally to the *actual effort choices* of their workers. At a cost of 1, an employer could *increase* or *decrease* his employee's payoff by 2.5. If employers were self-interested, they would of course do neither, since they do not interact with the same worker a second time. However, 68% of the time employers punished employees who did not fulfil their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded 44% of employees who *exactly* fulfilled their contracts. Moreover, employees *expected* this behavior on the part of their employers, as shown by the fact that their effort levels *increased significantly* when their bosses gained the power to punish and reward them. Underfulfilled contracts dropped from 86% to 26% of the exchanges, and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in the net payoffs to all subjects, even when the payoff reductions resulting from employer punishment of employees are taken into account.

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

3.2. Strong Reciprocity in the Public Goods Game

The *public goods game* is designed to illuminate such problems as the voluntary payment of taxes and contribution to team and community goals (Ledyard 1995). The following is a common variant of the game. Ten subjects are told that $1 will be deposited in each of their 'private accounts' as a reward for participating in each round of the experiment. For every $1 that a subject moves from his 'private account' to the 'public account,' the experimenter will deposit one half dollar in the private accounts of each of the subjects at the end of the game. This process will be repeated 10 times, and at the end the subjects can take home whatever they have in their private accounts.

If all 10 subjects are perfectly cooperative, each puts $1 in the public account at the end of each round, generating a public pool of $10; the experimenter then puts $5 in the private account of each subject. After 10 rounds of this, each subject has $50. However, every $1 a player contributes to the public account, while benefiting others by an amount $4.50, costs the contributor $0.50. Therefore the dominant strategy for a self-interested player is to contribute nothing to the pool. Each subject then earns just $10.

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

Could we not explain the decay of public contribution by *learning*: the participants really do not understand the game at first, but once they hit upon the free-riding strategy, they apply it? One indication that learning does not account for the decay of cooperation is a result obtained by Andreoni (1988). Andreoni finds that when the public goods game is played with several groups, but after every series of rounds group membership is reshuffled, after each reshuffling, subjects begin by contributing about half, but once again cooperation decays as the game progresses. Yet surely subjects did not 'unlearn' the money-maximizing behavior between shuffles! Andreoni (1995) suggests a strong reciprocity explanation for the decay of cooperation: public-spirited contributors want to retaliate against free-riders, and the only way available to them in the game

is by not contributing themselves. Subjects often report this reason for the unraveling of cooperation retrospectively. More compelling, however, is the fact that when subjects are given a more direct way of retaliating against defectors, they use it in a way that helps sustain cooperation (Dawes et al. 1986; Sato 1987; Yamagishi 1988a, b, 1992).

These studies do not completely rule out the possibility that retaliation is *strategic*, however, since in the above studies the subjects remained together for several rounds of play. Fehr and Gächter (2000, 2002) remedied this defect, setting up a repeated public goods game with the possibility of costly retaliation, but they ensured that group composition changed *in every period* so subjects knew that costly retaliation could not confer any pecuniary benefit to those who punish. Nonetheless, punishment of free-riding was prevalent and gave rise to a large and sustainable increase in cooperation levels.

Fehr and Gächter (2000) used six- and ten-round public goods games with groups of size four, and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. They ran between 10 and 18 groups simultaneously. Under the *Partner* treatment, the 4 subjects remained in the same group for all 10 periods. Under the *Stranger* treatment, the subjects were randomly reassigned after each round. Finally, under the *Perfect Stranger* treatment the subjects were randomly reassigned and assured that they would never meet the same subject more than once. Subjects earned an average of about $35 for an experimental session.

Fehr and Gächter ran the experiment for 10 rounds with punishment and 10 rounds without. Their results are illustrated in Figure 3. We see that *when costly punishment is permitted, cooperation does not deteriorate*, and in the Partner treatment, despite strict anonymity, *cooperation increases almost to full cooperation, even on the final round*. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games.

3.3. The Ultimatum Game

In the *ultimatum game*, under conditions of anonymity, two players are shown a sum of money, say $10. One of the players, called the 'proposer', is instructed to offer any number of dollars, from $1 to

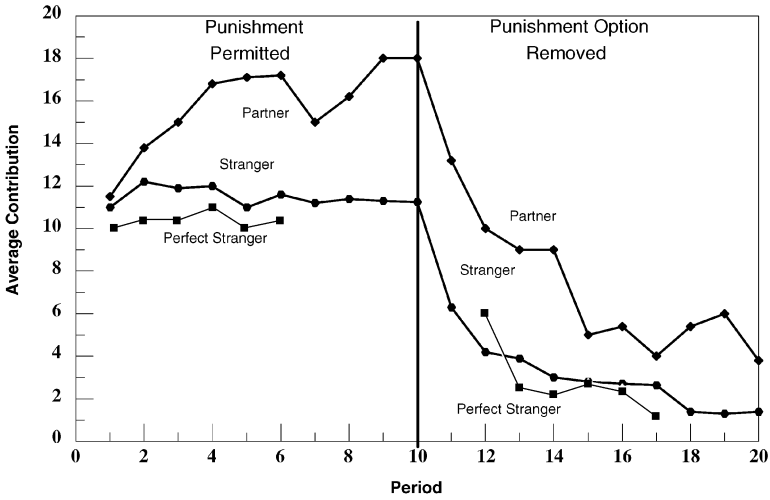

Figure 3. Average Contributions over Time in the Partner, Stranger, and Perfect Stranger Treatments When the Punishment Condition Is Played First (Fehr and Gächter 2000).

$10, to the second player, who is called the 'responder'. The responder, again under conditions of anonymity, can either accept the offer or reject it. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

Since the game is played only once and the players do not know each other's identity, a self-interested responder will accept *any positive amount of money*.

Knowing this, a self-interested proposer will offer the minimum possible amount, $1, and this will be accepted. However, when actually played, *the self-interested outcome is almost never attained or even approximated*. In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts, 50% of the total generally being the modal offer. Respondents frequently reject offers below 30% (Gintis 2000a).

Economists were at first astonished at this behavior. Why would people reject a positive amount of money? They suggested that perhaps players did not understand the game. But by changing the rules a bit experimenters showed that this was not the case. For

instance, if the game is changed so that the offer is generated by a computer (and the responder is told this fact), the rejection rate becomes very low, however small a share is offered to the responder (Blount 1995). Similarly, if the game is changed so that if the responder rejects the proposer's offer he receives zero payoff, but the proposer still receives the share he proposed for himself, responders almost never reject offers.

The ultimatum game reveals strong reciprocity on the part of the respondent, since the game is not repeated, and since the respondent's rejection of low offers is contingent upon the proposer having acted unfairly, and upon being able to punish the proposer. Moreover, to the extent that proposers make offers that are larger than necessary to induce assent on the part of the respondent, proposers also exhibit strong reciprocity in the form of a predisposition to cooperate.

3.4. The Cultural Variability of Cooperative Behavior

How generalizable are these results? Most ultimatum game experiments use college students as subjects. The fact that the results are similar around the world could either be because strong reciprocity is a uniformly expressed human behavior, or because college students have very similar cultures around the world. If the latter, which economic and social conditions are involved? Is strong reciprocity better explained statistically by individual attributes such as sex, age, or wealth, or by the attributes of the group to which the individuals belong? Are there cultures that approximate the self-interested actor of standard economics and biology? To answer these questions, I and a group of colleagues undertook a large crosscultural study of behavior in ultimatum and public goods games (Henrich et al. 2001). Twelve experienced anthropological field researchers, working in 12 countries on 4 continents, recruited subjects from 15 small-scale societies exhibiting a wide variety of economic and cultural conditions. These societies consisted of three foraging groups (the Hadza of East Africa, the Au and Gnau of Papua New Guinea, and the Lamalera of Indonesia), six slash-and-burn horticulturists (the Aché, Machiguenga, Quichua, and Achuar of South America, and the Tsimané and Orma of East Africa), four nomadic herding groups (the Turguud, Mongols, and Kazakhs of Central Asia, and the Sangu of East Africa) and two

sedentary, small-scale agricultural societies (the Mapuche of South America and Zimbabwe farmers in Africa).

The results of this study can be summarized as follows. First, the self-interested actor model is not supported in any society studied. Second, there is considerably more behavioral variability across groups than has been found in previous crosscultural research and the self-interested actor model fails in a wider variety of ways than in previous experiments. Third, group-level differences in economic organization and the degree of market integration explain a substantial portion (more than 60%) of the behavioral variation across societies: the higher the *degree of market integration* and the higher the *payoffs to cooperation* in production, the greater the level of cooperation in experimental games. Fourth, individual-level economic and demographic variables do not explain behavior either within or across groups. Fifth, *behavior in the experiments is generally consistent with economic patterns of everyday life in these societies*.

To illustrate this fifth point, which is critical for our argument, consider two Papua New Guinea groups, the Au and the Gnau. These groups are neighboring forager-horticulturalists who have a culture of competitive gift-giving. In the ultimatum game, about one fifth of proposers made *hyperfair offers* of 60% or more of the pie to responders. Moreover, *nearly half of the hyperfair offers were rejected by responders*! These results, never before seen in ultimatum game experiments, reflect the everyday life experience of self-aggrandizement through making generous offers, and self-abasement by accepting such offers.[8]

While much more research must be done before anything conclusive can be drawn from such studies, we can tentatively suggest that strong reciprocity is a virtually universal behavioral template, but that its expression is strongly affected by the culture of the particular society in which it is expressed. This remains true, moreover, even in contexts of anonymity and non-repeatedness that strongly favor the expression of self-interested behaviors. There is no question but that subjects fully understood the nature of the game they were playing, since each was tested on understanding before being permitted to participate. Therefore the fact that they made generous offers or rejected ungenerous offers (or in the case of the Au and the Gnau, rejected even generous offers), and did so broadly in conformance with the norms and practices of the societies in which they live, is a strong vindication of the sociological model of the

internalization of norms: individuals maximize an objective function that includes, in addition to material payoffs, the norms of fairness prevalent in their societies.

4. A Model of Cultural Evolution

Consider a group in which members can either adopt or fail to adopt a certain cultural norm A. We call those who adopt norm A *altruists*, and those who do not adopt norm A *self-interested* types, or 'B-types'. Altruism is costly, in that self-interested types have fitness 1, as compared with altruists, who have fitness $1 - s$, where $0 < s < 1$.[9] We assume in each period that agents pair off randomly, mate, and have offspring in proportion to their fitness, after which they die (we call this a *biologically adaptive dynamic*). Families pass on their cultural norms to their offspring, so offspring of AA parents are altruists, offspring of BB parents are self-interested, and half of the offspring of AB-families are altruists, the other half self-interested (we call this *vertical transmission*). We also assume that the self-interested offspring of AB- and BB-families are susceptible to influence by community institutions promoting altruistic norms, a fraction of such offspring becoming altruists (we call this *oblique transmission*).

For the first stage, suppose there are n males and n females at the beginning of the period. If the fraction of altruists is α, there will be $n\alpha^2$ AA-families, who will have $n\alpha^2(1-s)^2\beta$ offspring, all of whom are altruists, where we choose β so that the population grows at some rate $g(\alpha)$, depending on the degree of prosociality of norm A. There will also be $2n\alpha(1-\alpha)$ AB-families, who will have $2n\alpha(1-\alpha)(1-s)\beta$ offspring, half of whom are altruists. Finally, there will be $n(1-\alpha)^2$ BB-families who will have $n(1-\alpha)^2\beta$ offspring. Adding up the number of offspring, we see that we must have $\beta = g(\alpha)/(1-s\alpha)^2$. Thus the frequencies of AA, AB, and BB offspring are given by

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

Second, a fraction $\alpha\gamma$ of offspring of AB- and BB-families who are self-interested switch to being altruists under the influence of

the oblique transmission of cultural norm A, where γ is a measure of the strength of the oblique transmission process. Note that we have made the conservative assumption that oblique transmission is proportional to the level of altruism. It is easy to check that the change in the fraction of altruists in the next generation is given by

$$\dot{\alpha} = f(\alpha) = \frac{\alpha(1-\alpha)(\gamma - s)}{1 - s\alpha}. \tag{2}$$

Third, each group member i observes 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, and agents' objective functions do not perfectly track fitness, so it is reasonable to assume that 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 $p = \sigma s$ for some proportionality constant $\sigma > 0$.

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

$$\alpha' = \alpha - \sigma\alpha(1-\alpha)s,$$

which, expressed in differential equation form, is

$$\dot{\alpha} = -\sigma\alpha(1-\alpha)s \tag{3}$$

This is a special case of the *replicator dynamic* in cultural evolution.[10] We now combine these two sources of change in the fraction of altruists, giving

$$\dot{\alpha} = h(\alpha) = f(\alpha) - \sigma\alpha(1-\alpha)s \tag{4}$$

where σ now represents the relative strength of the replicator dynamic, which is biased against the altruistic norm, in comparison with the cultural transmission mechanisms, which may favor this norm. In reduced form, we now have

$$\dot{\alpha} = \frac{\alpha(1-\alpha)}{1 - s\alpha}(\gamma - s - s\sigma)(1 - s\alpha). \tag{5}$$

We call the situation $\dot{\alpha} = 0, \alpha \in [0, 1]$ a *cultural equilibrium* of the dynamical system. We then have

THEOREM 1. *We assume $\gamma \geq 0$ is given and fixed throughout.*

(a) If

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

$\alpha = 1$ is a globally stable altruistic equilibrium.

(b) If

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

both $\alpha = 0$ and $\alpha = 1$ are locally stable equilibria of the system and there is third unstable equilibrium $\alpha^* \in (0,1)$ separating the basins of attraction of the two stable equilibria: both self-interested and altruistic equilibria are stable.

(c) If $s > s_{\max}$, then $\alpha = 0$ is a stable self-interested equilibrium of the system.

Proof: There are three zeros of (5), of which two are $\alpha = 0$ and $\alpha = 1$. The third is $\alpha^* = (s(1+\sigma)-\gamma)/s^2\sigma$. If $s < s_{\min}$, then $h'(0) > 0$, $\alpha^* < 0$, and $h'(1) > 0$, so the unique stable equilibrium is $\alpha = 1$, proving (a). If $s_{\min} < s < s_{\max}$, then $\alpha^* \in (0,1)$, $h'(0)$, $h'(1) < 0$, so both $\alpha = 0$ and $\alpha = 1$ are stable. α^* must then be unstable, proving (b). Finally, if $s > s_{\max}$, $\alpha^* > 1$, $h'(0) < 0$, and $h'(1) > 0$, so $\alpha = 0$ is the only stable cultural equilibrium, proving (c). ■

Theorem 1 implies what might be called the *Fundamental Theorem of Sociology*:

COROLLARY 1.1. *Altruistic norms persist in a cultural equilibrium only if there is a strictly positive rate of cultural transmission of altruism* via *social institutions.*

Proof: If $\gamma = 0$, then $s_{\max} = 0$, so $s > s_{\max}$. By Theorem 6, $\alpha = 0$ is the only stable cultural equilibrium. ■

Theorem 1 shows that the higher the personal cost of altruistic behavior, the more stringent the conditions under which altruism will emerge. This illustrates the power of a theory that models the

tension between prosocial socialization institutions and the psychological mechanism of norm internalization on the one hand, and the replicator dynamic that induces agents to shift to higher payoff behaviors, whatever the effect of these behaviors on others, and on society as a whole, on the other hand. This tension is also revealed in the following:

COROLLARY 1.2. *If the strength of the replicator dynamic σ satisfies*

$$\sigma < \frac{\gamma - s}{s},$$

the altruistic cultural equilibrium is globally stable. If

$$\frac{\gamma - s}{s} < \sigma < \frac{\gamma - s}{s(1 - s)},$$

both the self-interested and the altruistic cultural equilibria are locally stable, and the basin of attraction of the altruistic equilibrium shrinks as ó increases. Finally, if

$$\sigma > \frac{\gamma - s}{s(1 - s)},$$

the self-interested cultural equilibrium is globally stable.

5. Maintaining Cooperation Through Altruistic Punishment

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; Henrich and Boyd 2001; Bowles 2001; Boyd et al. 2001) or repeated interactions with a suitably low rate of discounting future benefits (Fudenberg and Maskin 1986; Hirshleifer and Rasmusen 1989; Bowles and Gintis 1998; 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, and even when c is very large.

Let A be an altruistic trait that induces its bearer to cooperate in the public goods game, while trait B induces its bearer to defect. The fitness deficit of altruistic trait is now $s = c$, so Theorem 1 implies that if

$$c < \frac{\gamma}{1+\sigma}, \tag{8}$$

complete cooperation in the Public Goods Game is a stable cultural equilibrium. If this inequality fails, but we have

$$c \leq \left(1+\sigma-\sqrt{(1+\sigma)^2-4\gamma\sigma}\right)/2\sigma, \tag{9}$$

then full cooperation remains a stable equilibrium, but there is another stable equilibrium with complete defection. If (9) fails, cooperation cannot be sustained.

The model in its current form is unrealistic, however, in that most social groups that rely on cooperation have forms of punishment of defectors that considerably reduce the need for altruism and increase the range of parameters over which high levels of cooperation can be maintained.[11]

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 self-interested, while A-type individuals are strong reciprocators: they cooperate unconditionally and punish defectors, provided the threat of punishment leads self-interested types to cooperate.

If punishment cannot deter defectors, then strong reciprocators neither cooperate nor punish.

Suppose that while defectors are always detected, a certain fraction $\beta > 0$ of cooperators accidentally defect, or appear to defect. 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

$$\alpha \geq \alpha_{\min} = \frac{c}{(c_p+c)(1-\beta)}. \tag{10}$$

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

$$s(\alpha) = \begin{cases} 0 & \alpha < \alpha_{\min} \\ w(1-(1-\beta)^{n-1}) & \alpha \geq \alpha_{\min} \end{cases} \tag{11}$$

The dynamics of the system are now given by

$$\dot{\alpha} = h(\alpha), \tag{12}$$

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

$$h(\alpha) = \alpha(1-\alpha)\gamma, \tag{13}$$

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

We cannot use Theorem 1 to analyze the behavior of this cultural system, since s is now frequency dependent. However, we have the following theorem.

THEOREM 2. *If the fraction α of strong reciprocators obeys (12), then*

$$c_p < \frac{\beta c}{1-\beta}, \tag{14}$$

is necessary and sufficient for there to be a locally stable full cooperation equilibrium. Moreover, assuming (14), the following holds.

(a) If $\gamma > s(1+\sigma(1-\alpha_{\min}s))$, then $\alpha = 1$ is a full cooperation globally stable equilibrium.

(b) If $s(1+\sigma(1-\alpha_{\min}s)) > \gamma > s(1+\sigma(1-s))$, then $\alpha = 1$ is a full cooperation locally stable equilibrium. There is a second full cooperation locally stable equilibrium at $\alpha = \alpha_{\min}$.

(c) If $\gamma < s(1+\sigma(1-s))$, then there is a globally stable full cooperation equilibrium at $\alpha = \alpha^* = (s(1+\sigma)-\gamma)/s^2\sigma$.

Proof: Note first that when $s = 0$, $h(\alpha) > 0$ for $\alpha \in [0, 1)$. Therefore the fraction of strong reciprocators increases for $\alpha < \alpha_{\min}$. To prove (a) and (b), we note that the same argument as in Theorem 1 shows that the equilibrium $\alpha = 1$ is locally stable for $s = w(1-(1-\beta)^{n-1}) < s_{\max}$, where $s_{\max}$ is given by (7). It is easy to check that this condition is equivalent to $\gamma > s(1+\sigma(1-s))$. But this condition is implied by the inequality assumption on γ.

As in Theorem 1, $h(\alpha)$ has a zero at $\alpha^* = (s(1+\sigma)-\gamma)/s^2\sigma$. But $\alpha^* > \alpha_{\min}$ if and only if $\gamma < s(1+\sigma) - \alpha_{\min}s^2\sigma$. Hence if this inequality holds, then there is an unstable equilibrium at α^* which implies $h(\alpha) < 0$ to the left of $\alpha = \alpha_{\min}$, so this equilibrium is also stable. If the opposite inequality holds for γ, then $\alpha = 1$ is globally stable. To prove part c, note that $\alpha = 1$ is unstable in this case. Moreover, $\alpha^* < \alpha_{\min}$ leads to a contradiction, since this inequality implies $\gamma > s(1+\sigma) - \alpha_{\min}s^2\sigma > s(1+\sigma(1-s))$, which violates the assumption of (c). Hence $\alpha^* > \alpha_{\min}$ must be a stable equilibrium, in which case it is also globally stable, since an equilibrium at $\alpha_{\min}$ must be unstable. ■

The argument surrounding (8) and (9) implies that if the cost c of contributing to the public good is sufficiently high, cooperation will not take place in a large group, however great the benefits of cooperation. Theorem 2 shows the power of altruistic punishment and the contribution of strong reciprocity to maintaining cooperation in such a situation. As long as (14) holds (which will be the case under a great range of parameter values even when (8) fails), there is a full cooperation equilibrium with a positive level of strong reciprocity. Moreover, we expect c_p to be large, since humans are unique among species that live in groups and recognize individuals, in their capacity to inflict heavy punishment at low cost to the punisher (Bingham 1999), as a result of their superior tool-making and hunting ability (Goodall 1964; Darlington 1975; Plooij 1978; Fifer 1987; Isaac 1987).

6. Cultural Dynamics When Payoffs Are Frequency Dependent

If the payoffs to A and B are frequency dependent, as would be the case if they represented strategies in a non-cooperative game, then we will have in general the functional relationship $s = s(\alpha)$. A specific case of such a functional relationship was presented in the previous section. While we could extend Theorem 1 broadly to this new situation, we will deal with only partial results. We have

THEOREM 3. *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 A, $s = s(\alpha)$. Let $s_{\min}$ and $s_{\max}$ be given by (6) and (7). Then if $s(1) < s_{\max}$, there is a stable altruistic cultural*

equilibrium, and if $s(0) > s_{\min}$*, there is a stable self-interested cultural equilibrium.*

Proof: Note that $h(1) = 0$ and $h'(1) = (s(1) - \gamma)/(1 - s(1)) + \sigma s$, which is negative for $s(1) < s_{\max}$. Thus $\alpha = 1$ is a stable equilibrium. Moreover, $h(0) = 0$ and $h'(0) = \gamma - s(0)(1 + \sigma)$, which is negative for $s(0) > s_{\min}$. Thus $\alpha = 0$ is a stable cultural equilibrium. ■

For an interesting example of frequency dependent cultural equilibrium, suppose in each period members of the population pair off randomly and play a *hawk-dove game* as follows. The two players must share a resource whose value is 2. The 'dove' strategy is to share equally, so when two doves meet each receives a payoff of 1. The 'hawk' strategy is to try to steal the whole resource. When a hawk meets a dove, the hawk takes everything, so has payoff 2, while the dove has payoff 0. When two hawks meet, however, they fight, one (randomly) gets the resource, but each faces an expected net cost (due to energy and injury losses) $a > 0$. It is easy to show that there is a unique, evolutionary stable, equilibrium to this game in which the frequency of doves is $\alpha^* = a/(1 + a)$, and the expected payoff to all players is $\pi^* = a/(1 + a)$. We treat the dove strategy as altruistic, because switching from the hawk to the dove strategy will increase the payoff of all the player's partners, although at a cost to the player himself, if there is a greater than equilibrium fraction of doves.

Suppose the dove norm is culturally transmitted, as in the previous models. If the fraction of doves is α, the expected payoff to a dove is α, and the expected payoff to a hawk is $2\alpha - a(1 - \alpha)$, so the loss associated with being a dove is $s(\alpha) = \alpha(1 + a) - a$. The expression for $h(\alpha)$ is quite complicated in this case, but we do have the following theorem.

THEOREM 4. *Suppose in each period members of the population pair off randomly and play a hawk-dove game, with payoffs as given in the previous paragraph. Then if* $\gamma > 0$*, there is a locally stable cultural equilibrium in which the payoffs are higher than in the self-interested equilibrium.*

Proof: We have $h(1) = -\sigma((2 - \gamma) + a(1 - \gamma))/(2 + a)^2 < 0$ and $h(a/(1 + a)) = a^2\gamma\sigma/(1 + a)^2 > 0$. Thus $h(\alpha)$ must have a positive-to-negative crossing in the interval $(a/(1 + a), 1)$, which is then a

locally stable equilibrium. Since this has a greater than equilibrium fraction of doves, all agents have higher payoffs than in the self-interested equilibrium. ■

7. Cultural Dynamics When Payoffs Do Not Affect Fitness

In all models to this point we have treated the payoffs to cultural traits as biological fitness. If we assume payoffs represent the subjective utility of agents rather than their biological fitness, s can be negative as well as positive, and we must then replace (1) with

$$f_{AA} = \alpha^2, \quad f_{AB} = 2\alpha(1-\alpha), \quad f_{BB} = (1-\alpha)^2. \tag{15}$$

In this case the equation of motion becomes

$$\dot{\alpha} = h(\alpha) = \alpha(1-\alpha)(\alpha\gamma - s\sigma). \tag{16}$$

THEOREM 5. *Suppose the conditions of Theorem 1 hold, except now payoffs represent subjective utility rather than biological fitness. Then there is a cultural equilibrium at* $\alpha = 1$*, which is globally stable if* $s \leq 0$*. If* $s > 0$*, there is always a locally stable self-interested equilibrium at* $\alpha = 0$*, but if* $\gamma > s\sigma$*, there is also a locally stable altruistic equilibrium at* $\alpha = 1$*. In case there are two locally stable equilibria, their basins of attraction are separated by* $\alpha = s\alpha/\gamma$*.*

Proof: The zeros of $h(\alpha)$ are $\alpha = 0, 1, s\sigma/\gamma$, and the last lies in the unit interval when $0 < s\sigma < \gamma$. We also have $h'(\alpha) = -3\alpha^2 - s\sigma + 2\alpha(\gamma + s\sigma)$, which must be negative for a stable equilibrium. Evaluating this expression at the zeros of $h(\alpha)$ verifies the statements in the theorem. ■

If the cost of altruism, s, is frequency dependent, the behavior of the system is more interesting. Consider, for example, the hawk-dove game described in the previous section. Now α is governed by

$$\dot{\alpha} = h(\alpha) = \alpha(1-\alpha)(a\sigma - \alpha(1+a) - \gamma)),$$

and we have the following theorem.

THEOREM 6. *Suppose in each period members of the population pair off randomly and play a hawk-dove game, with payoffs as given in the previous section. Then*

(a) *If $\gamma > \sigma$, there is a globally stable altruistic cultural equilibrium in which the payoffs to all agents is 1, which is an increase of $1/(1 + a)$ over the self-interested equilibrium.*

(b) *If $\gamma < \sigma$, there is a stable cultural equilibrium with both altruists and self-interested types present, and $\alpha = a\sigma/(\sigma(1 + a) - \gamma)$. All agents have payoff $\pi = a\sigma/(\sigma(1 + a) - \gamma)$, which is an increase of $a\gamma/(\sigma(1 + a) - \gamma) > 0$ over the self-interested equilibrium.*

(c) *In the previous case, more oblique transmission (greater γ) or less copying the successful strategy (lower σ) entails a greater gain from attaining the altruistic equilibrium.*

Proof: In this case $h(\alpha)$ has zeros $\alpha = 0, 1$, and $a\sigma/(\sigma(1 + a) - \gamma)$. We have $h'(\alpha) = a\sigma + 3\alpha^2(\sigma(1 + a) - \gamma) + 2\alpha(\gamma - (1 + 2a)\sigma)$. Evaluating this expression at the three equilibria gives the asserted results. ■

8. Explanation of Norm Internalization

Why do we have the generalized capacity to internalize norms? Integrating sociology, economics, and biology to explain altruism will not be effective unless this question is answered, because the concept of internalization sits uneasily with both the economist's model of the self-interested rational actor and the biologist's requirement that the concept be evolutionarily grounded. The capacity to internalize is certainly curious, something akin to the capacity of a digital computer to be programmed, albeit only within certain strict limits. From a biological standpoint, internalization may be an elaboration of imprinting and imitation mechanisms found in several species of birds and mammals, but its highly developed form in humans indicates it probably had great adaptive value during our evolutionary emergence as a species. Moreover, from an economic standpoint, the everyday observation that people who exhibit a strongly internalized moral code lead happier and more fulfilled lives than those who subject all actions to a narrow calculation of personal costs and benefits of norm compliance, suggests it might not be 'rational' to be self-interested.

A related paper, Gintis (2003), shows that *if* internalization of *some* norms is personally fitness-enhancing (e.g. preparing for the future, having good personal hygiene, positive work habits, and/or

control of emotions), *then* genes promoting the capacity to internalize can evolve. Given this genetic capacity, we have seen above, altruistic norms will be internalized as well, provided their fitness costs are not excessive. In effect, altruism 'hitchhikes' on the personal fitness-enhancing capacity of norm internalization.[12] Altruistic behavior, then, is an *exception*, in the sense of Gould and Vrba (1981).

But, why should the internalization of *any* norms be individually fitness-enhancing? The following is a possible explanation, based on the observation that internalization alters the agent's *goals*, whereas instrumental and conventional cultural forms merely aid the individual in attaining *pre-given* goals. In humans, as much as in other species, these goals are related to, but not reducible to, biological fitness.

Biological fitness is a theoretical abstraction unknown to virtually every real-life organism. Organisms therefore do not, in any circumstance, literally maximize fitness. Rather, organisms have a relatively simple state-dependent preference function that is itself subject to selection according to its ability to promote individual fitness (Alcock 1993). In a slowly changing environment, this preference function will track fitness closely. In a rapidly changing environment, however, natural selection will be too slow, and the preference function will not track fitness well.

The development of cultural transmission, in the form of instrumental techniques and conventions, and the ensuing increase in social complexity of hominid society, doubtless produced such a rapidly changing environment, thus conferring high fitness value on the development of a *non-genetic mechanism for altering the agent's preference function*. Internalization is adaptive because it allows the human preference function to shift in directions conducive to higher personal fitness. The internalization of norms is thus adaptive because it facilitates the transformation of drives, needs, desires, and pleasures (arguments in the human objective function) into forms that are more closely aligned with fitness maximization. Internalization is limited to our species, moreover, because no other species places such great emphasis on cultural transmission.

We humans thus have a 'primordial' preference function that does not well serve our fitness interests, and which is more or less successfully 'overridden' by our internalized norms. This primordial preference function knows nothing of 'thinking ahead', but rather satisfies immediate desires. Lying, cheating, killing, stealing, and satisfying short-term bodily needs (wrath, lust, greed, gluttony, sloth) are all

actions that produce immediate pleasure and drive-reduction at the expense of our overall well-being in the long run. This fact explains the congenital weakness of human nature in its tendency to succumb to the unruly temptations of the flesh.

This evolutionary argument is meant to apply to the long period in the Pleistocene during which the human character was formed. Social change since the agricultural revolution some 10,000 years ago has been far too swift to permit even the internalization of norms to produce a close fit between preference and fitness. Indeed, with the advent of modern societies, the internalization of norms has been systematically diverted from *fitness* (expected number of offspring) to *welfare* (net degree of contentment) maximization. This, of course, is precisely what we would expect when humans obtain control over the content of ethical norms. Indeed, this *misfit* between welfare and fitness is doubtless a necessary precondition for civilization and a high level of *per capita* income. This is true because, were we fitness maximizers, every technical advance would have been accompanied by an equivalent increase in the rate of population growth, thus nullifying its contribution to human welfare, as predicted long ago by Thomas Malthus. The demographic transition, which has led to dramatically reduced human birth rates throughout most of the world, is a testimonial to the gap between welfare and fitness. Perhaps the most important form of prosocial cultural transmission in the world today is the norm of having few, but high-quality, offspring.

9. Conclusion

Economics and biology offer a wealth of powerful analytical tools for analyzing dynamic strategic interaction. But the key to solving the puzzle of prosociality comes from a venerable branch of sociology: socialization theory, and more specifically, the theory of *internalization of norms*, first stressed by Durkheim (the quote is from *Suicide*):

> The influence of society is what had roused in us the sentiments of sympathy and solidarity drawing us toward others; it is society which, fashioning us in its image, fills us with religious, political and moral beliefs that control our actions.

Biologists, who are most comfortable with models of Darwinian competition, and economists, who are no less at home with models

of market competition, have ignored this phenomenon, doubtless because individuals who internalize norms that reduce their fitness (biology) or material rewards (economics) should be outcompeted by others who are immune to society's psychological manipulations.

This article shows that the internalization of norms is compatible with the economist's concept of optimization of a preference function by choosing from a set of available actions and copying the successful behavior of others (the replicator dynamic), as well as the biologist's treatment of adaptive dynamics and cultural transmission. When appropriately combined, they solve the problem of prosociality.

NOTES

I thank Samuel Bowles, Rob Boyd, and Ernst Fehr for helpful comments, and the John D. and CatherineT. MacArthur Foundation for financial support.

1. It is curious that these disciplines, while their models are incompatible, and each discipline considers itself 'scientific,' have not witnessed sustained attempts at adjudicating their differences in explaining the same object of knowledge: cooperation in human society. This would not be tolerated in the natural sciences and, indeed, is sorely in need of repair.
2. The same idea might be more evocatively phrased by saying that internalized norms are *constitutive of the self*.
3. Population contractions were likely common in the evolutionary history of *Homo sapiens* (Boone and Kessler 1999). The very low rate of growth of the human population over the whole prehistoric period, plus the high rate of human population growth in even poor contemporary foraging societies in good times (Keckler 1997), suggests that periodic crises occurred in the past. Moreover, flattened mortality profiles of prehistoric skeletal populations indicate population crashes ranging from 10% to 54% at a mean rate of once per 30 years (Keckler 1997). Finally, optimal foraging models of hunter-gatherer societies often predict limit cycles (Belovsky 1988).
4. We provide experimental evidence for norm internalization in Section 3.3.
5. In the biological literature, *group selection* applied to a trait has tended to mean that the trait suffers a within-group fitness deficit, but nevertheless grows in the population because groups in which the trait is prevalent outcompete other groups in which the trait has low frequency (Wilson 1980). There is a weaker sense of group selection in which a trait is associated with, or facilitates, a certain group structure and groups with this structure outcompete groups without the structure, but the selected trait does not have a within-group fitness disadvantage. The biological critiques of group selection do not apply to this weaker notion. Our model is a group selection model in this weaker, uncontroversial, sense.
6. There is little doubt but that humans have prosocial human genes. We know, for instance, that brain structures in the prefontal lobes are required for normal

sociality (Damasio 1994; Damasio et al. 1994). Moreover, sociopathy is heritable (Mealey 1995). Finally, some types of prosocial behavior, especially revenge-seeking impulses, are counterindicated in many cultural systems (e.g. the Judeo-Christian), though these altruistic behaviors are nevertheless quite common, indicating that they are expressed even when they are not formally taught.

7. The observed behavior was predicted by Akerlof (1982) and Blau (1964).
8. Also interesting is the fact that the modal offer in these groups was 30%, much lower than the mode among college students (usually 50%), and a full third of such offers were rejected. The Au and the Gnau even rejected about half of offers of 40% of the pie.
9. We assume s is the same for all agents. It is plausible that more fit agents will have lower s because they exhaust fewer resources in contributing to the group. Indeed, behaving altruistically could then serve as a costly signal of quality. See Gintis et al. (2001) for details.
10. For a more general derivation, see Gintis (2000a).
11. 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 et al. (1992), Boyd and Richerson (1992), Gintis (2000b), Fehr and Gächter (2000), Henrich and Boyd (2001), and Henrich et al. (2001).
12. This mechanism was asserted by Simon (1990), who instead of 'internalization of norms' used the term 'docility', in the sense of 'capable of being easily led or influenced'.

REFERENCES

Akerlof, George A. 1982. 'Labor Contracts as Partial Gift Exchange.' *Quarterly Journal of Economics* 97: 543–69.

Alcock, John. 1993. *Animal Behavior: An Evolutionary Approach*. Sunderland, MA: Sinauer.

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

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

Andreoni, James. 1995. 'Cooperation in Public Goods Experiments: Kindness or Confusion.' *American Economic Review* 85: 891–904.

Axelrod, Robert and William D. Hamilton. 1981. 'The Evolution of Cooperation.' *Science* 211: 1390–6.

Belovsky, G. 1988. 'An Optimal Foraging-Based Model of Hunter-Gatherer Population Dynamics.' *Journal of Anthropological Archaeology* 7: 329–72.

Benedict, Ruth. 1934. *Patterns of Culture*. Boston, MA: Houghton Mifflin.

Bingham, Paul M. 1999. 'Human Uniqueness: A General Theory.' *Quarterly Review of Biology* 74: 133–69.

Blau, Peter. 1964. *Exchange and Power in Social Life*. New York: John Wiley.

Blount, Sally. 1995. 'When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences.' *Organizational Behavior & Human Decision Processes* 63: 131–44.

Bonner, John Tyler. 1984. *The Evolution of Culture in Animals*. Princeton, NJ: Princeton University Press.

Boone, James L. and Karen L. Kessler. 1999. 'More Status or More Children? Social Status, Fertility Reduction, and Long-Term Fitness.' *Evolution and Human Behavior* 20: 257–77.

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

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

Bowles, Samuel and Herbert Gintis. 1998. 'The Moral Economy of Community: Structured Populations and the Evolution of Prosocial Norms.' *Evolution and Human Behavior* 19: 3–25.

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

Boyd, Robert and Peter J. Richerson. 1988. 'The Evolution of Reciprocity in Sizable Groups.' *Journal of Theoretical Biology* 132: 337–56.

Boyd, Robert and Peter J. Richerson. 1990. 'Group Selection among Alternative Evolutionarily Stable Strategies.' *Journal of Theoretical Biology* 145: 331–42.

Boyd, Robert and Peter J. Richerson. 1992. 'Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups.' *Ethology and Sociobiology* 113: 171–95.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. 'Evolution of Altruistic Punishment.' *Proceedings of the National Academy of Sciences* 100: 3531–35.

Brown, Donald E. 1991. *Human Universals*. New York: McGraw-Hill.

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

Clutton-Brock, T. H. and G. A. Parker. 1995. 'Punishment in Animal Societies.' *Nature* 373: 58–60.

Conlisk, John. 1988. 'Optimization Cost.' *Journal of Economic Behavior and Organization* 9: 213–28.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Damasio, H., T. Grabowski, R. Frank, A. M. Galaburda and A. R. Damasio. 1994. 'The Return of Phineas Gage: Clues about the Brain from the Skull of a Famous Patient.' *Science* 264: 1102–5.

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

Dawes, Robyn M., John M. Orbell and J. C. Van de Kragt. 1986. 'Organizing Groups for Collective Action.' *American Political Science Review* 80: 1171–85.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

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

Edgerton, Robert B. 1992. *Sick Societies: Challenging the Myth of Primitive Harmony*. New York: The Free Press.

Fehr, Ernst and Simon Gächter. 2000. 'Cooperation and Punishment.' *American Economic Review* 90: 980–94.
Fehr, Ernst and Simon Gächter. 2002. 'Altruistic Punishment in Humans.' *Nature* 415: 137–40.
Fehr, Ernst, Simon Gächter and Georg Kirchsteiger. 1997. 'Reciprocity as a Contract Enforcement Device: Experimental Evidence.' *Econometrica* 65: 833–60.
Fifer, F. C. 1987. 'The Adoption of Bipedalism by the Hominids: a New Hypothesis.' *Human Evolution* 2: 135–47.
Frank, Steven A. 1995. 'Mutual Policing and Repression of Competition in the Evolution of Cooperative Groups.' *Nature* 377: 520–2.
Fudenberg, Drew and Eric Maskin. 1986. 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information.' *Econometrica* 54: 533–54.
Gadagkar, Raghavendra. 1991. 'On Testing the Role of Genetic Asymmetries Created by Haplodiploidy in the Evolution of Eusociality in the Hymenoptera.' *Journal of Genetics* 70: 131.
Gintis, Herbert. 1975. 'Welfare Economics and Individual Development: A Reply to Talcott Parsons.' *Quarterly Journal of Economics* 89: 291–302.
Gintis, Herbert. 2000a. *Game Theory Evolving*. Princeton, NJ: Princeton University Press.
Gintis, Herbert. 2000b 'Strong Reciprocity and Human Sociality.' *Journal of Theoretical Biology* 206: 169–79.
Gintis, Herbert. 2003. 'The Hitchhiker's Guide to Altruism: Genes and Culture, and the Internalization of Norms.' *Journal of Theoretical Biology*.
Gintis, Herbert, Eric Alden Smith and Samuel Bowles. 2001. 'Costly Signaling and Cooperation.' *Journal of Theoretical Biology* 213: 103–19.
Goodall, Jane. 1964. 'Tool-using and Aimed Throwing in a Community of Free-Living Chimpanzees.' *Nature* 201:1264–6.
Gould, Stephen J. and Elizabeth Vrba. 1981. 'Exaption: A Missing Term in the Science of Form.' *Baleobiology* 8: 415.
Grusec, Joan E. and Leon Kuczynski. 1997. *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*. New York: John Wiley & Sons.
Hamilton, W. D. 1963. 'The Evolution of Altruistic Behavior.' *American Naturalist* 96: 354–6.
Henrich, Joseph and Robert Boyd. 2001. 'Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas.' *Journal of Theoretical Biology* 208: 79–89.
Henrich, Joseph and Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. 'Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies.' *American Economic Review* 91: 73–8.
Hirshleifer, David and Eric Rasmusen. 1989. 'Cooperation in a Repeated Prisoners' Dilemma with Ostracism.' *Journal of Economic Behavior and Organization* 12: 87–106.
Isaac, B. 1987. 'Throwing and Human Evolution.' *African Archeological Review* 5: 3–17.
Keckler, C.N.W. 1997. 'Catastrophic Mortality in Simulations of Forager Age-of-Death: Where Did all the Humans Go?.' In *Integrating Archaeological Demography: Multidisciplinary Approaches to Prehistoric Populations. Center for Archaeological Investigations, Occasional Papers No. 24*, ed. R. Paine, pp. 205–28. Carbondale, IL: Southern Illinois Univerisity Press.

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

Ledyard, J. O. 1995. 'Public Goods: A Survey of Experimental Research.' In *The Handbook of Experimental Economics*, eds J. H. Kagel and A. E. Roth , pp. 111–94. Princeton, NJ: Princeton University Press.

Maynard Smith, John. 1998. 'The Origin of Altruism.' *Nature* 393: 639–40.

Maynard Smith, John and Eors Szathmary. 1997. *The Major Transitions in Evolution*. Oxford: Oxford University Press.

Mead, Margaret. 1963. *Sex and Temperament in Three Primitive Societies*. New York: Morrow.

Mealey, Linda 1995. 'The Sociobiology of Sociopathy.' *Behavioral and Brain Sciences* 18: 523–41.

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

Nowak, Martin A. and Karl Sigmund. 1998. 'Evolution of Indirect Reciprocity by Image Scoring.' *Nature* 393: 573–7.

Ostrom, Elinor, James Walker and Roy Gardner. 1992. 'Covenants With and Without a Sword: Self-Governance Is Possible.' *American Political Science Review* 86: 404–17.

Parsons, Talcott. 1964. 'Evolutionary Universals in Society.' *American Sociological Review* 29: 339–57.

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

Plooij, F. X. 1978. 'Tool-using during Chimpanzees' Bushpig Hunt.' *Carnivore* 1: 103–6.

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

Simon, Herbert. 1990. 'A Mechanism for Social Selection and Successful Altruism.' *Science* 250: 1665–8.

Soltis, Joseph, Robert Boyd and Peter Richerson. 1995. 'Can Group-functional Behaviors Evolve by Cultural Group Selection: An Empirical Test.' *Current Anthropology* 36: 473–83.

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

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

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

Wilson, David Sloan. 1980. *The Natural Selection of Populations and Communities*. Menlo Park, CA: Benjamin Cummings.

Wrong, Dennis H. 1961. 'The Oversocialized Conception of Man in Modern Sociology.' *American Sociological Review* 26: 183–93.

Yamagishi, Toshio. 1988a. 'The Provision of a Sanctioning System in the United States and Japan.' *Social Psychology Quarterly* 51: 265–71.

Yamagishi, Toshio. 1988b. 'Seriousness of Social Dilemmas and the Provision of a Sanctioning System.' *Social Psychology Quarterly* 51: 32–42.

Yamagishi, Toshio. 1992. 'Group Size and the Provision of a Sanctioning System in a Social Dilemma.' In *Social Dilemmas: Theoretical Issues and Research Findings*, eds W. B. G. Liebrand, David M. Messick, and H. A. M. Wilke, pp. 267–87. Oxford: Pergamon Press.

HERBERT GINTIS is External Faculty, Santa Fe Institute. His research interests include evolutionary game theory, behavioral and experimental economics, behavioral labor economics, and intergenerational status transmission. Among his recent publications (not mentioned above) are 'A Markov Model of Production, Trade, and Money: Theory and Artificial Life Simulation' (*Computational and Mathematical Organization Theory* 3 (1997): 19–41); with Samuel Bowles and Melissa Osborne, 'Incentive-Enhancing Preferences: Personality, Behavior and Earnings' (*American Economic Review* 91 (2001): 155–8); with Samuel Bowles and Melissa Osborne, 'The Determinants of Individual Earnings: Skills, Preferences, and Schooling' (*Journal of Economic Literature* (2002): 1137–76); with Samuel Bowles, 'Intergenerational Inequality' (*Journal of Economic Perspectives* 3 (2002): 3–30).

ADDRESS: 15 Forbes Avenue, Northampton, MA 01060, USA
[email: hgintis@attbi.com; http://www-unix.oit.umass.edu/~gintis]