J. theor. Biol. (2001) **208,** 79–89

Why People Punish Defectors

Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas

JOSEPH HENRICH* AND ROBERT BOYD†

**University of Michigan,* 701 *Tappan Road, D3276, Ann Arbor, MI* 48109-1234, *U.S.A. and*
†Department of Anthropology, University of California, Los Angeles, 405 *Hilgard Ave, Los Angeles, CA* 90024, *U.S.A.*

(*Received on* 8 *June* 2000, *Accepted in revised form on* 22 *September* 2000)

In this paper, we present a cultural evolutionary model in which norms for cooperation and punishment are acquired via two cognitive mechanisms: (1) payoff-biased transmission—a tendency to copy the most successful individual; and (2) conformist transmission—a tendency to copy the most frequent behavior in the population. We first show that if a finite number of punishment stages is permitted (e.g. two stages of punishment occur if some individuals punish people who fail to punish non-cooperators), then an arbitrarily small amount of conformist transmission will stabilize cooperative behavior by stabilizing punishment at some *n*-th stage. We then explain how, once cooperation is stabilized in one group, it may spread through a multi-group population via cultural group selection. Finally, once cooperation is prevalent, we show how prosocial genes favoring cooperation and punishment may invade in the wake of cultural group selection.

Introduction

In many societies, humans cooperate in large groups of unrelated individuals. Most evolutionary explanations for cooperation combine kinship (Hamilton, 1964) and reciprocity ("reciprocal altruism" Trivers, 1971). These mechanisms seem to explain the evolution of cooperation in many species including ants, bees, naked mole rats and vampire bats. However, because social interaction among humans often involves large groups of mostly unrelated individuals, explaining cooperation has proved a tricky problem for both evolutionary and rational choice theorists. Evolutionary models of cooperation using the repeated *n*-person prisoner's dilemma predict that cooperation is not likely to be favored by natural selection if groups are larger than around 10, unless relatedness is very high (Boyd & Richerson, 1988). As group size rises above 10, to 100 or 1000, cooperation is virtually impossible to evolve or maintain with only reciprocity and kinship.*

E-mails: henrich@umich.edu; rboyd@anthro.ucla.edu

*Two other explanations for cooperation go by the handles *by-product mutualism* (Brown, 1983) and *group selection* (Sober & Wilson, 1998). In by-production mutualism, individuals who "cooperate" get a higher payoff (have a higher expected fitness) than non-cooperators. The cooperative contribution to the fitness of others is simply a by-product of narrow self-interest. That is, in the process of helping myself, I also help you "by accident". Hence, although this situation may abound in nature, it is not the situation we are interested in (and not cooperation by many definitions). And, while genetic group selection may explain some cooperation in nature (e.g. honeybees, see Seeley, 1995), we believe that gene flow rates between human populations, relative to selection, are too high to maintain the required variation between groups (Richerson & Boyd, 1998).

Many students of human behavior believe that large-scale human cooperation is maintained by the threat of punishment. From this view, cooperation persists because the penalties for failing to cooperate are sufficiently large that defection "doesn't pay". However, explaining cooperation in this way leads to a new problem: why do people punish non-cooperators? If the private benefits derived from punishing are greater than the costs of administering it, punishment may initially increase, but cannot exceed a modest frequency (Boyd & Richerson, 1992). Individuals who punish defectors provide a public good, and thus can be exploited by non-punishing cooperators if punishment is costly. Second-order free riders cooperate in the main activity, but cheat when it comes time to punish non-cooperators. As a consequence, second-order free riders receive higher payoffs than punishers do, and thus punishment is not evolutionarily stable. Adding third (third-order punishers punish second-order free riders) or higher-order punishers only pushes the problem back to higher orders. Solving this problem is important because there is widespread agreement that the threat of punishment plays an important role in the maintenance of cooperation in many human societies.

Social scientists have explained the maintenance of punishment in three ways: (1) many authors assume that a state or some other external institution does the punishing; (2) others assume punishing is costless (McAdams, 1997; Hirshleifer & Rasmussen, 1989); and (3) a few scholars incorporate a recursive punishing method in which punishers punish defectors, individuals who fail to punish defectors, individuals who fail to punish non-punishers, and so on in an infinite regress (Boyd & Richerson, 1992; Fundenberg & Maskin, 1986). However, none of these solutions are satisfactory. While it is useful to assume institutional enforcement in modern contexts, it leaves the evolution and maintenance of punishment unexplained because at some point in the past there were no states or institutions. Furthermore, the state plays a very small role in many contemporary small-scale societies that nonetheless exhibit a great deal of cooperative behavior. This solution avoids the problem of punishment by relocating the costs of punishment outside the problem. The second solution, instead of relocating the costs, assumes that punishment is costless. This seems unrealistic because any attempt to inflict costs on another must be accompanied by at least some tiny cost—and any non-zero cost lands both genetic evolutionary and rational choice approaches back on the horns of the original punishment dilemma. The third solution, pushing the cost of punishment out to infinity, also seems unrealistic. Do people really punish people who fail to punish other non-punishers, and do people punish people who fail to punish people, who fail to punish non-punishers of defectors and so on, *ad infinitum*? Although the infinite recursion is cogent, it seems like a mathematical trick.

Conformist Transmission in Social Learning can Stabilize Punishment

In this paper, we argue that the evolution of cooperation and punishment are plausibly a side effect of a tendency to adopt common behaviors during enculturation. Humans are unique among primates in that they acquire *much* of their behavior from other humans via social learning. However, both theory and evidence suggest that humans do not simply copy their parents, nor do they copy other individuals at random (Henrich & Boyd, 1998; Takahasi, 1998; Harris, 1998). Instead, people seem to use social learning rules like "copy the successful" (termed pay-off biased or prestige-biased transmission, see Henrich & Gil-White, 2000) and "copy the majority" (termed conformist transmission, Boyd & Richerson, 1985; Henrich & Boyd, 1998), which allow them to short-cut the costs of individual learning and experimentation, and leapfrog directly to adaptive behaviors. These specialized social learning mechanisms provide a generalized means of rapidly sifting through the wash of information available in the social world and inexpensively extracting adaptive behaviors. These social learning short-cuts do not always result in the best behaviors, nor do they prevent the acquisition of maladaptive behaviors. Nevertheless, when averaged over many environments and behavioral domains (e.g. foraging, hunting, social interaction, etc.), these cultural transmission mechanisms provide fast and frugal means