Chapter from:
Genetic and Cultural Evolution of Cooperation
Edited by P. Hammerstein, 2003
ISBN 0-262-08326-4

22

Origins of Human Cooperation

Samuel Bowles and Herbert Gintis
Santa Fe Institute, Santa Fe, NM 87501, U.S.A.

ABSTRACT

Biological explanations of cooperation are based on kin altruism, reciprocal altruism, and mutualism, all of which apply to human and nonhuman species alike. Human cooperation, however, is based in part on capacities that are unique to, or at least much more highly developed in, *Homo sapiens*. In this chapter, an explanation of cooperation is sought that works for humans but does not work for other species, or works substantially less well. Central to this explanation will be human cognitive, linguistic, and physical capacities that allow the formulation of general norms of social conduct, the emergence of social institutions regulating this conduct, the psychological capacity to internalize norms, and the formation of groups based on such nonkin characteristics as ethnicity and linguistic behavior, which facilitates highly costly conflicts among groups. Agent-based modeling shows that these practices could have coevolved with other human traits in a plausible representation of the relevant environments. The forms of cooperation to be explained are confirmed by natural observation, historical accounts, and behavioral experiments and are based on a plausible evolutionary dynamic involving some combination of genetic and cultural elements, the consistency of which can be demonstrated through formal modeling. Moreover, the workings of the models developed account for human cooperation under parameter values consistent with what can be reasonably inferred about the environments in which humans have lived.

INTRODUCTION

> The Americans ... are fond of explaining almost all the actions of their lives by the principle of self interest rightly understood; they show with complacency how an enlightened regard for themselves constantly prompts them to assist one another and inclines them willingly to sacrifice a portion of their time and property to the welfare of the state. In this respect I think they frequently fail to do themselves justice; in the United States as well as elsewhere people are sometimes seen to give way to those disinterested and spontaneous impulses that are natural to man; but the Americans seldom admit that they yield to emotions of this kind.
>
> — Alexis de Tocqueville
> (*Democracy in America*, 1830, Book II, chapter VII)

Cooperation among humans is unique in nature, extending to a large number of unrelated individuals and taking a vast array of forms. By cooperation we mean an individual behavior that incurs personal costs to engage in a joint activity that

confers benefits exceeding these costs to other members of one's group. This applies, for example, to contributing in a public goods game.[1] Although the absence of this unique type of cooperation in other species could be an evolutionary accident, a more plausible explanation is that human cooperation is the result of human capacities that are unique to our species.

Common explanations of cooperation in other species based on genetic relatedness (kin altruism) and repeated interactions (e.g., reciprocal altruism) certainly apply to cooperation in humans as well. However, the capacities underlying these mechanisms are not unique to humans: repeated interactions and interactions among kin are common in many species. We do not seek to diminish the importance of these familiar modes or to suggest that extensions of them to account for uniquely human aspects of cooperation are uninteresting. Rather we suggest that it would be fruitful to seek an explanation of cooperation that works for humans but, because it centrally involves attributes unique to humans, does not work for other species, or works substantially less well.

Central to our explanation will be human cognitive, linguistic, and physical capacities that allow the formulation of general norms of social conduct, the emergence of social institutions regulating this conduct, the psychological capacity to internalize norms, and the basing of group membership on such nonkin characteristics as ethnicity and linguistic behavior, which facilitates highly costly conflicts among groups. Of course, it will not do to posit these rules and institutions a priori. Rather, we must show that these could have coevolved with other human traits in a plausible representation of the relevant environments.

Our thinking, while necessarily speculative, has been disciplined in three ways. First, the forms of cooperation we seek to explain are confirmed by natural observation, historical accounts, and behavioral experiments. Second, we require that our account be based on a plausible evolutionary dynamic involving some combination of genetic and cultural elements, the consistency of which can be demonstrated through formal modeling. Third, the workings of the models we develop must account for human cooperation under parameter values consistent with what can be reasonably inferred about the environments in which humans have lived. When the models in question resist analytical solution (because they are complicated and highly nonlinear), this third requirement entails computer simulation under plausible parameter values.

The chapter is structured as follows:

- We support our assertion that explanations based on kin and reciprocal altruism are incomplete.
- We characterize key individual behavioral traits that we think account for much of human cooperation. We term this *strong reciprocity.*

[1] This definition of cooperation excludes mutually beneficial interactions (mutualisms), the evolutionary explanation of which is relatively simple; nonproductive forms of altruism (in which the benefit received does not exceed the cost to the altruist); and those lacking the common benefits of joint activity that are characteristic of the behaviors we wish to explain.

- We explain why multilevel selection among human groups operating on both cultural and genetic variability must play an important role in our explanation.
- We show that some common human institutions create the conditions under which multilevel selection is especially powerful. This provides a reason why group-level institutions, such as resource sharing as well as warfare, may have coevolved with the individual behaviors we call strong reciprocity.
- We explain why strong reciprocators may have been favored evolutionarily under conditions where their actions constituted a difficult-to-fake (costly) signal of their otherwise unobservable qualities as a mate, coalition partner, or opponent.

We argue that the maintenance of group boundaries through the parochial exclusion of "outsiders" may have contributed to the evolutionary success of cooperative behaviors. This, in turn, may provide part of the explanation of the salience of group membership as a determinant of the scope of cooperative relationships.

We develop the idea that human capacities to internalize norms and mobilize emotions in support of cooperative behavior have attenuated the conflict between individual-interest and group-benefit, and have thus supported cooperative interactions even under conditions when multilevel selection and the cooperation-inducing effects of costly signaling are weak.

In the spirit of this gathering, we concentrate on expressing a point of view, without giving the full attention to the more nuanced and formal arguments that a more extended presentation would allow. Nor do we take note of our immense debt to the work of other scholars, many of them joining us in this workshop, except to say that what follows is the result of a sustained collaboration in recent years with Ernst Fehr, Simon Gächter, Armin Falk, Urs Fischbacher, and their coauthors, as well as with Robert Boyd, Marcus Feldman, Joe Henrich, Peter Richerson, and Eric Alden Smith. Our own contributions to the ideas expressed here are summarized in our recent synthetic works (Gintis 2000a; Bowles 2003).

WHY EXPLANATIONS BASED ON KIN AND RECIPROCAL ALTRUISM ARE INCOMPLETE

We do not doubt that relatedness is an important part of the explanation of human cooperation, as it is among other animals, and that cooperation among kin may have been a template whose gradual extension contributed to cooperation among nonkin. However, to explain human cooperation among large numbers of unrelated individuals in this way is implausible.

Similarly, repeated interactions allowing retaliation against antisocial actions undoubtedly contribute to sustaining cooperation among humans and perhaps among some other animals. Some have suggested that the evolution of cooperation among entirely self-interested humans is explained in this manner.

This, however, is false. First, much of the experimental evidence about human behaviors contributing to cooperation comes from nonrepeated interactions, or from the final round of a repeated interaction. We do not think that subjects are unaware of the one-shot setting, or unable to leave their real-world experiences with repeated interactions at the laboratory door. Indeed, evidence is overwhelming that humans readily distinguish between repeated and nonrepeated interactions and adapt their behavior accordingly. Nonexperimental evidence is equally telling: common behaviors in warfare as in everyday life are not easily explained by the expectation of future reciprocation.

Second, conditions of early humans may have made the repetition–retaliation mechanism an ineffective support for cooperation. Members of mobile foraging bands could often escape retaliation by relocating to other groups. Moreover, in many situations critical to human evolution, repetition of an interaction was quite unlikely, as when groups faced dissolution as the result of group conflict or an adverse environment.

Third, the conditions under which repetition and retaliation can explain why self-regarding individuals would cooperate are not met in settings where large numbers interact. The celebrated "folk theorem," which is frequently invoked to show that repeated interactions among self-regarding individuals can support seemingly other-regarding behaviors, does not extend plausibly from two-person to n-person groups for large n. Critical differences between dyadic and n-person interactions in this respect are that (a) the number of accidental defections or perceived defections increases with n, and such "trembles" dramatically increase the cost of punishing defectors; (b) probability that a sufficiently large fraction of a large group of heterogeneous agents will be sufficiently forward-looking to make cooperation profitable decreases exponentially as n rises; and (c) coordination and incentive mechanisms required to ensure punishment of defectors by self-regarding group members become increasingly complex and unwieldy as n increases.[2] Although many important human interactions are dyadic (e.g., mutualistic exchange of goods), many important examples of cooperation (e.g., risk reduction through co-insurance, information sharing, maintenance of group-beneficial social norms, and group defense) are large group interactions. For these cases, the folk theorem provides no reason to expect cooperation to be common and durable rather than rare and ephemeral.

PSYCHOLOGICAL AND BEHAVIORAL ASPECTS OF ALTRUISM: PROSOCIAL EMOTIONS AND STRONG RECIPROCITY

Prosocial emotions are physiological and psychological reactions that induce agents to engage in cooperative behaviors as we have defined them above. Some

[2] A well-known theorem showing that repetition among a large number of agents can support efficient cooperative equilibria (Fudenberg and Maskin 1990) effectively requires group members to be infinitely lived and does not apply even approximately to human groups under the most optimistic assumptions concerning longevity and future orientation.

prosocial emotions, including shame, guilt, empathy, and sensitivity to social sanction, induce agents to undertake constructive social interactions; others, such as the desire to punish norm violators, reduce free riding when the prosocial emotions fail to induce sufficiently cooperative behavior in some fraction of members of the social group (Frank 1987; Hirshleifer 1987).

Without prosocial emotions, we would all be sociopaths, and human society would not exist, however strong the institutions of contract, governmental law enforcement, and reputation. Sociopaths have no mental deficit except that their capacity to experience shame, guilt, empathy, and remorse is severely attenuated or absent. They comprise 3–4% percent of the male population in the United States (Mealey 1995), but account for approximately 20% of the United States' prison population and between 33% and 80% of the population of chronic criminal offenders.

Prosocial emotions are responsible for the host of civil and caring acts that enrich our daily lives and render living, working, shopping, and traveling among strangers feasible and pleasant. Moreover, representative government, civil liberties, due process, women's rights, respect for minorities, to name a few of the key institutions without which human dignity would be impossible in the modern world, were brought about by people involved in collective action, pursuing not only their personal ends but also a vision for all of humanity. Our freedoms and comforts alike are based on the emotional dispositions of generations past.

Whereas we think evidence is strong that prosocial emotions account for important forms of human cooperation, there is no universally accepted model of how emotions combine with more cognitive processes to affect behaviors. Nor is there much agreement on how best to represent the prosocial emotions that support cooperative behaviors, although we (Bowles and Gintis 2002) have attempted one in this direction. It is uncontroversial, however, to assert that there are many civic-minded acts that cannot be explained by self-regarding preferences, including why people vote, why they give anonymously to charity, and why they sacrifice themselves in battle. In dealing with these areas of social life, a suggestive body of evidence points to a behavior that we call *strong reciprocity*. 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 reasonably expect future personal gains from such retaliation. The strong reciprocator is thus both a *conditionally altruistic cooperator* and a *conditionally altruistic punisher* whose actions benefit other group members at a personal cost. We call this reciprocity "strong" to distinguish it from such forms of "weak" reciprocity as reciprocal altruism, indirect reciprocity, and other such interactions that posit individually self-regarding behavior sustained by repeated interactions or positive assortation (see Fehr et al. 2002).

MULTILEVEL SELECTION

In populations composed of groups characterized by a markedly higher level of interaction among members than with outsiders, it has long been recognized that evolutionary processes may be decomposed into between-group and within-group selection effects (Price 1970). Where the rate of replication of a trait depends on the composition of the group, and where group differences in composition exist, group selection contributes to the pace and direction of evolutionary change. Until recently, however, most who modeled evolutionary processes under the joint influence of group and individual selection have concluded that the former cannot offset the latter, except where special circumstances (small group size, limited migration) heighten and sustain differences between groups relative to within-group differences.

Thus, group selection models are widely judged to have failed to explain evolutionary success of individually costly forms of group-beneficial sociality. But group selection operating on genetic and cultural variation may be of considerably greater importance among humans than other animals. Among the distinctive human characteristics that enhance the relevance of group selection is our capacity to suppress within-group phenotypic differences (e.g., via resource sharing, co-insurance, consensus decision making), conformist cultural transmission, ethnocentrism (which supports positive assortation within groups and helps maintain group boundaries), and the high frequency of intergroup conflict.

In Gintis (2000b) we develop an analytical model showing that under plausible conditions strong reciprocity can emerge from reciprocal altruism, through group selection. The paper models cooperation as a repeated *n*-person public goods game in which, under normal conditions, when agents are sufficiently attentive to future gains from group membership, cooperation is sustained by trigger strategies, as asserted in the folk theorem. However, when the group is threatened with extinction or dispersal, say through war, pestilence, or famine, cooperation is most needed for survival. Probability of one's contributions being repaid in the future, however, decreases sharply when the group is threatened, since the probability that the group will dissolve increases and hence the incentive to cooperate will dissolve. Thus, *precisely when a group is most in need of prosocial behavior, cooperation based on reciprocal altruism will collapse*. Such critical periods were common in the evolutionary history of our species. A small number of strong reciprocators, who punish defectors *without regard for the probability of future repayment*, can dramatically improve the survival chances of human groups. Moreover, 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, as a result of their superior tool-making and hunting ability. Indeed, and in sharp contrast to nonhuman primates, even the strongest man can be killed while sleeping by the weakest, at low cost to the punisher. A simple argument using Price's equation then shows that under these

conditions, strong reciprocators can invade a population of self-regarding types and can persist in equilibrium.

Our joint work with Boyd and Richerson (Boyd et al. 2003) shows, through agent-based simulations, that for some cooperative behaviors — notably punishing those who violate cooperative norms — group selection on culturally transmitted traits can be decisive even for very large groups and for substantial rates of migration. The reason for this surprising result is that if most members of a group are adhering to the norm, the costs incurred by those predisposed to punish violators are very small for the simple reason that violations are infrequent. Thus while within-group selection against the cooperative behavior exists, it is very weak in the neighborhood of the cooperative equilibrium. This supports the persistence over long periods of substantial between-group differences in composition, some with virtually all cooperative agents predisposed to cooperate and punish those who do not, and other groups composed of virtually all self-regarding individuals Additional between-group variance is provided by intergroup conflicts following which winning groups absorb losers and then divide.

One particularly attractive property of these models is that they predict a heterogeneous equilibrium with a considerable fraction of both self-regarding and strong reciprocator types, as is often found in the experimental literature (Fehr and Gächter 2002).

COEVOLUTION OF INSTITUTIONS AND BEHAVIORS

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

The idea that the suppression of within-group competition may be a strong influence on evolutionary dynamics has been widely recognized in eusocial insects and other species. Alexander (1979), Boehm (1982), and Eibl-Eibesfeldt (1982) first applied this reasoning to human evolution, exploring the role of culturally transmitted practices that reduce phenotypic variation within groups. Examples of such practices are leveling institutions, such as resource sharing among nonkin, namely those which reduce within-group differences in reproductive fitness or material well-being. These practices are leveling to the extent that they result in less pronounced within-group differences in material well-being or fitness than would have obtained in their absence. Thus, the fact that good

hunters who are generous toward other group members may experience higher fitness than other hunters and enjoy improved nutrition (as a result of consumption smoothing) does not indicate a lack of leveling unless these practices also result in lesser fitness and worse nutrition among less successful hunters (which seems highly unlikely).

By reducing within-group differences in individual success, such practices may have attenuated within-group genetic or cultural selection operating against individually costly but group-beneficial practices, thus giving the groups adopting them advantages in intergroup contests. Group-level institutions are thus constructed environments capable of imparting distinctive direction and pace to the process of biological evolution and cultural change. Hence, evolutionary success of social institutions that reduce phenotypic variation within groups may be explained by the fact that they retard selection pressures working against in-group-beneficial individual traits and the fact that high frequencies of bearers of these traits reduces the likelihood of group extinctions.

We have modeled an evolutionary dynamic along these lines with the novel features that genetically and culturally transmitted individual behaviors as well as culturally transmitted group-level institutional characteristics are subject to selection, with intergroup contests playing a decisive role in group-level selection (Bowles 2001; Bowles et al. 2003). We show that intergroup conflicts may explain the evolutionary success of both (a) altruistic forms of human sociality toward nonkin and (b) group-level institutional structures such as resource sharing that have emerged and diffused repeatedly in a wide variety of ecologies during the course of human history. In-group-beneficial behaviors may evolve if they inflict sufficient costs on outgroup individuals and group-level institutions limit the individual costs of these behaviors and thereby attenuate within-group selection against these behaviors.

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

STRONG RECIPROCITY AS A SIGNAL OF QUALITY

Cooperative behaviors may be favored in evolution because they enhance the individual's opportunities for mating and coalition building. This would be the

case, for example, if sharing valuable information or incurring dangers in defense of the group were taken by others as an honest signal of the individual's otherwise unobservable traits as a mate or political ally. Much of the literature on costly signaling and human evolution explains such behaviors as good hunters contributing their prey to others, but the same reasoning applies to cooperative behaviors. Cooperative behaviors would thus result in advantageous alliances for those signaling in this manner, and the resulting enhanced fitness or material success would then account for the proliferation of the cooperative behaviors constituting the signal. With Eric Alden Smith (Gintis et al. 2001), we have modeled this process as a multiplayer public goods game that involves no repeated or assortative interactions, so that noncooperation would be the dominant strategy if there were no signaling benefits. We show that honest signaling of underlying quality by providing a public good to the rest of the group can be evolutionarily stable and proliferate in a population in which it is initially rare, provided that certain plausible conditions hold. Behaviors conforming to what we call strong reciprocity could have thus evolved in this way.

Our signaling equilibrium alone, however, does not require that the signal confer benefits on other group members. Antisocial behaviors could perform the same function: beating up one's neighbor can demonstrate prowess just as much as behaving bravely in defense of the group. If signaling is to be an explanation of group-beneficial behavior, the logic of the model must be complemented by a demonstration that group-beneficial signaling is favored over antisocial signaling. We supply this by noting that the level of public benefit provided may be positively correlated with the individual benefit the signaler provides to those who respond to the signal. For instance, the signaler who defends the group is more likely to confer a benefit (say, protection) on his partner or allies than the signaler who beats up his neighbor. Group-beneficial signals may attract larger audiences than antisocial signals. Finally, group selection among competing groups would favor those at group-beneficial signaling equilibria over those either at nonsignaling equilibria or those at antisocial signaling equilibria.

As this last reason suggests, the effects of signaling and group selection on cooperation may be synergistic rather than simply additive. Group selection provides a reason why signaling may be prosocial, whereas signaling theory provides a reason why group-beneficial behaviors may be evolutionarily stable in a within-group dynamic, thus contributing to between-group variance in behavior and thereby enhancing the force of group selection.

PAROCHIALISM AND RECIPROCITY

The predisposition of individuals to behave cooperatively often depends on the identities of the individuals with whom they are interacting: "insiders" are favored over "outsiders." Insider-outsider distinctions play a critical role in the above models. In our group selection models, cooperative behaviors conferring benefits on fellow group members allowed highly cooperative groups to prevail

in intergroup conflicts. Thus the very behaviors that are beneficial to one's own group are costly or even lethal to members of other groups. In addition, as we have seen, maintenance of group boundaries to limit the extent of migration and the frequency of intergroup conflict contribute substantially to the force of group selection in promoting cooperation within groups. Thus it seems likely that within-group cooperation and hostility toward "outsiders" coevolved.

In-group favoritism is often supported by the cultural salience of physical, linguistic, and other behavioral markers identifying insiders and outsiders in conjunction with exclusionary practices, which we call parochialism. We have modeled parochialism as a filter on the ascriptive traits of those with whom one might interact, a particular filter excluding those with "objectionable" traits (Bowles and Gintis 2000). Members of groups benefit in two ways from the adoption of more parochial filters: equilibrium group size and cultural heterogeneity of group members is thereby reduced, and this enhances the effectiveness of mutual monitoring and reputation-building in supporting high levels of within-group cooperation. More parochial groups forego the economies of scale, gains from exchange, and possible collective cognition benefits of larger and more diverse membership. The degree of parochialism observed in a population will depend on the balance of these benefits and costs of exclusionary practices. As these have evolved over time with the effects of changing environments and technologies, our analysis of "optimal parochialism" may provide a way of modeling the coevolution of cooperation and out-group hostility, though we have not attempted this ambitious project.

PROSOCIAL EMOTIONS: MODELS AND EXPERIMENTAL EVIDENCE

As we have argued above, adherence to social norms is underwritten not only by the cognitively mediated pursuit of self-interest but also by emotions. Shame, guilt, empathy, and other visceral reactions play a central role in sustaining cooperative relations. The puzzle is that prosocial emotions are at least *prima facie* altruistic, benefiting others at a cost to oneself. Thus, under any payoff-monotone dynamic in which the self-regarding trait tends to increase in frequency, prosociality should atrophy.

Pain is a presocial emotion. Shame is a social emotion: a distress that is experienced when one is devalued in eyes of one's consociates because of a value that one has violated or a behavioral norm that one has not lived up to.

Does shame serve a purpose similar to that of pain? If being socially devalued has fitness costs, and if the amount of shame is closely correlated with the level of these fitness costs, then the answer is affirmative. Shame, like pain, is an aversive stimulus that leads the agent experiencing it to repair the situation that led to the stimulus and to avoid such situations in the future. Shame, like pain, replaces an involved optimization process with a simple message: whatever you did, undo it if possible, and do not do it again. Of course, the individual can

override the unpleasurable shame sensation if the benefits are sufficiently great, but the emotion nevertheless, on average, will reduce the frequency of shame-inducing social behaviors.

Since shame is evolutionarily selected and is costly to use, it must on the average confer a selective advantage on those who experience it. Two types of selective advantage are at work here. First, shame may raise the fitness of an agent who has incomplete information (e.g., as to how fitness-reducing a particular antisocial action is), limited or imperfect information-processing capacity, and/or a tendency to undervalue costs and benefit that accrue in the future. Probably all three conditions conspire to react suboptimally to social disapprobation in the absence of shame, and shame brings us closer to the optimum. The role of shame in alerting us to negative consequences in the future, of course, presupposes that society is organized to impose those costs on rule violators. Shame may have coevolved with the emotions motivating punishment of antisocial actions (the reciprocity motive in our model).

The second selective advantage to those experiencing shame arises through the effects of group competition. Where the emotion of shame is common, punishment of antisocial actions will be particularly effective and, as a result, seldom used. Thus groups in which shame is common can sustain high levels of group cooperation at limited cost and will be more likely to spread through interdemic group selection. Shame thus serves as a means of economizing on costly within-group punishment.

Shame can be investigated in the laboratory. In Bowles and Gintis (2002) we consider a public goods game where agents maximize a utility function that captures five distinct motives: personal material payoffs, one's valuation of the payoffs to others, which depend both on one's altruism and one's degree of reciprocity, and one's sense of guilt or shame when failing to contribute one's fair share to the collective effort of the group. Shame is evident if players who are punished by others respond by behaving more cooperatively than is optimal for a material payoff-maximizing agent. We present indirect empirical evidence suggesting that such emotions play a role in the public goods game. However, direct evidence on the role of emotions in experimental games remains scanty.

INTERNALIZATION OF NORMS

An *internal norm* is a pattern of behavior enforced in part by internal sanctions, including shame and guilt as outlined above. People follow internal norms when they value certain behaviors for their own sake in addition to, or despite, the effects these behaviors have on personal fitness and/or perceived well-being. The ability to internalize norms is nearly universal among humans. While widely studied in the sociology literature (socialization theory), it has been virtually ignored outside this field (but see Caporael et al. 1989 and Simon 1990).

Socialization models have been strongly criticized for suggesting that people adopt norms independent of their perceived payoffs. In fact, people do not

always blindly follow the norms that have been inculcated in them; instead, at times, they treat compliance as a strategic choice (Gintis 1975). The "oversocialized" model of the individual presented in the sociology literature can be counteracted by adding a phenotypic copying process reflecting the fact that agents shift from lower to higher payoff strategies (Gintis 2003b).

All successful cultures foster internal norms that enhance personal fitness, such as future orientation, good personal hygiene, positive work habits, and control of emotions. Cultures also universally promote altruistic norms that subordinate the individual to group welfare, fostering such behaviors as bravery, honesty, fairness, willingness to cooperate, and empathy with distress of others.

Given that most cultures promote cooperative behaviors, and if we accept the sociological notion that individuals internalize norms that are passed to them by parents and other influential elders, it becomes easy to explain human cooperation. If even a fraction of society internalized the norms of cooperation and punished free riders and other norm violators, a high degree of cooperation could be maintained in the long run. Thus we are left with two puzzles: why do we internalize norms, and why do cultures promote cooperative behaviors?

We provide an evolutionary model in which the capacity to internalize norms develops because this capacity enhances individual fitness in a world in which social behavior has become too complex and multifaceted to be fruitfully evaluated piecemeal through individual rational assessment (Gintis 2003a). Internalization moves norms from constraints that one can treat instrumentally toward maximizing well-being to norms that are then valued as ends rather than means. It is not difficult to show that if an internal norm enhances fitness, then for plausible patterns of socialization, the allele for internalization of norms is evolutionarily stable.

We (Gintis 2003a) use this framework to model Herbert Simon's (1990) explanation of altruism. Simon suggested that altruistic norms could "hitchhike" on the general tendency of internal norms to be fitness enhancing. However, Simon provided no formal model of this process and his ideas have been widely ignored. This chapter shows that Simon's insight can be analytically modeled and is valid under plausible conditions. A straightforward gene–culture coevolution argument then explains why fitness-reducing internal norms are likely to be prosocial as opposed to socially harmful: groups with prosocial internal norms will outcompete groups with antisocial, or socially neutral, internal norms.

CONCLUSION

Two themes run through our account of the origin of cooperation among humans: (a) the importance of groups in human evolution and the power of multilevel selection; (b) the underlying dynamic of gene–culture coevolution. We close with comments on what we consider two mistaken approaches: the tautological extension of self-interest to the status of the fundamental law of

evolution and the representation of culture as an epiphenomenal expression of the interaction of genes and environments.

Like de Tocqueville's "Americans," a distinguished tradition in biology and the social sciences has sought to explain cooperative behavior "by the principle of self-interest, rightly understood." From J.B.S. Haldane's quip that he would risk his life to save eight drowning cousins to the folk theorem of modern game theory, this tradition has clarified the ways that relatedness, repeated play, and other aspects of social interactions among members of a group might confer fitness advantages on those engaging in seemingly unselfish behaviors. The point is sometimes extended considerably by noting that if the differential replication of traits by selection operating on either culturally or genetically transmitted traits is monotonic in payoffs, only traits that on average have higher payoffs will be evolutionarily successful. If selfish behaviors are then *defined* as those that on average have higher payoffs, the principle of self-interest becomes the fundamental law of evolution.

Some prominent researchers in evolutionary biology have taken precisely this tack. Richard Dawkins (1989), for instance, states in the course of the first four pages of *The Selfish Gene* that "a predominant quality to be expected in a successful gene is ruthless selfishness. This gene selfishness will usually give rise to selfishness in individual behavior .... Let us try to *teach* generosity and altruism, because we are born selfish."[3] Similarly, drawing out the philosophical implications of the evolutionary analysis of human behavior, Richard Alexander (1987) says, "ethics, morality, human conduct, and the human psyche are to be understood only if societies are seen as collections of individuals seeking their own self-interest ... That people are in general following what they perceive to be their own interests is, I believe, the most general principle of human behavior." (pp. 3, 35).

Like de Tocqueville, we object to the tautological extension of the principle of self-interest. Our concern is not with the fitness-based or other payoff monotonic dynamic process assumed in this approach. It goes without saying that traits that experience lower fitness in a population will be handicapped in any plausible evolutionary dynamic: even cultural evolution may be strongly biased toward proliferation of behaviors leading to individual material success. Rather, our concern is with the distortion of the term "self-interest." Those who, in Darwin's words, were "ready to warn each other of danger, to aid and defend each other" would tautologically be deemed "selfish" if, as Darwin (1871/1973) suggested, tribes in which these behaviors were common would "spread and be victorious over other tribes." We have eschewed the terms "selfish" or "self-interested" to avoid confusions and have instead defined cooperative behaviors in terms of their costs to the individual and their beneficial consequences for group

[3] Note the tendency in this last sentence to identify self-interest with on-average higher payoffs, but then the use of the term in its everyday sense, which is completely unwarranted.

members. Our models and simulations show that these behaviors may proliferate under plausible conditions as the result of the group structure of human populations and success of groups in which cooperators are common.

Turning to our second point, we note that reduction of culture to an effect of the interaction of genes and natural environments is a common, if rarely explicit, aspect of accounts from such diverse authors as Karl Marx and some modern-day sociobiologists. Like the principle of self-interest, the hypothesis that the interaction of natural environments and genes affects the evolution of cultures has yielded numerous insights. It is also true, however, that culture affects the natural and social environments in which the relative fitness of genetically transmitted behavioral traits is determined. Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1985), Durham (1991) and others have provided compelling examples of these cultural effects on genetic evolution. Our own models of the coevolution of genetically transmitted individual behaviors and culturally transmitted group-level institutions are but some of the many models of this process. In one of our models, for example, we have seen that the presence of a culturally transmitted convention (resource sharing) is essential to the evolution of a genetically transmitted altruistic trait governed by natural selection. It may be helpful to represent human cultures, especially the institutional structures they support, as a case of niche construction, i.e., the creation of a particular environment such that genetic evolution is affected (Laland et al. 2000; Bowles 2000).

The challenge of explaining the origins of human cooperation has led us to the study of the social and environmental conditions of life of mobile foraging bands and other stateless simple societies which arguably made up human society for most of the history of anatomically modern humans. The same quest has made noncooperative game theory (which assumes the absence of enforceable pre-play agreements) an essential tool. But as several authors have pointed out, most forms of contemporary cooperation are supported by incentives and sanctions based on a mixture of multilateral peer interactions and third party enforcement, often accomplished by the modern nation state. It would be modest and perhaps even wise to resist drawing strong conclusions about cooperation in the 21st century on the basis of our thinking about the origins of cooperation in the Late Pleistocene.

ACKNOWLEDGMENTS

We are grateful to Eric Alden Smith for helpful comments and to the Santa Fe Institute and the John D. and Catherine E. MacArthur Foundation for support of this research.

REFERENCES

Alexander, R.D. 1979. Biology and Human Affairs. Seattle: Univ. of Washington Press.
Alexander, R.D. 1987. The Biology of Moral Systems. Hawthorne, NY: de Gruyter.
Boehm, C. 1982. The evolutionary development of morality as an effect of dominance behavior and conflict interference. *J. Soc. Biol. Struct.* **5**:413–421.
Bowles, S. 2000. Economic institutions as ecological niches. *Behav. Brain Sci.* **23**.

Bowles, S. 2001. Individual interactions, group conflicts, and the evolution of preferences. In: Social Dynamics, ed. S.N. Durlauf and H.P. Young, pp. 155–190. Cambridge, MA: MIT Press.
Bowles, S. 2003. Microeconomics: Behavior, Institutions, and Evolution. Princeton, NJ: Princeton Univ. Press.
Bowles, S., and H. Gintis. 2000. Persistent parochialism: The dynamics of trust and exclusion in networks. Santa Fe Institute Working Paper 00-03-017. Santa Fe, NM: Santa Fe Institute.
Bowles, S., and H. Gintis. 2002. Prosocial emotions. Santa Fe Institute Working Paper 02–07–028. Santa Fe, NM: Santa Fe Institute.
Bowles, S., J.-K. Choi, and A. Hopfensitz. 2003. The coevolution of individual behaviors and group level institutions. *J. Theor. Biol.*, in press
Boyd, R., H. Gintis, S. Bowles, and P.J. Richerson. 2003. Evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* **100**:3531–3535.
Boyd, R., and P.J. Richerson. 1985. Culture and the Evolutionary Process. Chicago: Univ. of Chicago Press.
Caporael, L., R. Dawes, J. Orbell, and J.C. van de Kragt. 1989. Selfishness examined: Cooperation in the absence of egoistic incentives. *Behav. Brain Sci.* **12**:683–738.
Cavalli-Sforza, L.L., and M.W. Feldman. 1981. Cultural Transmission and Evolution. Princeton, NJ: Princeton Univ. Press.
Darwin, C. 1871/1973. The Descent of Man. New York: Appleton Press.
Dawkins, R. 1989. The Selfish Gene. 2d ed. Oxford: Oxford Univ. Press.
Durham, W.H. 1991. Coevolution: Genes, Culture, and Human Diversity. Stanford: Stanford Univ. Press.
Eibl-Eibesfeldt, I. 1982. Warfare, man's indoctrinability and group selection. *J. Comp. Ethnol.* **60**:177–198.
Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. *Nature* **415**:137–140.
Fehr, E., U. Fischbacher, and S. Gächter. 2002. Strong reciprocity, human cooperation and the enforcement of social norms. *Nature* **13**:1–25.
Frank, R.H. 1987. If *Homo economicus* could choose his own utility function, would he want one with a conscience? *Am. Econ. Rev.* **77**:593–604.
Fudenberg, D., and E. Maskin. 1990. Evolution and cooperation in noisy repeated games. *Am. Econ. Rev.* **80**:275–279.
Gintis, H. 1975. Welfare economics and individual development: A reply to Talcott Parsons. *Qtly. J. Econ.* **89**:291–302.
Gintis, H. 2000a. Game Theory Evolving. Princeton, NJ: Princeton Univ. Press.
Gintis, H. 2000b. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**:169–179.
Gintis, H. 2003a. The hitchhiker's guide to altruism: Gene–culture coevolution and the internalization of norms. *J. Theor. Biol.* **220**:407–418.
Gintis, H. 2003b. Solving the puzzle of human prosociality. *Ration. Soc.* **15**.
Gintis, H., E.A. Smith, and S. Bowles. 2001. Costly signaling and cooperation. *J. Theor. Biol.* **213**:103–119.
Hirshleifer, J. 1987. Economics from a biological viewpoint. In: Organizational Economics, ed. J.B. Barney and W.G. Ouchi, pp. 319–371. San Francisco: Jossey-Bass.
Laland, K., F.J. Olding-Smee, and M. Feldman. 2000. Group selection: A niche construction perspective. *J. Consc. St.* **7**:221–224.
Mealey, L. 1995. The sociobiology of sociopathy. *Behav. Brain Sci.* **18**:523–541.
Price, G.R. 1970. Selection and covariance. *Nature* **227**:520–521.
Simon, H. 1990. A mechanism for social selection and successful altruism. *Science* **250**:1665–1668.