Biology and Philosophy (2005) 20:767–789
DOI 10.1007/s10539-004-5155-2

# Punishment and the strategic structure of moral systems

CHANDRA SEKHAR SRIPADA
*Department of Philosophy, Rutgers University, 26 Nichol Avenue, New Brunswick NJ 08901, USA; (e-mail: sripada@eden.rutgers.edu)*

Received 6 October 2004; accepted in revised form 22 October 2004

**Key words:** Moral norms, Evolution of morality, Reciprocity, Punishment, Cooperation, Evolutionary game theory, Moral diversity

**Abstract.** The *problem of moral compliance* is the problem of explaining how moral norms are sustained over extented stretches of time despite the existence of selfish evolutionary incentives that favor their violation. There are, broadly speaking, two kinds of solutions that have been offered to the problem of moral compliance, the reciprocity-based account and the punishment-based account. In this paper, I argue that though the reciprocity-based account has been widely endorsed by evolutionary theorists, the account is in fact deeply implausible. I provide three arguments that suggest that moral norms are sustained by punishment, not reciprocity. But in addition to solving the problem of moral compliance, the punishment-based account provides an additional important theoretical dividend. It points the way for how theorists might build an evolutionary account of a feature of human groups that has long fascinated and troubled social scientists and moral philosophers – the existence of moral diversity.

Moral norms are universally present in all human groups, but the existence of norms presents an evolutionary puzzle. Many moral norms direct individuals to undertake actions that are not ostensibly in their selfish evolutionary interest. For example moral norms often require people to share, help others or sacrifice for the group. They also forbid certain specific kinds of behavior, for example violence, theft and adultery. Additionally, they provide boundaries and constraints on social relations in the form of rules of authority, hierarchy, status and kinship. Moral norms also express a range of other prohibitions, for example taboos and ritualistic rules that restrict people in multitudinous, sometimes burdensome ways.[1]

In all these cases, individuals may have a selfish interest in violating moral norms, at least on some occasions. Most obviously, a person may want to violate rules about sharing or sacrificing. But a person might also have a selfish interest in murdering a rival, stealing a neighbor's goods, skirting obligations to an authority figure, or otherwise escaping the multitude of prohibitions and taboos that regulate daily life.

[1] See Westermark (1937), especially chapter 8, and Edel and Edel (2000) for further discussion of the moral norms listed in this paragraph and other examples of moral norms commonly found in human groups.

If individuals often have a selfish interest in violating moral norms, then what accounts for their long-term stability? Why don't moral norms simply collapse from routine violation? I'll call the fact that moral norms don't in fact collapse in this way, and indeed that they are routinely complied with, and have been for perhaps tens of thousands of years, '*the problem of moral compliance*.' The problem of moral compliance is a fundamental problem in the evolutionary-focused investigation of morality. Moral norms regulate a host of domains in a way that makes an *enormous* impact on individual-level reproductive fitness. To the extent that individuals have a selfish evolutionary interest in violating moral norms, then the fact that they do in fact routinely comply with norms is genuinely puzzling and requires explanation.

Despite the importance of the problem of moral compliance, it has not been *directly* discussed or dealt with by evolutionary-minded theorists. Instead, most theorists interested in the origins and workings of morality have focused on a related, though, as we shall see, importantly different, problem, *the problem of cooperation*. Very roughly, the problem of cooperation is the problem of how people sustain cooperative outcomes in situations called *collective action problems*. In such problems, there is a conflict between collective benefit and selfish interest. Theorists use models derived from *game theory* to explain, in a mathematically precise way, how cooperation can be sustained against selfish incentives to free ride.

Broadly speaking, there are two kinds of solutions to the problem of cooperation that have loomed large in the literature. One approach is the so-called *reciprocity-based account*. A number of theorists have shown that altruistic actions, actions that confer a benefit on others at one's own expense, can be sustained if helping others is made contingent on receiving like acts of helping in return. A second solution to the problem of cooperation is the so-called *punishment-based account*. According to this account, individuals cooperate because the threat of punishment makes it in their selfish interest to do so. The punishment-based account has only recently been made precise in the form of evolutionary models. Thus, it is less well known among evolutionary-minded theorists.

Theorists frequently suggest that game-theoretic models designed to explain cooperation in the context of collective action problems can be generalized to provide an explanation for the origins and operation of human moral systems more broadly (and I'll discuss these suggestions in a later section). Unfortunately, these suggestions are almost never made very precise. So an important question remains: What exactly is the relationship between game-theoretic models of cooperation and human moral systems? The aim of this paper, is to systematically address this question. In the course of this paper, I'll defend two main claims. First, the strategic structure of the problem of cooperation is importantly different than the problem of moral compliance. Second, once this difference is made explicit, it is clear that the punishment-based account, and not the reciprocity-based account, provides the correct solution to the problem of moral compliance.

This paper is divided into five parts. In part I, I'll clarify the problem of cooperation and collective action problems. In part II, I'll introduce the reciprocity-based account of cooperation. I'll then attempt to generalize the reciprocity-based account to the problem of moral compliance. I'll show that the reciprocity-based account suffers from three problems that are more or less decisive against the account as a solution to the problem of moral compliance. In part III, I'll introduce and explain the alternative punishment-based account of cooperation. I'll show that the punishment-based account can be appropriately generalized as a solution to the problem of moral compliance, and that it is empirically plausible as a solution to the problem of moral compliance in its own right. But the punishment-based account raises an important question: What makes it the case that self-interested agents will carry out costly punishment? In part IV, I'll discuss some of the theoretical and empirical issues that are relevant to how costly punishment is supported. In the final section of the paper, I'll suggest that in addition to solving the problem of moral compliance, the punishment-based account provides an additional theoretical dividend. It points the way for how theorists might eventually build an evolutionary account of a feature of human groups that has long fascinated and troubled social scientists and moral philosophers – the existence of *moral diversity*.

## Part I: Collective action problems and the problem of cooperation

A collective action problem is a situation in which there are at least two alternative actions available to individuals, where these two actions have the following features. One of the actions, 'cooperate,' provides a benefit to everyone at a cost to the cooperator, such that the cost to the cooperator exceeds the selfish benefits she receives from her own act of cooperation. The other action, 'defect,' provides no collective benefit, and costs nothing for the defector. The benefits from cooperation are such that if everyone (or most everyone) chooses to cooperate, then everyone will be *substantially* better off, by her own estimation, than in an alternative case where everyone defects. Cooperation can nevertheless be difficult to sustain. The problem is that from the perspective of each individual, it appears that she is better off choosing to defect, *regardless of what the others choose*. But if everyone reasons this way, then cooperation collapses and everyone receives an outcome far worse than had they all cooperated. Thus collective action problems are situations in which there is a conflict between collective benefit and selfish interest. The *problem of cooperation* is how can cooperation be sustained despite each individual's temptation to free ride?

For the past 30 years, theorists in many disciplines have studied the problem of cooperation in the context of collective action problems using models derived from game theory. Figure 1 depicts the pay-off matrix for a two-person collective action problem. According to this matrix, if both players choose cooperate, they both receive a payoff of 3. If both choose defect, they receive a

**Player B**

|  |  | Cooperate | Defect |
|---|---|---|---|
| **Player A** | Cooperate | A=3, B=3 | A=0, B=5 |
|  | Defect | A=5, B=0 | A=1, B=1 |

*Figure 1.* Prisoner's dilemma.

payoff of 1. If player A chooses cooperate while player B chooses defect, A receives the worst possible pay-off, 0, while B receives the best payoff, 5 (and if B chooses cooperate while A chooses defect, these payoffs are reversed). Thus the matrix specified here captures the underlying strategic structure of a collective action problem. Games whose payoff matrices capture the structure of a collective action problem in this way are called 'Prisoner's Dilemmas.'[2]

## Part II: The Reciprocity-based account

### *Reciprocity as a solution to the problem of cooperation*

The reciprocity-based account of cooperation has an extensive history in a number of disciplines. In this section, I'll focus on the way the account has been elaborated by evolutionary-minded theorists in particular. In 1971, Robert Trivers published a seminal paper on how altruistic behavior in the animal world might be sustained by means of reciprocity (Trivers 1971). One of Trivers important innovations was that he conceptualized the problem of sustaining altruism as a two-person Prisoner's Dilemma. He recognized that in a Prisoner's Dilemma situation, if interactions occurred just once, cooperation cannot be sustained as both players will have reason to defect regardless of what the other player chooses. But if interactions are *repeated*, agents can make their cooperation contingent on cooperation from the other party, and make defection contingent on acts of defection from the other party. This pattern of 'like actions beget like actions' is the hallmark of reciprocity. Finally, Trivers showed that the strategy of reciprocity could yield a net long-term evolutionary gain for a player vs. the strategy of not engaging in reciprocity. Reciprocators receive a large benefit when interacting with other reciprocators, which, under the appropriate conditions, can more than make up for losses suffered when they interact with defectors.

Building on the work of Trivers, Robert Axelrod pioneered the use of *evolutionary modeling* to study the evolutionary dynamics of strategies for the two-person repeated Prisoners' Dilemma. In a *classical* game, one assumes that players are ideally rational deliberators, who possess sophisticated knowledge of the game, its payoffs, and the beliefs of the other agents. In an *evolutionary*

[2] It is usually stipulated that in a Prisoner's Dilemma, the average payoffs of the cooperate-defect and defect-cooperate outcome must be less than the payoffs of the cooperate–cooperate outcome.

game, these rationality and knowledge assumptions are dispensed with. Instead, in an evolutionary game, there is a large population of players who each deploy a single fixed strategy. Players are randomly picked from this population to interact in a game, and the pay-offs they receive determine their representation in the population in subsequent generations. For example, players that do better vs. the existing pool of other players are 'selected,' and they are represented in greater numbers in the next generation, while players that fare poorly against the existing pool of players are selected against.

One strategy that did particularly well in Axelrod's study was 'Tit-for-tat,' a strategy that embodies the fundamental idea behind reciprocity (Axelrod 1984). Tit-for-tat always cooperates with its partner on the first round of the repeated Prisoner's Dilemma. Thereafter, it plays cooperate if its partner cooperates on the previous round, and plays defect if its partner defects on the previous round.

The fact that Tit-for-tat follows the principle of reciprocity, and makes its cooperation *contingent* on like cooperation from the other party on the previous round, plays an important role in explaining its success in Axelrod's study. Strategies which don't make their cooperation contingent on like cooperation from the other party, for example, an indiscriminate altruist that just always cooperates, will be exploited by defectors on every interaction. However, Tit-for-tat cannot be exploited in this way because defectors instead find that their defection is greeted with like acts of defection. Axelrod and Hamilton (1981) showed that the mechanism of reciprocity can indeed be a powerful mechanism for stabilizing cooperation. In particular, they used evolutionary modeling to show that in a population in which the frequency of Tit-for-tat approaches one, rare mutant strategies that deploy alternative strategies, including the simple non-cooperative strategy that defects in all interactions (so-called All Defect), cannot displace Tit-for-tat. The technical term for this property of Tit-for-tat is that it is an *evolutionarily stable strategy*, or ESS (Maynard Smith and Price 1973).[3] When Tit-for-tat is common, it remains common. Rare strategies that don't follow the principle of reciprocity, like All Defect, cannot exploit Tit-for-tat and are ultimately weeded away.

The work of Trivers and Axelrod became instant classics among scholars and also, interestingly, among the wider public. The appeal of the idea of reciprocity is hardly surprising. Darwinian theory, rightly or wrongly, had historically been associated with a picture in which organisms are incessantly engaged in a ruthless competition to survive – nature red in tooth and claw. The work of Trivers and Axelrod was the first to show in a mathematically precise way how this picture was importantly wrong, and that certain kinds of altruistic behavior could in fact be sustained (among unrelated individuals) despite Darwinian pressures to maximize one's own selfish evolutionary

[3] Tit-for-tat is not actually itself an ESS. However, several strategies closely related to Tit-for-tat, including so-called 'contrite Tit-for-tat' are ESS's in environments in which players at least occasionally make mistakes (Boyd 1989). For the purposes of this paper, this complication can safely be ignored.

interests. Indeed, Trivers and Axelrod showed that helping behavior is actually *favored* for Darwinian reasons. When helping others is made contingent on like acts of helping, helping others is to one's own evolutionary advantage![4]

### *Reciprocity as a solution to the problem of moral compliance*

Recall the problem of moral compliance, the problem of how moral norms are supported against incentives to deviate, which I posed at the start of this essay. What is the relationship between the reciprocity and the problem of moral compliance? Many contemporary evolutionary theorists seem to endorse the idea that reciprocity can be generalized from its role as a solution to the problem of cooperation to serve as a solution to the problem of moral compliance as well. Though these claims are seldom made very precise, the attraction of the idea is clear. If human moral systems are supported by reciprocity, then compliance with morality wouldn't be puzzling – compliance would be in one's own long-term evolutionary interests.

The idea that morality is anchored by reciprocity actually has a fairly old pedigree. David Hume provided an early account of how rules of property are sustained by reciprocity.

> It is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods, provided he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behavior. And this may properly enough be call'd a convention or agreement betwixt us, tho' without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are perform'd upon the supposition, that something is to be perform'd on the other part (Hume 1992 [1739], p. 490).

Hume's idea is an intuitively appealing one, and many evolutionary theorists have pursued the idea that morality is based on reciprocity in various ways. Trivers himself was the first to make suggestions along these lines. One of Trivers' most influential proposals was that reciprocity could serve to explain the origin and function of human *moral sentiments*. He argued that moral attitudes and emotions, for example friendship, hatred, gratitude, sympathy, and guilt could be explained as adaptations that emerge to regulate systems of

[4] There is some controversy about whether Tit-for-tat is maintained as an ESS by natural selection operating at the level of the individual, as opposed to natural selection operating at the level of the group. Here I follow the standard view that Tit-for-tat is maintained by individual-level selection. See Uyenoyama and Feldman (1992) and Sober and Wilson (1998) for further discussion.

reciprocity (Trivers 1971, p. 49). Trivers also proposed that reciprocity could help explain the origin and function of *rules of exchange* that apply to larger groups. His theory was that humans living in close-knit groups form complex reciprocal relationships involving multiple other people. As systems of reciprocation become larger and more elaborate, rules of exchange would be codified to coordinate people's expectations about what kinds of behaviors are normatively acceptable Trivers 1971, p. 52).

Perhaps the most explicit and ambitious attempt to derive human morality from reciprocity was made by Richard Alexander, in his important work, *The Biology of Moral Systems* Alexander 1987). Alexander viewed human moral systems as systems of what he called *indirect reciprocity*. One of his original insights was that reciprocity might be arranged in the form of a chain. Person A helps person B and person B helps person C, and so on. Eventually, person A is in turn helped by someone else, say person X, who may never have been directly helped by person A. According to Alexander, such chains of indirect reciprocation serve to show how reciprocity can in fact explain more generalized kinds of moral behavior. One problem with Alexander's account is that he does not provide a formal model for how precisely this suggestion is supposed to work. Another problem is that Alexander frequently invokes various forms of *punishment*, for example reputational sanctions and ostracism, to explain how moral rules are maintained. Thus, it is unclear whether to interpret Alexander's work as an account of how the problem of moral compliance is solved by means of reciprocity, as opposed to other means.

Even if the notion of reciprocity isn't always made precise and the manner in which reciprocity is supposed to support moral systems isn't spelled out in detail, overall, the idea that reciprocity plays a crucial role in sustaining human moral systems is fairly widespread among evolutionary-minded theorists. Nevertheless, I believe that this idea is mistaken. In particular, I believe that if we understand reciprocity in terms of the models described by Trivers and Axelrod, then it cannot be the case that moral systems are supported by reciprocity. In what follows, I'll argue that reciprocity is untenable as a solution to the problem of moral compliance in its full generality because of three basic problems. I call the first two problems the *scaling-up problem* and the *incompleteness problem*, and I'll explain them in the next two sections. The third problem is one of *empirical inadequacy* – reciprocity makes predictions about how moral norms are enforced that are contravened by the evidence.

## *Three problems for reciprocity as a solution to the problem of moral compliance*

### *Problem 1: The scaling-up problem*

The work of Trivers and Axelrod, and related work by other theorists, provided a compelling case that reciprocity can sustain cooperation in the context of collective action problems restricted to *two players*. Thus reciprocity is plausible as a mechanism by which humans cooperate in dyadic contexts, for example the reciprocal exchange of goods and services that routinely occurs in human groups (and there is solid empirical evidence that this is in fact the case). But can reciprocity be scaled up to sustain cooperation in the context of collective action problems with many players?

The reason this question is important is because moral norms often apply to large numbers of individuals interacting collectively. For example, people in human groups routinely follow group-level moral norms that require collective defense of the group, or collective hunting and resource distribution (Cashdan 1980; Boehm 1999). People also follow moral norms that require participation in community works projects such as home building and forest clearing, which are well described in the ethnographic literature (see Fiske 1991). Additionally, people also follow moral norms that regulate so-called *common pool resources*, i.e. resources such as land, water, and plant and animal species that are protected from over-utilization by rules that allocate to each person of the group a limited share (Ostrom 1990). The existence of moral norms that apply to large numbers of individuals interacting collectively suggests that in order to solve the problem of moral compliance, we must account for cooperation in large groups, and not just cooperation in dyads.

Unlike in the two-person case, sustaining cooperation in large groups by means of reciprocity has been found to be deeply problematic. The main problem is that reciprocity relies on non-reciprocation as its only form of deterrence, and non-reciprocation is a highly '*unselective*' deterrent. We can illustrate the notion of a deterrent being selective vs. unselective by considering a case. Suppose 10 players will interact repeatedly in a Prisoner's Dilemma, and in order to sustain cooperation by reciprocity, the 10 players agree to adopt the following strategy: Each person will cooperate in the first round and will continue to cooperate in subsequent rounds conditional on cooperation from *all* the others in previous rounds. Suppose further that in the second round one player defects while the other nine cooperate. In this case, the strategy adopted by these players requires that the nine cooperators must themselves defect in the subsequent rounds. But by defecting in subsequent rounds, the nine cooperators unselectively generate a harm not only for the one defector, but also for the other cooperators – the outcome is worse for *everybody*. A selective deterrence is one that can be deployed exclusively against defectors, without creating a harm for others. Non-reciprocation is a highly unselective form of deterrence.

Because non-reciprocation is unselective in this way, the mechanism of sustaining cooperation by reciprocity in large groups encounters multiple problems, in particular if we focus on evolutionary games. The first problem is that in the context of large groups, reciprocators must be highly '*strict*' – they must cooperate only if *everyone* else in the group cooperates. Suppose they are less than strict and continue to cooperate even if just one or two others defect. Then these defectors who are generously allowed to reap the rewards of cooperation will multiply and inevitably displace the existing population of cooperators over time. So a necessary condition for sustaining cooperation in large groups is that the cooperators must be strict and cease cooperation if there is even a single defector in their midst (Boyd and Richerson 1989 provide a simulation that confirms this informal argument). But highly strict reciprocators, in particular, face a number of problems. One problem is that strict reciprocators will cooperate only if the formation of groups is extremely *homogenous* so that no defectors are present whatsoever. But obtaining such homogenous groups is typically not feasible in real world conditions. Another problem is that even in groups composed exclusively of strict reciprocators, the resulting cooperative equilibrium is extremely sensitive to *errors*. For example, if we assume that players occasionally make errors (due to poor information or execution errors, etc.) then the model predicts that if one person defects due to an error, everyone else defects in retaliation, cooperation collapses and everyone is worse off; *and this pattern of collapse occurs each time such errors are made*. To sum up, in large groups, reciprocity requires stringent conditions in order to operate, and even then it is liable to repeated collapse due to errors. For these reasons, sustaining cooperation in large groups by means of reciprocity is for all practical purposes unfeasible.

*Problem 2: The incompleteness problem*

A second problem for reciprocity as a solution to the problem of moral compliance is what I call the *incompleteness problem*. In any given human group, there will be myriad moral norms that apply to disparate domains of social life. Some moral norms prescribe cooperative solutions to collective action problems and we can call these ‘CAP moral norms.’ A moral norm that requires that successful hunters should share meat with unsuccessful hunters is an instance of a CAP moral norm (assuming each hunter has a more or less equal chance of success on any given day). The work of Trivers and Axelrod and others shows that reciprocity can sustain compliance with CAP moral norms, though, as we’ve seen, their account founders in large groups.

But there are a large number of moral norms that don’t regulate collective action problems and thus are not CAP moral norms. To give an example, virtually all human groups have moral norms that forbid consanguineous sexual relations within the nuclear family, i.e. incest (see Murdock 1949). But a moral norm forbidding incest is, of course, not a cooperative norm that applies to a collective action problem. For this reason, we can call this norm that forbids incest a ‘non-CAP’ moral norm. Non-CAP moral norms are extremely common. Many moral norms that pertain to social domains such as violence, adultery, sexual behavior, kin relations, status relations, authority relations, food habits, ritualistic practices and many others, are non-CAP moral norms. Just as in the case of CAP moral norms, individuals routinely possess selfish incentives to deviate from non-CAP moral norms. So what accounts for compliance with non-CAP moral norms?

As it turns out reciprocity is unable to explain compliance with non-CAP moral norms. To see this, let us return to the case of the moral norm forbidding incest. This norm cannot be sustained by reciprocity. If a person desires to violate an incest prohibition, this person cannot be deterred by reciprocal violation of the prohibition by others. Reciprocal violation of this sort simply does not act as a deterrent. Consider another example. Suppose there is a moral norm in a community against abusing one's child. This norm cannot be sustained by reciprocity because I cannot deter my neighbor from beating his child by beating my own child. Other examples include moral norms that prohibit the harming of certain animals, forbid certain kinds of sexual activities or taboo the consumption of certain food items. In all these cases, it is not possible to enforce the moral norm by means of the threat of reciprocal defection. Examples like these could be multiplied easily.

What these examples show is that there are many moral rules whose payoff structure for complying with the rule vs.violating the rule are such that reciprocity could never, *even in principle*, explain how compliance with these rules is sustained. Reciprocity operates on the principle of like begets like – compliance begets compliance and defection begets defection. Moral norms that apply to collective action problems have a very specific pay-off structure that permits cooperative solutions to be sustained by reciprocity. But the crucial point is that moral norms that apply to situations that are not collective action problems, like the ones cited above, don't have this payoff structure for compliance and violation, and thus can't be sustained by reciprocity.

It is a significant liability for reciprocity-based accounts that they can't even in principle explain compliance with many kinds of moral norms. The reason is that moral norms appear to be a *unified kind.* There is no evidence that there are fundamental cleavages among moral norms as they apply to some domains vs. other domains in the way that norms are complied with, enforced, and in the way that the mechanisms which underwrite norm psychology acquire and utilize norms. For example, it is not the case that people conceptualize moral norms against assault, promise breaking and theft in fundamentally different ways than moral norms that deal with sharing meat.[5] It is a striking fact that moral rules with very different pay-off structures for compliance and violation are all supported by what appears to be a unified mechanism. But reciprocity cannot be this unified mechanism since it is incapable of sustaining many kinds of moral norms.

*Problem 3: Empirical inadequacy*

The third problem with the reciprocity-based account as a solution to the problem of moral compliance is *empirical inadequacy*. The reciprocity-based account predicts a pattern of behavior in the context of enforcement of moral norms that is contravened by the evidence of how moral norms are actually enforced in human groups.

Consider a moral norm that regulates some common pool resource, for example the use of water from a local watering hole. Each family leader is

[5] See Sripada and Stich forthcoming, for further support for the claim that moral norms constitute a unified kind, and are underwritten by a unified psychology.

allotted a limited share of the resource; by limiting each person's share in this way, the resource is protected from over-exploitation. Suppose one person takes more than his or her allotted share of water. The reciprocity account predicts that an act of defection of this sort will lead to reciprocal acts of defection by others. That is, the others will also defect and attempt to take more than their allotted share of water as well. But this predicted scenario is simply not what actually happens in real world situations. Instead, a person that fails to follow moral norms will typically be *punished*. For example, the person who exceeds his or her allotted share of water may be criticized and condemned for being selfish or greedy. The person may also be excluded from receiving water, excluded from the group altogether, or even hit, hurt or harmed in other ways (see, for example, the case studies in Ostrom 1990). A large body of evidence indicates that it is punishment and not the threat of non-reciprocation that actually sustains moral norms in human groups. I'll discuss this evidence in the following section, where I introduce and clarify the *punishment-based account* of moral compliance.

Before moving on, let me make an important clarification. I am arguing against the hypothesis that the reciprocity-based account provides the correct account of how *moral norms* are sustained in human groups. I am not denying that reciprocity provides a fully adequate solution to many kinds of cooperative dilemmas. For example, I accept that in all human groups, reciprocity plays a crucial role in maintaining friendships, alliances, economic partnerships, especially when these interactions occur in dyads. So reciprocity-based social practices clearly exist and are important. What I am denying, quite specifically, is the widely held hypothesis that reciprocity plays an important role in supporting *moral norms*.

## Part III: The punishment-based account

### *Punishment as a solution to the problem of cooperation*

The core idea of the punishment-based account of cooperation is that cooperation is sustained by punishment for defection, making it in each person's selfish interest to cooperate. Among evolutionary-minded theorists, the punishment-based account has received relatively little attention. One reason is the early success of the work of Trivers and Axelrod, which had the effect of influencing many theorists to pursue reciprocity-related ideas. A second reason, perhaps just as important, is that many theorists don't recognize that there is a distinction to be made between reciprocity and punishment. For example, I noted earlier that Richard Alexander freely intermixes reciprocity and punishment throughout his discussion of how moral systems are sustained. So an important goal of this part of this paper will be to clarify the structure of the punishment-based account of cooperation and show how it is importantly distinct from the reciprocity-based account.

We can illustrate how the punishment-based account of cooperation works by examining in detail a model developed in Boyd and Richerson (1992). Consider the following game that incorporates punishment as a mechanism for sustaining cooperation: A large number of players engage in an indefinitely repeated game. Each constituent game of the repeated game has two phases, a cooperation phase and a punishment phase. The cooperation phase has the payoffs of a standard Prisoner's Dilemma, and players choose to either Cooperate or Defect. In the punishment phase, each player chooses either to Punish or Not Punish each other player, conditional on that player's previous actions. Punishment is a costly act, and it costs $k$ for the punisher, while it produces harm $h$ for the punished. We can call games broadly structured along these lines *punishment-based games*, and the structure of the game is depicted in Figure 2.

Boyd and Richerson analyzed strategies for sustaining cooperation in a repeated punishment-based game with $n$ players, where $n$ is *an arbitrarily large number* (Boyd and Richerson 1992). They demonstrated that the strategy that cooperates and punishes non-cooperators (and also stabilizes punishment by punishing non-punishers) is an ESS, even when the number of players is very large.

Punishment-based models are different from reciprocity-based models in several ways. Perhaps the most important difference is that in the case of punishment, the relationship between punishment and the behavior that punishment is used to enforce can be quite *arbitrary*, while this is not the case with reciprocity. The basic principle in reciprocity is *like begets like* – cooperation begets cooperation and defection begets defection. Punishment operates on a very different basic principle, *defection begets harm.* For example, if a person fails to share meat with the group, the person may be punished by being beaten or publicly humiliated – the relationship between the violation and the harm imposed on the violator need not be connected in the 'like begets like' fashion of reciprocity.

Another difference between punishment and reciprocity is that reciprocity-based models are structured around a Prisoner's Dilemma, as depicted in Figure 1. But the punishment-based game depicted in Figure 2 is not itself a

**Player B**

**Cooperation Phase**

| **Player A** | | Cooperate | Defect |
|---|---|---|---|
| | Cooperate | A= 3, B= 3 | A= 0, B= 5 |
| | Defect | A= 5, B= 0 | A= 1, B= 1 |

**Punishment Phase**

Each player chooses to either Punish or Not Punish each other player (conditional on that player's previous actions). Punishment costs $k$ for the punisher, and reduces the punished player's pay-offs by $h$.

*Figure 2.* Punishment-based game. Note: This is an $n$-player game. Since $n$-player matrices cannot be rendered easily in 2-dimensions, only two players are depicted.

Prisoner's Dilemma because there is a second phase of the game, the so-called punishment phase, that is not present in a Prisoner's Dilemma game. While the punishment-based game shown in Figure 2 *does* have a Prisoner's Dilemma as a constituent, later I'll show that it is wholly inessential to the model that it has a Prisoner's Dilemma as a component (see section ' Punishment as a solution to the problem of moral compliance' below). Thus punishment-based models, suitably generalized, have no essential connection to a Prisoner's Dilemma at all.

Another difference between reciprocity and punishment is that in the case of punishment, the harm that is applied to defectors can be *selective*, i.e., individual free riders can be harmed without other individuals also being harmed. As a consequence of selective punishment, free riders receive lower payoffs than those who cooperate and are weeded away. Thus punishment can sustain cooperation in large groups, whereas, as we've seen, reciprocity has trouble in this regard. Punishment is unique in that it is plausibly the *only* mechanism that can explain how cooperation is sustained in large groups of unrelated individuals.

Punishment and reciprocity also differ in the respect that punishment is invariably a *costly* action, whereas non-reciprocation typically isn't costly to carry out. Since punishment is invariably costly, there will always be incentives to avoid carrying out punishment. So what sustains costly punishment? In the preceding Boyd and Richerson model, incentives to avoid carrying out punishment are curtailed by the use of punishment for failing to punish. I call this strategy *higher-order punishment*, and I'll discuss it in Part IV.

An important bit of clarification remains. In the punishment-based game depicted in Figure 2, how is the action labeled 'Punish' to be interpreted in specific behavioral terms? Since in this model, punishment results in a *direct* reduction in the punished player's pay-offs, we can call the kind of punishment represented in this model *direct punishment*. Direct punishment corresponds to actions such as hitting, hurting, seizure of property, destruction of property, fining or any other action which directly lowers the pay-offs for another individual. As I'll argue in the next section, direct punishment routinely occurs in simple societies, and it can be quite devastating to the recipients.

There are other types of punishment that are importantly different from direct punishment. One that stands out as being particularly widespread is *exclusion-based punishment*, for example ostracism and banishment. Exclusion-based punishment involves removing a defector completely from an ongoing interaction among players, such that the removed player is excluded from interacting with *everybody*. Exclusion-based punishment such as ostracism and banishment are routinely used in human groups, and they can deliver potent harms to the recipients of the punishment (Brown 1991).

Perhaps the most commonly used kind of punishment in human societies is *reputation-based punishment*. In reputation-based punishment, when a person behaves uncooperatively, others sanction him or her by expressing attitudes of condemnation and blame. These attitudes can be expressed publicly, as in denouncing or shaming, or privately as in gossiping. All the players keep track of reputational information and use this valuable information as a basis for assortively forming groups with fellow cooperators, and avoiding being in groups with defectors. In addition, many other kinds of social interactions can also be based on reputation. For example, a person with a poor reputation may be avoided for the purposes of marriage or friendship, and may be regarded as having less status or may be denied certain privileges. Reputation-based punishment is extremely widespread because it can be quite potent in delivering harms to defectors, and yet it can also be relatively low-cost. But it's worth emphasizing that reputational punishment is certainly *not costless*. Those who reputationally sanction defectors essentially function as 'whistle-blowers,' and they pay costs in terms of monitoring and effort, as well as bearing the inevitable risk of acrimony and backlash from those whose reputation is smeared (see Horne 2001).

Both exclusion-based punishment and reputation-based punishment work, at least in part, by denying the punished person an opportunity to engage in materially beneficial relationships with others in the social group. In this respect, they are reminiscent of reciprocity models because they use the threat of *withholding benefits* to enforce certain behaviors. However, other aspects of exclusion and reputation-based punishment are much more in the spirit of punishment models. Exclusion and reputation-based punishment are *selective* and they are typically *costly* to impose. Most importantly, exclusion and reputation-based punishment can be used to enforce *arbitrary* behaviors. For example, a person who eats a tabooed food item may be excluded or reputationally sanctioned. In this case, exclusion and reputational sanctioning clearly function as a kind of punishment, and it would be quite odd to classify them as a form of non-reciprocation.[6]

The three categories of punishment I've described are extremely common, and can be realized by a *very wide variety of actual human behaviors*. Hitting, harming, destroying property, expulsion, excommunication, ostracism, extremely subtle forms of avoidance and non-inclusion, shaming, scolding, slandering, and rebukes of various forms and degrees all count as instances of punishment. Indeed, merely crooking one's eyebrow at another person at a public gathering in a way that conveys blame counts as punishment.[7]The fact that even subtle public displays can exact tremendous damage to another person's reputation suggests that such behavior rightfully counts as a form of punishment, on par with other behaviors whose potential for harm is much more obvious.

[6] It's worth emphasizing that the boundaries between reciprocity and punishment are certainly not sharp. While there are certain cases that are clear instances of punishment and certain cases that are clear instances of reciprocity, there will inevitably be cases that lie in between, and thus are difficult to classify.

[7] I thank Dan Sperber for this felicitous way of putting the point about the eyebrow crooking.

**Player B**

**Compliance Phase**

| | | Comply | Violate |
|---|---|---|---|
| **Player A** | Comply | A= ?, B= ? | A= ?, B= ? |
| | Violate | A= ?, B= ? | A= ?, B= ? |

**Punishment Phase**

Each player chooses to either Punish or Not Punish each other player (conditional on that player's previous actions). Punishment costs *k* for the punisher, and reduces the punished player's pay-offs by *h*.

*Figure 3*. Generalized version of punishement-based game. Note: This is an *n*-player game. Since *n*-player matrices cannot be rendered easily in 2-dimensions, only two players are depicted.

### *Punishment as a solution to the problem of moral compliance*

Earlier I listed three problems that the reciprocity-based account faced in its being generalized from an account of cooperation to a comprehensive account of moral compliance. As we've seen, punishment can sustain cooperation in large groups. Thus punishment avoids the first kind of problem that plagued reciprocity models, *the scaling-up problem*. But punishment can also deal with the second kind of problem that plagued reciprocity models, *the incompleteness problem*. Recall that reciprocity can only sustain compliance with a subset of moral norms, so-called 'CAP moral norms.' Punishment, however, provides a general mechanism for stabilizing any kind of moral norm. The crucial feature of punishment that allows this flexibility is that punishment can bear an *arbitrary* relationship to the rule it is used to enforce. Thus, while punishment can be used to curtail incentives to violate rules that regulate collective action dilemmas, it needn't be restricted to enforcing just these rules. Punishment can also be used to curtail incentives to violate *just about any rule*.

We can represent the strategic structure of the problem of enforcing *any arbitrary social rule* by modifying the punishment-based model depicted in Figure 3. In order to represent compliance with arbitrary rules, we need to replace the first phase of the constituent game, the so-called 'cooperation phase' with another game. In this game, players have two actions, Comply and Violate. The content of any social rule can then be abstractly represented in terms of the pay-offs that ensue to each player given that each other player either chooses to comply with or violate the rule. We keep the second phase, the punishment phase, as it is. The result is a modified punishment-based model in which compliance with any arbitrary social rule can be represented.

In this modified punishment-based game, it's nevertheless probably the case that so long as the punishment is severe enough, there is an evolutionary equilibrium in which everyone follows the rule and punishes rule violators, and each individual is made *strictly worse off* by unilateral deviation (Boyd and Richerson 1992).[8] Of course, for this result to obtain, punishment must itself be

[8] This result also follows from the so-called 'folk theorems' of game theory. See Fudenberg and Maskin (1986)

stabilized by some mechanism, for example, by higher-order punishment. The intuitive reason why punishment can support just about any rule should be clear. Regardless of how the payoffs are structured for following or violating a rule, so long as punishment is severe enough, it overwhelms any selfish incentive to violate the rule, making rule compliance in each person's overall selfish interests. This is true even for rules that don't apply to collective action domains, for example social rules that regulate theft, murder, sexual relations, authority relations or food taboos.

Punishment can sustain moral norms that regulate collective action domains as well as non-collective action domains. But a further fact, which I noted in section 'Three problems for reciprocity as a solution to the problem of moral compliance,' is that there appears to be no fundamental cleavage between these two domains in the ways that rules are complied with and enforced, and in the psychology that underwrites these rules. Moral norms in both domains seem to be of a unified kind. It is a virtue of punishment models that they provide a unified explanation of what ostensibly appears to be a unified phenomenon.

Now let's turn to the third problem that faced the reciprocity-based account, the *problem of empirical inadequacy*. The punishment-based account of moral compliance has strong empirical support that arises from a number of sources. One important source of support comes from the ethnographic record, which reveals that moral norms are universally supported by punishment directed at those that violate norms (Roberts 1979; Black 1998; Sober and Wilson 1998; Boehm 1999). In particular, the three categories of punishment discussed earlier, direct punishment, exclusion-based punishment, and reputation-based punishment, are each universally present in all human groups and are used to enforce moral norms (Brown 1991; Dunbar 1997; Boehm 1999; Wilson et al. 2000).

Another source of evidence comes from experimental economics and other experimental disciplines. Recent studies have found that in various experimental situations and games, people reliably display punitive reactions in the context of violations of moral norms (Henrich et al. 2001). Also, consistent with the punishment-based account, a large and growing body of evidence indicates that people punish even if it is *costly* (Fehr and Gachter 2002). One study has even found that people engage in costly punishment when they are *merely observers* of norm violations (and are not directly harmed) (Carpenter et al. forthcoming). Finally, there is evidence that motivations to punish violations of moral norms are mediated by species-typical emotional reactions, in particular reactions of anger, disgust, and contempt (Fehr and Gachter 2002; see Haidt (2000) for a review). These findings are suggestive that punitive reactions have a robust, universal basis in innate human psychology. Overall, the accumulated empirical evidence strongly suggests that the punishment-based account correctly describes how moral norms are supported against incentives to deviate in real-world situations.

To sum up, the punishment-based account gives us a theoretically coherent and empirically plausible account of how compliance with moral norms is sustained. In this respect, it fares better than the reciprocity-based account, which faced three more or less decisive problems as a solution to the problem of moral compliance. Before moving on, it's worth briefly addressing the following concern. Given the theoretical and empirical reasons that favor the punishment-based account, one might naturally wonder why so many theorists have thought that reciprocity plays the primary role in supporting moral systems. I believe one reason for the popularity of the reciprocity-based account is some theorists have failed to distinguish the role of reciprocity in *sustaining* moral systems from *other roles* that reciprocity might play in relation to morality. One such role is that reciprocity is likely to have been an *evolutionary precursor* for human moral systems. A nice feature of reciprocity-based models, especially simple two-person models of reciprocity, is that there are well-developed evolutionary scenarios that show how systems of reciprocity might have first emerged from a state in which reciprocity was absent in a population. For example, Axelrod and Hamilton (1981) showed that even very low levels of assortive interactions (i.e., interactions in which reciprocators disproportionately interact with other reciprocators) are sufficient to allow rare reciprocators to invade a population of defectors. A natural hypothesis is that these simple systems of reciprocity might serve as evolutionary precursors for more complex elements of human moral systems, and, indeed, a number of theorists have made claims along these lines (for example, see De Waal 1996).

But we need to be careful in how we interpret these claims. Even if systems of reciprocity serve as evolutionary precursors for human moral systems, it does not follow that moral systems are *maintained* by systems of reciprocity. The phylogenetic history of human moral systems needs to be kept separate from the question of how moral systems are currently supported. As I've argued, there are strong reasons to believe that human moral systems are maintained by systems of punishment (and not reciprocity), and this claim is quite consistent with the hypothesis that moral systems have their phylogenetic roots in reciprocity.

## Part IV: The problem of sustaining costly punishment

According to the punishment-based account, moral norms are supported by means of punishment directed at those who violate moral norms. However, punishment, as opposed to non-reciprocation, is invariably costly to the punisher. What makes it the case that punishment is individually rational to carry out, even if the act of punishment is costly? Put another way, punishment curtails the incentive to violate moral norms, but what curtails the incentive to be a non-punisher?

It is only recently, as punishment-based models have become better known as alternatives to standard reciprocity-based models, that the question of punishment stabilization has garnered significant attention. Thus an important benefit from clarifying the difference between reciprocity and punishment-based models is that to the extent that we find that punishment-based models are overall more plausible as accounts of how moral norms are sustained in human groups, our attention is appropriately directed towards the heretofore-neglected issue of punishment stabilization. Though punishment stabilization is relatively new as a topic of investigation, important things have already been learned. There are a number of *intrinsic features* of punishment that lower its costs dramatically, as well as a number of *punishment stabilization mechanisms* that make it the case that self-interested agents will in fact carry out punishment.

One feature of punishment that lowers its cost is that for many kinds of punishment, there exists a striking *asymmetry* in which the costs to the dispenser of the punishment are a small fraction of the harm delivered to the receiver of the punishment (Sober and Wilson 1998; Bingham 1999). For example, for the cost of ostracizing one person, a man's social life and livelihood can be wrecked. Or, for the cost of starting a fire, a man's house can be destroyed (of course, there are likely to be at least some further costs that arise from the risk of retaliation). Because of the asymmetry between costs and harms associated with many kinds of punishment, for any given incentive to free ride, the cost of punishment needed to negate this incentive needs to be only a small fraction of the size of the incentive.

A second feature of punishment that lowers its cost is that punishment is a *conditional strategy* – one actually punishes (and pays the costs) only when someone violates moral norms and not otherwise. Consider a group in which most everyone cooperates and punishes non-cooperators. In this group, non-cooperators fare poorly and their numbers are quickly depleted (recall from section 'Punishment as a solution to the problem of cooperation' that punishment is a *selective* deterrent that can target non-cooperators specifically). With few non-cooperators, there are few acts of defection and punishment only rarely needs to be actually executed. As a result, the cost of sustaining punishment is dramatically lowered.[9]

In addition to the preceding intrinsic features of punishment that lower its cost, there are various *mechanisms* that serve to stabilize punishment. One of the most important mechanisms is *higher-order punishment*, in which those that fail to punish rule violations are themselves subject to punishment. This is the punishment stabilization mechanism invoked in the Boyd and Richerson (1992) model discussed earlier. The crucial idea in higher-order punishment is that people *treat the execution of punishment as a moral duty*. In other words, in addition to any moral rules they may hold, people view punishment for the violation of a moral rule as itself morally required. When punishment is conceptualized as a moral duty in this way, the result is the recursive generation of higher-order punishments.[10]For example, if a moral violation occurs, people in the community

[9] This argument appears in Olson's (1965) classic *The Logic of Collective Action*, is formalized in Oliver (1980) and is the basis for an evolutionary game-theoretic simulation in Boyd et al. (2003).

[10] Put another way, in my view, human moral psychology embodies the following recursive schema for generating higher-order moral requirements: *If X is morally required, punishing violations of X is morally required.* After writing this section, I was pleased to find that Alan Gibbard made a very similar point more than a decade ago in Gibbard (1990).

punish the wrongdoer, which they regard as their moral duty. But if a person does not punish the wrongdoer, that person has violated his moral duty to punish and is himself labeled as a wrongdoer, and is susceptible to punishment.

Does higher-order punishment require *infinite hierarchies* of punishment to stabilize punishment? Unfortunately, the answer to this question is not at all clear. Suppose people treat punishment as a moral duty, and as a consequence, they attempt to ensure that those that fail to punish are subject to punishment. Of course, because of cognitive limitations and other factors, individuals deploying this strategy will actually implement perhaps one or two layers of higher-order punishment, with the implementation of all higher orders of punishment being much less likely to occur. Even if higher-order punishment exhibits this pattern, that is, it is reliable at lower levels of the punishment hierarchy but much more inconsistent thereafter, it may nevertheless be evolutionarily stable.

Two factors tend to work to stabilize higher-order punishment. One is the asymmetry of costs and harms associated with punishment, which was discussed earlier. This tends to stabilize higher-order punishment by making the benefits from failing to punish significantly less than the harms associated with being punished for failing to punish, even if these harms are inconsistently delivered. The second factor is that higher-order non-punishers only very rarely get the opportunity to benefit from avoiding paying the costs of punishment. In general, in order for a person to benefit from failing to execute an $n$th-order punishment, a rule violation must occur *and* at least one person must fail to appropriately execute punishment at each level up to level $n - 1$. But since rule violation and non-punishment are each relatively rare events, the occurrence of opportunities for higher-order non-punishers to benefit from non-punishment are even rarer still.

It may be the case that despite the operation of the preceding factors, there is some level on the punishment hierarchy at which agents reliably cease to have an incentive to carry out costly punishment. In classical game theory, an equilibrium is said to fail the test of *subgame perfection* if it is maintained by threats that an agent does not have an incentive to actually carry out. In an evolutionary game, however, evolutionary dynamics can, in some circumstances, sustain equilibriae that are not subgame perfect (see Samuelson 1997, esp. Ch. 8 and Skyrms 1990). Much further work is needed to ascertain whether higher-order punishment, implemented in a manner consistent with human cognitive limitations, is subgame perfect, and if it is not, whether it can nevertheless be sustained by evolutionary dynamics.

Since the evolutionary dynamics of higher-order punishment are still poorly understood, empirical evidence for the real-world existence and operation of higher-order punishment would be quite useful. Though it has never been directly studied, *indirect evidence* for the existence of higher-order punishment comes from a large empirical literature that finds that people conceptualize punishment as a moral duty.[11] For example, in a review of the literature in sociology and criminology, Vidmar and Miller (1980) conclude that people have powerful *retributivist sentiments* – they see punishment as morally required because the violator deserves to suffer.

> Although some of these retributive reactions may derive directly from perceiving a threat to group or individual values, they also arise from deeply held beliefs of 'justice' or 'oughtness': the reactor feels it would not be right for the offender to escape with impunity. The offender has violated a moral rule that transcends the specific victim or even the social group… The affective reaction in these instances is strong: a compelling need to see the moral order set right has been aroused (Vidmar and Miller 1980, p. 580–581).

In addition to higher-order punishment, theorists have proposed various other mechanisms that might play a role in punishment stabilization. One proposal is based on the theory of *costly signaling*. According to this proposal, individuals who appropriately punish defectors signal that they are good cooperative partners (or good in other respects), while individuals who fail to appropriately punish reveal that they are not good cooperative partners (or not good in other respects) (Gintis et al. 2001). Henrich and Boyd (2001) have proposed a punishment stabilization mechanism that is based on *conformist cultural transmission*, the tendency for people to adopt cultural variants based on the variant's commonness in the population. Recently, Boyd et al. (2003) have shown that low levels of *cultural group selection* can serve to stabilize punishment. Overall, the issue of punishment stabilization is of central importance in understanding the strategic structure of moral systems. Much has already been learned, and the issue deserves continued study.

## Part V: Towards an evolutionary account of moral diversity

The *problem of moral compliance* is the problem of explaining how moral norms are sustained despite the existence of selfish evolutionary incentives that favor their violation. In this paper, I argued that the prevailing view among evolutionary-minded theorists is false; the problem of moral compliance is

[11] Philosophers have long produced normative theories that emphasize that punishment is a moral duty. For example, in a memorable passage from *Critique of Practical Reason*, Kant writes:Even if a civil society resolved to dissolve itself with the consent of all its members- as might be supposed in the case of a people inhabiting an island resolving to separate and scatter themselves throughout the world- the last murderer lying in prison ought to be executed before the resolution was carried out. This ought to be done in order that every one may realize the desert of his deeds, and the bloodguiltiness may not remain on the people; for otherwise they might all be regarded as participators in the murder as a public violation of justice (Kant 1972[1887], p. 105–106).Kant's claim that a society that fails to punish a murderer is itself guilty of a crime makes clear the connection between viewing punishment as a moral duty and higher-order punishment.

solved by punishment, not reciprocity. In this concluding section, I'll argue that in addition to providing a solution to the problem of moral compliance, the punishment-based account provides another important theoretical dividend. It points the way for how theorists might eventually build an evolutionary account of a feature of human groups that has long fascinated and troubled social scientists and moral philosophers – the existence of *moral diversity*.

The topic of moral diversity is conspicuously missing from the literature on the evolution of morality. I believe the excessive focus on reciprocity-based models has been the primary reason for this pattern of neglect. As we've seen, the main problem with reciprocity is that it is fundamentally a theory of how cooperation can be sustained in collective action problems. As a consequence, reciprocity-based models of morality create the impression that human moral systems prescribe a highly *uniform* set of rules across human groups, where these rules invariably lead to cooperative or *group beneficial* outcomes.

However, the picture of moral systems associated with the reciprocity-based account is in fact deeply misleading in two ways. First, the contents of moral norms are not uniform across human groups and, second, moral norms do not always lead to group beneficial outcomes. Let me take up each of these points in turn. A closer look at the ethnographic record suggests that moral norms are in fact quite *variable* across human groups. There is solid evidence that in domains such as social exchange (Fiske 1991; Henrich et al. 2001), violence (Robarcheck and Robarcheck 1992; Keeley 1996), hierarchy and social stratification (Boehm 1999), marriage (Durham 1991), sexual rules (Bourguignon and Greenbaum 1973) and many others, moral norms differ substantially across human groups (see Sripada and Stich forthcoming for a brief review of some evidence for moral diversity). Because moral diversity of this sort is difficult to accommodate in the reciprocity-based framework, it has been largely ignored by evolutionary-minded theorists. It is unfortunate, to say the least, that one of the most fascinating and important aspects of human morality is simply not discussed in the evolution of morality literature.

Another misleading feature of the reciprocity-based account is that it fosters a kind of *naïve optimism* about the social consequences of moral systems. The reciprocity-based account encourages us to believe moral systems invariably produce outcomes that confer a *mutual benefit* on interacting parties. But the ethnographic record calls this Panglossian picture into question. Actual moral norms that prevail in human groups – for example, asymmetric gender norms, grossly hierarchical rules of authority, burdensome food or sexual taboos – are often sub-optimal in the sense that they diverge from what one might call a 'cooperative' or 'mutually beneficial' outcome (see Edgerton 1992). Thus, far from being uniformly *group beneficial*, moral norms differ substantially across human groups in the extent to which they lead to prosocial outcomes. This is another important aspect of moral diversity, and it too has been largely ignored by evolutionary-minded theorists.

The origins and maintenance of moral diversity can, in contrast, be quite naturally investigated from the perspective of the punishment-based account. As we've seen, punishment is a quite general method for sustaining compliance with *just about any moral rule*. For this reason, the punishment-based account provides a natural way of modeling how different rules might emerge and remain stable in different groups. Additionally, the flexibility of punishment as a rule-stabilizer provides a natural way of modeling the emergence and stability of sub-optimal moral norms, including asymmetric or burdensome rules of the kind routinely found in human societies. Thus, in contrast to the reciprocity-based account, which has difficulty accommodating moral diversity and thus encourages its neglect, the punishment-based account provides researchers with a powerful framework for constructing sophisticated evolutionary models of moral diversity. More generally, the punishment-based account calls into question the prevailing optimism about the nature of moral systems fostered by the reciprocity-based account. When moral systems are investigated from the perspective of the punishment-based account, they are seen to exhibit a fundamental kind of *arbitrariness* that demands careful further exploration.

## References

Alexander R.D. 1987. The Biology of Moral Systems. Aldine De Gruyter, New York.

Axelrod R. 1984. The Evolution of Cooperation. Basic Books, New York.

Axelrod R. and Hamilton W.D. 1981. The evolution of cooperation. Science 211: 1390–1396.

Bingham P. 1999. Human uniqueness: a general theory. Quart. Rev. Biol. 74(2): 133–169.

Black D. 1998. The Social Structure of Right and Wrong. Academic Press, London.

Boehm C. 1999. Hierarchy in the Forest. Harvard University Press, Cambridge, MA.

Bourguignon E. and Greenbaum L. 1973. Diversity and Homogeneity in World Societies. HRAF Press, New Haven, CT.

Boyd R. 1989. Mistakes allow evolutionary stability in the repeated Prisoner's Dilemma. J. Theor. Biol. 136: 47–56.

Boyd R., Bowles S., Gintis H. and Richerson P. 2003. The evolution of altruistic punishment. Proc. Natl. Acad. Sci. 100(6): 3531–3535.

Boyd R. and Richerson P. 1989. The evolution of reciprocity in sizable groups. J. Theor. Biol. 132: 337–356.

Boyd R. and Richerson P. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol. Sociobiol. 13: 171–195.

Brown D. 1991. Human Universals. McGraw-Hill, New York.

Carpenter J., Mathews P. and Okomboli O.Why punish? Social reciprocity and the enforcement of prosocial norms. J. Evol. Econ. 14(4): 407–429.

Cashdan E. 1980. Egalitarianism among hunters and gatherers. Am. Anthropol. 82: 116–120.

De Waal F. 1996. Good Natured: The Origin of Right and Wrong in Humans and Other Animals. Harvard University Press, Cambridge, MA.

Dunbar R. 1997. Grooming, Gossip and the Evolution of Language. Harvard University Press, Cambridge, MA.

Durham W. 1991. Coevolution. Stanford University Press, Stanford, CA.

Edel M. and Edel A. 2000. Anthropology and Ethics. Transaction Publishers, New Brunswick, NJ.

Edgerton R.B. 1992. Sick Societies. The Free Press, New York.

Fehr E. and Gachter S. 2002. Altruistic punishment in humans. Nature 415: 137–140.

Fiske A.P. 1991. Structures of Social Life. The Free Press, New York.

Fudenberg D. and Maskin E. 1986. The folk theorem in repeated games with discounting and incomplete information. Econometrica 54: 533–554.

Gibbard A. 1990. Norms, discussion, and ritual: evolutionary puzzles. Ethics 100(4): 787–802.

Gintis H., Smith E.A. and Bowles S. 2001. Costly signaling and cooperation. J. Theor. Biol. 213: 103–119.

Haidt J. 2000. The moral emotions. In: Davidson R.J., Scherer K. and Goldsmith H.H. (eds), Handbook of Affective Sciences. Oxford University Press, New York.

Henrich J. and Boyd R. 2001. Why people punish defectors. J. Theor. Biology 208: 79–89.

Henrich J., Boyd R., Bowles S., Camerer C., Fehr E. and Gintis H. 2001. Foundations of Human Sociality. Oxford University Press, New York.

Horne C. 2001. The enforcement of norms: group cohesion and meta-norms. Soc. Psychol. Quart. 64(3): 253–266.

Hume D. 1992 [1739]. A Treatise of Human Nature. Prometheus Books, Buffalo, New York.

Kant I. 1972 [1887]. Justice and punishment (from *Critique of Practical Reason*). In: Ezorsky G. (ed.), Philosophical Perspectives on Punishment. State University of New York Press, Albany, NY.

Keeley L. 1996. War Before Civilization: The Myth of the Peaceful Savage. Oxford University Press, New York.

Maynard S.J. and Price G.R. 1973. The logic of animal conflict. Nature 246: 15–18.

Murdock G.P. 1949. Social Structure. Free Press, New York.

Oliver P. 1980. Rewards and punishment as selective incentives for collective action: theoretical investigations. Am. J. Sociol. 85(6): 1356–1375.

Olson M. 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Ostrom E. 1990. Governing the Commons. Cambridge University Press, New York.

Robarcheck C.A. and Robarchek C.J. 1992. Cultures of war and peace: a comparative study of Waorani and Semai. In: Silverberg J. and Gray P. (eds), Aggression and Peacefulness in Humans and Other Primates. Oxford University Press, New York.

Roberts S. 1979. Order and Dispute. St. Martin's Press, New York.

Samuelson L. 1997. Evolutionary Games and Equilibrium Selection. MIT Press, Cambridge, MA.

Skyrms B. 1990. The Dynamics of Rational Deliberation. Harvard University Press, Cambridge, MA.

Sober E. and Wilson D.S. 1998. Unto Others. Harvard University Press, Cambridge, MA.

Sripada C.S., Stich S. A framework for the psychology of norms. To appear In: Carruthers S., Laurence S. and Stich S. (eds), Innateness and the Structure of the Mind, Vol. 2. Oxford University Press, London.

Trivers R. 1971. The evolution of reciprocal altruism. Quart. J. Biol. 46(1): 35–57.

Uyenoyama M.R. and Feldman M.W. 1992. Altruism: some theoretical ambiguities. In: Keller E.F. and Lloyd E.A. (eds), Keywords in Evolutionary Biology. Harvard University Press, Cambridge, MA.

Vidmar N. and Miller D.T. 1980. Sociological processes underlying attitudes toward legal punishment. Law Soc. Rev. 14(3): 565–602.

Westermark E. 1937. Ethical Relativity. Brace and Company, Harcourt, New York.

Wilson D.S., Wilczynski C., Wells A. and Weiser L. 2000. Gossip and other aspects of language as group-level adaptations. In: Heyes C. and Huber L. (eds), The Evolution of Cognition. MIT Press, Cambridge, MA.