

## 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.

© 2001 Academic Press

#### 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.\*

\*Two other explanations for cooperation go by the handles *by-product mutualism* (Brown, 1983) and *group selection* (Sober & Wilson, 1998). In by-product 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 & Rasmusen, 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; Fudenberg & 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

to acquire complex, highly adaptive behavioral repertoires.

Both theoretical and empirical research indicates that conformist transmission plays an important role in human social learning. We have already shown that a heavy reliance on conformist transmission outcompetes both unbiased (i.e. vertical) transmission and individual learning under a wide range of conditions (Henrich & Boyd, 1998), and especially when problems are difficult. Second, empirical research by psychologists, economists and sociologists shows that people are likely to adopt common behaviors across a wide range of decision domains. Although much of this work focuses on easy perceptual tasks (Asch, 1951), and confounds normative conformity (going with the popular choice to avoid appearing deviant) with conformist transmission (using the popularity of a choice as an indirect measure of its worth), more recent work shows that social learning and conformist transmission are important in difficult individual problems (Baron *et al.*, 1996; Insko *et al.*, 1985; Campbell & Fairey, 1989), voting situations (Wit, 1999) and cooperative dilemmas (Smith & Bell, 1994).

Conformist transmission can stabilize costly cooperation without punishment, but only if it is very strong. All other things being equal, pay-off biased transmission causes higher payoff variants to increase in frequency, and thus cooperation is not evolutionarily stable under plausible conditions—because not-cooperating leads to higher payoffs than cooperating. Thus, pay-off biased transmission, alone, suffers the same problem as natural selection in genetic evolution. However, under conformist transmission individuals preferentially adopt common behaviors, which acts to increase the frequency of the most common behavior in the population. Thus, if cooperation is common, conformist transmission will oppose payoff-biased transmission, and, as long as cooperation is not too costly, maintain cooperative strategies in the population. However, if the costs of cooperation are substantial, it is less likely that conformist transmission will be able to maintain cooperation.

A quite different logic applies to the maintenance of punishment. Suppose that both punishers and cooperators are common, and that being

punished is sufficiently costly that cooperators have higher payoffs than defectors. Rare invading second-order free riders who cooperate but do not punish will achieve higher payoffs than punishers because they avoid the costs of punishing. However, because defection does not pay, the only defections will be due to rare mistakes, and thus the *difference* between the payoffs of punishers and second-order free riders will be relatively small. Hence, conformist transmission is more likely to stabilize the punishment of non-cooperators than cooperation itself. As we ascend to higher-order punishing, the difference between the payoffs to punishing vs. non-punishing decreases geometrically towards zero because the occasions that require the administration of punishment become increasingly rare. Second-order punishing is required only if someone erroneously fails to cooperate, and then someone else erroneously fails to punish that mistake. For third-order punishment to be necessary, yet another failure to punish must occur. As the number of punishing stages ( $i$ ) increases, conformist transmission, no matter how weak, will at some stage overpower payoff-biased imitation and stabilize common  $i$ -th order punishment. Once punishment is stable at the  $i$ -th stage, payoffs will favor strategies that punish at the  $(i - 1)$ -th order, because common punishers at the  $i$ -th order will punish non-punishers at stage  $i - 1$ . Stable punishment at stage  $(i - 1)$ -th order means payoffs at stage  $i - 2$  will favor punishing strategies, and so on down the cascade of punishment. Eventually, common first-order punishers will stabilize cooperation at stage 0.

It is important to see that the stabilization of punishment is, from the gene's point of view, a maladaptive side-effect of conformist transmission. If there were genetic variability in the strength of conformist transmission ( $\alpha$ ) and cooperative dilemmas were the *only* problem humans faced, then conformist transmission might never evolve. However, human social learning mechanisms were selected for their capability to efficiently acquire adaptive behaviors over a wide range of behavioral domains and environmental circumstances—from figuring out what foods to eat, to deciding what kind of person to marry—precisely because it is costly for individuals to determine the best behavior. Hence, we should

expect conformist transmission to be important in cooperation as long as distinguishing cooperative dilemmas from other kinds of problems is difficult, costly or error prone. Looking across human societies we find that cooperative dilemmas come in an immense variety of forms, including harvest rituals among agriculturalists, barbasco fishing among Amazonian peoples, warfare, irrigation projects, taxes, voting, meat sharing and anti-smoking pressure in public places. It is difficult to imagine a cognitive mechanism capable of distinguishing cooperative circumstances from the myriad of other problems and social interactions that people encounter.

In what is to come, we formalize this argument. Our goal is to demonstrate the soundness of our reasoning and show how very weak conformist transmission can stabilize cooperation and punishment. After demonstrating this, we will describe how cooperation, once it is stabilized in one group, can spread across many populations via *cultural group selection*. We will also briefly show how genes for prosocial behavior may eventually spread in the wake of cultural evolution.

### A Cultural Evolutionary Model of Cooperation and Punishment

In this model, a large number of groups each consisting of  $N$  individuals are drawn at random from a very large population. Individuals within each group interact with one another in an  $i + 1$  stage game. The first stage is a one-shot cooperative dilemma, which is followed by  $i$  stages in which individuals can punish others. We number the first, cooperative stage as “0” and the punishment stages as  $1, \dots, i$ . The behavior of individuals during each stage is determined by a separate culturally acquired trait with two variants, P (prosocial variant) and NP (not prosocial variant).

During the initial cooperative dilemma, individuals can either “cooperate”—contribute to a public good—or “defect”—not contribute and free-ride on the contributions of others. Each cooperator pays a cost  $C$  to contribute a benefit  $B$  ( $B > C$ ) to the group—this  $B$  is divided equally among *all* group members. Defectors do not pay the cost of cooperation ( $C$ ), but do share equally

in the total benefits. The variable  $p_0$  represents the frequency of individuals in the population with the cooperative variant in stage 0. People with the cooperative variant “intend” to cooperate, but mistakenly defect with probability  $e$ . Individuals who have the defecting variant always defect. This makes sense because, in the real world, people may intend to cooperate, but fail to for some reason. For example, a friend who plans to help you move, may forget to show up or have car trouble en route, etc. Defectors, however, are unlikely to mistakenly show-up on moving day and start carrying boxes. We will assume errors are rare, so that the value of  $e$  is small.

During the first punishment stage, individuals can punish those who defected during the cooperation stage. Doing this reduces the payoff of the individuals who are punished by an amount  $\rho$ , at a cost of  $\phi$  to the punisher ( $\phi < \rho < C$ ). Individuals with the punishing (P) variant at this stage intend to punish, but mistakenly fail to punish with probability  $e$ . Non-punishers, those with the NP-variant at stage 1, do nothing. We use  $p_1$  to stand for the frequency of first-stage punishers (i.e. individuals who have the P-variant at stage 1), and  $(1 - p_1)$  gives the frequency of first-stage free riders.

During the second punishment stage, individuals with the P-variant punish those who did not punish the non-cooperators during the previous stage with probability  $(1 - e)$ , and mistakenly fail to punish with probability  $e$ . And as before, punishment costs punishers  $\phi$  to administer, and costs those being punished an amount  $\rho$ . Those with the NP-variant at stage 2 do not punish. Let  $p_2$  be the frequency of second-stage punishers. At stage 3, individuals with the P-variant will punish individuals from stage 2 who failed to punish non-punishers from stage 1. The costs of punishment remain the same. Those with the NP-variant in stage 3 will not punish anyone from stage 2. The pattern repeats as one descends to stage  $i$  in Table 1 ( $p_i$  gives the frequency of punishers at stage  $i$ ). Because the interaction ends after stage  $i$ , individuals who fail to punish on stage  $i$  cannot be punished. Note that the trait that controls individual behavior at each stage has only two variants, and the values of variants at different stages are independent—so an individual could cooperate at stage 0 (have the

TABLE 1  
*Dichotomous traits for cooperation and punishment*

Stage	Frequency of P-variant	P-variant	NP-variant
0	$p_0$	Cooperate	Defect
1	$p_1$	Punish defectors	Do not punish defectors
2	$p_2$	Punish non-punishers at stage 1	Do not punish non-punishers at stage 1
3	$p_3$	Punish non-punishers at stage 2	Do not punish non-punishers at stage 2
$i$	$p_i$	Punish non-punishers at stage $i - 1$	Do not punish non-punishers at stage $i - 1$

P-variant), not punish at stage 1 (NP-variant), and punish at stage 2 (P-variant).

After all the punishments are complete, cultural transmission takes place. As we explained earlier, two components of human cognition create forces that change the frequency of the different variants: pay-off biased and conformist-biased imitation. Equation (1) gives the change in the frequency of stage 1 cooperators as a consequence of pay-off biased and conformist transmission (see Henrich, 1999).

$$\Delta p_0 =$$

$$p_0(1 - p_0) \left[ \underbrace{(1 - \alpha)\beta(b_C - b_D)}_{\text{Payoff-biased}} + \underbrace{\alpha(2p_0 - 1)}_{\text{Conformist}} \right]. \tag{1}$$

The parameter  $\alpha$  varies from 0 to 1 and represents the strength of conformist transmission in human psychology *relative* to pay-off biased transmission. We will generally assume  $\alpha$  is positive, but small. Practically speaking,  $\alpha$  must be less than 0.50, because otherwise beneficial variants would never spread—once a variant became common, it would remain common no matter how deleterious. The second term in eqn (1), labeled “conformist”, varies in magnitude from  $-\alpha$  to  $+\alpha$  and is the component of the overall bias contributed by conformist transmission. In the term labeled “payoff-biased”, the symbols  $b_C$  and  $b_D$  are the payoffs to cooperators and defectors, respectively. The quantity  $(b_C - b_D)$ , which we label  $\Delta b_0$ , gives the difference in payoffs between cooperation (P-variant) and defection (NP-variant) in stage 0. More generally,  $\Delta b_i$  is the differences in payoffs between the P- and NP-variants

during the  $i$ -th stage. The parameter  $\beta$  normalizes the quantity  $\Delta b_i$  so that it varies between  $-1$  and  $+1$ , and therefore  $\beta = 1/|\Delta b_i|_{max}$ . Thus, the term labeled “payoff-biased” varies between  $-(1 - \alpha)$  and  $+(1 - \alpha)$  and represents the component of the overall bias contributed by payoff-biased transmission.

The expected payoffs,  $b$ , to the P- and NP-variant at each stage depend on the rate of errors, the costs of cooperation and/or punishment, and the frequency of cooperators and punishers in the population. At stage 0, cooperators receive an average payoff of  $b_C$ , while defectors receive an average payoff of  $b_D$ :

$$b_C = (1 - e)(p_0 B(1 - e) - C + e(p_0 B - Np_1 \rho)),$$

$$b_D = (1 - e)(p_0 B - Np_1 \rho), \tag{2}$$

$$\Delta b_0 = b_C - b_D = (1 - e)(Np_1(1 - e)\rho - C).$$

Also as we mentioned, the term  $\Delta b_0$  gives the difference in payoffs between the two variants that control stage 0 behavior.

#### A HEURISTIC ANALYSIS

Let us first analyse eqn (1) by asking under what conditions will transmission favor cooperation ( $\Delta p_0 > 0$ ) in the absence of stage 1 punishers ( $p_1 = 0$ ). In this case,  $\Delta b_0 = -C(1 - e)$ , which is always negative; hence, payoff biased transmission never favors cooperation in the absence of punishment. So, to give cooperation its best chance, we assume that by some stochastic fluctuations the frequency of cooperators ends up

near one. How big does  $\alpha$  have to be so that conformist transmission overpowers payoff-biased transmission and increases the frequency of cooperators? The frequency of cooperators increases when

$$\alpha_0 > \frac{1}{1 + \beta_0 C(1 - e)}, \quad (3)$$

where  $\alpha_i$  (here,  $i = 0$ ) is the minimum value of  $\alpha$  that favors the spread or maintenance of the P-variant at stage  $i$  ( $\Delta p_i > 0$ ). With no punishment,  $\beta_i = 1/|\Delta b_i|_{max}$  means  $\beta_0 = 1/(C(1 - e))$ . As a consequence,  $\alpha_0$  must be greater than 0.50, and as we mentioned earlier,  $\alpha_i > 0.50$  seems extremely unlikely because such high values would prevent the diffusion of novel practices—cultures would be entirely static (see Henrich, 1999). Hence, conformist transmission, operating directly on cooperative strategies, is unlikely to maintain cooperation in the absence of punishment.

Now, let us examine the conditions under which first-stage punishment will increase in frequency. Again, the change in the frequency of first-stage punishers,  $\Delta p_1$ , is affected by both payoff biased and conformist transmission:

$$\begin{aligned} \Delta p_1 = & p_1(1 - p_1)[(1 - \alpha)\beta(b_{P1} - b_{NP1}) \\ & + \alpha(2p_1 - 1)]. \end{aligned} \quad (4)$$

The payoffs ( $b$ 's) to punishment and non-punishment depend on the cost of punishing ( $\phi$ ) and of being punished ( $\rho$ ), as well as the chance of mistakenly not punishing ( $e$ ). The subscript P1 indicates the P-variant at stage 1, while NP1 indicates the NP-variant at stage 1.

$$\begin{aligned} b_{P1} = & -(1 - e)N\phi(1 - p_0 + p_0e) \\ & - eNp_2\rho(1 - e), \\ b_{NP1} = & -Np_2(1 - e)\rho, \end{aligned} \quad (5)$$

$$\begin{aligned} \Delta b_1 = & b_{P1} - b_{NP1} = -N(1 - e) \\ & \times (\phi(1 - (1 - e)p_0) - p_2(1 - e)\rho). \end{aligned}$$

Assuming that there is only one punishment stage ( $i = 1$ ), and that cooperators and stage 1 punishers are initially common ( $p_0 = 1$  and  $p_1 = 1$ ), then  $\Delta b_1 = -N(1 - e)e\phi$ . If errors are rare enough such that terms involving  $e^2$  are negligible, then  $\Delta b_1 \approx -Ne\phi$ . Thus, the difference in payoff between the P-variant and the NP-variants at stage 1 is just the cost of punishing cooperators who make errors. If  $e < (1/N)$ , which is plausible unless groups are very large, then  $\Delta b_1$  is less than  $\phi$ —and smaller than  $\Delta b_0$  because  $\phi < \rho < C$ . Note that, when  $i > 0$ ,  $\beta = 1/(N(1 - e)(\rho(1 - e) + e\phi))$ , so the threshold value of  $\alpha$  necessary to stabilize cooperation in a two-stage game  $\alpha_1$ , is<sup>†</sup>

$$\alpha_1 = \frac{\phi e}{\rho(1 - e) + 2\phi e} \approx \frac{e\phi}{\rho}. \quad (6)$$

Equation (6) tells us that  $\alpha_1$  depends only on the error rate and the ratio of the cost of punishing to the cost of being punished. It also says that unless punishing is much more costly than being punished ( $2\phi e > \rho$ ), the threshold strength of conformism necessary to maintain first-stage punishment is small and less than the amount of conformism necessary to stabilize 0-th stage cooperation ( $\alpha_0 > \alpha_1 \approx e$ ).

If we do the same analysis for stage 2, we get the following expressions for  $\Delta p_2$  and  $\Delta b_2$ :

$$\Delta p_2 = p_2(1 - p_2)[(1 - \alpha)\beta\Delta b_2 + \alpha(2p_2 - 1)], \quad (7)$$

where

$$\begin{aligned} \Delta b_2 = & b_{P2} - b_{NP2} = -(1 - e)N[\phi(1 - p_1(1 - e)) \\ & \times (1 - p_0^N(1 - e)^N) - p_3(1 - e)\rho]. \end{aligned} \quad (8)$$

The first term inside the square brackets in eqn (8) is proportional to the number of individuals who did not punish during stage 1 ( $1 - p_1(1 - e)$ ), and to the probability that there was at least one defector during stage 0:  $(1 - p_0^N(1 - e)^N)$ . The quantity  $p_0(1 - e)$  is the

<sup>†</sup> Note, under a small range of conditions, when  $C > N(\rho(1 - e) + e\phi)$ , the system can still remain stable. Under these conditions, however,  $\beta$  becomes  $1/C(1 - e)$ . For simplicity, we leave this nuance until later in the paper.

expected frequency of cooperators who did not make a mistake, thus  $(p_0(1 - e))^N$  gives the probability that a group contains all cooperators who did not make a mistake—so, to get the probability that a group contains at least one defector, we simply subtract this probability from one. The second term inside the brackets is the cost of being punished during stage 2 for failing to punish during stage 1. If no third-stage punishers exist ( $p_3 = 0$ ), and first-stage punishers and cooperators are initially very common, then  $\Delta b_2 \approx -(eN)^2 \phi$ . Note, the difference in payoffs,  $\Delta b_2$ , is a factor of  $eN$  smaller than  $\Delta b_1$ , but the strength of conformist transmission remains constant. Calculating the required size of  $\alpha_2$  we get

$$\alpha_2 = \frac{N\phi e^2}{\rho(1 - e) + e\phi} \approx \frac{e\phi}{\rho} Ne. \quad (9)$$

Equation (9) demonstrates that  $0 < \alpha_2 < \alpha_1 < \alpha_0 = \frac{1}{2}$ . In this case  $\alpha_2 \approx Ne\alpha_1$ .

If we repeat this calculation for games with more punishment stages, we find that, although punishment during the last stage of the game is never favored by pay-off biased transmission *alone*, any positive amount of conformist transmission ( $\alpha > 0$ ) will, for some finite number of stages, overcome payoff-biased transmission and stabilize punishment. For any value  $i$  ( $i > 0$ ), the amount of conformist transmission required to stabilize punishment at the  $i$ -th stage is

$$\alpha_i = \frac{\phi e (Ne)^{i-1}}{\rho(1 - e) + e\phi(1 + (Ne)^{i-1})} \approx \frac{e\phi}{\rho} (Ne)^{i-1}. \quad (10)$$

Equation (10) shows that minimum amount of conformism necessary to stabilize punishment during the last stage,  $\alpha_i$ , gets smaller and smaller for greater values of  $i$  (assuming  $e < 1/N$ ).

Once conformist transmission overcomes payoff-biased transmission and stabilizes punishment at stage  $i$ , punishment at the stage  $i - 1$  will be stabilized because non-punishers at stage  $i - 1$  will be punished by frequent punishers during stage  $i$ . Once punishing strategies are common and stable at stage  $i - 1$ , frequent punishers at  $i - 1$  will cause pay-off biased transmission to favor the prosocial variant at stage  $i - 2$ . In most

cases, a combination of punishment and conformist transmission will eventually stabilize cooperation at stage 0. However, if  $C$  is sufficiently greater than  $N\rho(1 - e)$ , then stable punishment at stage 1 will not be able to overcome the costs of cooperation at stage 0, and cooperation will not be maintained, despite stable, high-frequency first-stage punishers.

#### FORMAL STABILITY ANALYSIS

A more rigorous local stability analysis of the complete set of recursions supports the heuristic argument just given. Consider the set of  $i + 1$  difference equations where  $\Delta p_j$  ( $j = 0, 1, \dots, i$ ; see the appendix) provides the dynamics of the behavioral traits at each stage. The cooperative equilibrium point ( $p_0 = 1, p_1 = 1, \dots, p_i = 1$ ) is locally stable under two distinct conditions:

**Stability Condition 1.** When  $i > 0$  and  $C < \rho(1 - e)N + (eN)^i\phi$  the cooperative equilibrium is locally stable when

$$\lambda_d = -\alpha + (1 - e)(1 - \alpha)\beta\phi(Ne)^i < 0, \quad (11)$$

where  $\beta = 1/(N(1 - e)((1 - e) + e\phi))$ . First, note that if  $\alpha = 0$ , the cooperative equilibrium is never stable because all the parameters involved are always positive. However, as long as  $\alpha$  is positive and  $e < 1/N$ , then the system of equations will be stable for some finite value of  $i$ . Substituting in the value of  $\beta$ , and solving eqn (11) for  $\alpha$ , we find that the minimum value of  $\alpha$  is

$$\alpha_i > \frac{e\phi(Ne)^{i-1}}{\rho(1 - e) + e\phi(1 + (Ne)^{i-1})}, \quad (12)$$

which is the same value [given in eqn (10)] derived using a less formal argument.

**Stability Condition 2.** However, if  $C > \rho(1 - e)N + (eN)^i\phi$  and  $i > 0$  then the cooperative equilibrium is stable when

$$\lambda_0 = -\alpha + (1 - \alpha)(1 - e)\beta(C - (1 - e)N\rho) < 0. \quad (13)$$

If we then solve this for the values of  $\alpha$  that create a stable cooperative equilibrium, we find

$$\alpha_i > \frac{\beta(1-e)(C - (1-e)N\rho)}{1 + \beta(1-e)(C - (1-e)N\rho)}. \quad (14)$$

Under Stability Condition 2,  $\beta = 1/(C(1-e))$ , so<sup>‡</sup>

$$\alpha_i > \frac{1 - [N\rho(1-e)/C]}{2 - [N\rho(1-e)/C]}. \quad (15)$$

The term  $N\rho(1-e)/C$  is always between zero and one, so the required  $\alpha$  is always less than  $\frac{1}{2}$ . This means that, even when the expected costs of being punished by everyone does *not* exceed the cost of cooperation (or the cost saved by defecting), the cooperative equilibrium can still be favored. Intuitively, this is the case in which conformist transmission and punishment combine to overcome the cost of cooperation. As with the previous condition, however, it is conformist transmission that stabilizes  $i$ -th stage punishment, which stabilizes first-stage punishment.

At first, Stability Condition 2 may seem strange, but the world is seemingly full of cases in which the costs of being punished seem insufficient to explain the observed degree of cooperation. Hence, this may illuminate such things as why Americans pay too much in taxes (i.e. more than they should assuming most people pay because they fear punishment; Skinner & Slemrod, 1985), why Americans wait in line, why the Aché share meat (Kaplan & Hill, 1985), and why people bother going to the voting booth (Mueller, 1989)—all of which seem overly cooperative, given the expected penalty. As we show, this may be important from a cultural group selection perspective because groups that minimize the costs of punishing and being punished ( $\rho$  and  $\phi$ ), while still maintaining cooperation, will do better than those that rely heavily on punishment to maintain cooperation.

<sup>‡</sup> Actually, there is a tiny range of  $(N\rho(1-e) + \phi(eN)^i) < C < (N\rho(1-e) + N\phi e)$  under which  $\beta$  still equals  $1/(N(1-e)(\phi(1-e) + e\phi))$ . Nothing particularly interesting happens in this range, so we will not discuss it. Note, if  $i = 1$ , the range is non-existent.

### Once Cooperation is Stabilized, it can Spread by Cultural Group Selection

By itself, the present model does not provide an explanation for human cooperation. We have shown that, under plausible conditions, a relatively weak conformist tendency can stabilize punishment, and therefore cooperation. However, non-cooperation and non-punishment is also an equilibrium of the model, and we have given no reason, so far, why most populations should stabilize at the cooperative equilibrium rather than the non-cooperative equilibrium. However, when there are multiple stable cultural equilibria with different average payoffs, *cultural group selection* can lead to the spread of the higher payoff equilibrium. As we have demonstrated above, cultural evolutionary processes will cause groups to exist at different behavioral equilibria. This means that different groups have different expected payoffs (due to different degrees of economic production, for example). The expected payoff of individuals from cooperative groups is  $b \approx (1-e)(B - C - eN(\phi + \rho(1+i)))$ , while the expected payoff of individuals in noncooperative/non-punishing groups is zero. Thus, cooperative groups will have a higher average payoff as long as the benefits of cooperation are bigger than the costs of cooperation and punishment. The combination of conformism and payoff biased transmission must also be strong enough to maintain stable cooperation in the face of migration between groups. Such persistent differences between groups creates the raw materials required by cultural group selection.

Cultural group selection can operate in a number of ways to spread prosocial behaviors. Cooperative groups will have higher total production, and consequently, more resources that can support more rapid population growth relative to non-cooperative groups. Or, cooperative groups may be better able to marshal and supply larger armies than non-cooperative groups, and hence be more successful in warfare and conquest. However, although these factors may be important (see Bowles, 2000), another, slightly subtler, cultural group selection process may also be significant. Pay-off-biased imitation means people will preferentially copy individuals who

get higher payoffs. The higher an individual's payoff, the more likely that individual is to be imitated. If individuals have occasion to imitate people in neighboring groups, people from cooperative populations will be preferentially imitated by individuals in non-cooperative populations because the average payoff to individuals from cooperative populations is much higher than the average payoff of individuals in non-cooperative populations. Boyd & Richerson (2000) have shown that, under a wide range of conditions (and fairly quickly), this form of cultural group selection will deterministically spread group-beneficial behaviors from a single group (at a group-beneficial equilibrium) through a meta-population of other groups, which were previously stuck at a more individualistic equilibrium.

### Culturally Evolved Cooperation may Cause Genes for Prosocial Behavior to Proliferate

Once the cooperative equilibrium becomes common, it is plausible that natural selection acting on genetic variation will favor genes that cause people to cooperate and punish—because such genes decrease an individual's chance of suffering costly punishment. This could arise in many ways. Individuals might develop a preference for cooperative or punishing behaviors that increases their likelihood of acquiring such behaviors. Or, alternatively, natural selection might increase the reliance on conformist transmission, making people more likely to acquire the most frequent behavior.

Here, we analyse the case in which the probability of mistakenly defecting or not-punishing,  $e$ , varies genetically. We assume that cultural evolution is much faster than genetic evolution, which implies that the population exists at a culturally evolved cooperative equilibrium. Further assume that while most individuals still make errors at the rate  $e$ , rare mutant individuals have a slightly different error probability of  $e'$  ( $= e - \varepsilon$ ), where  $\varepsilon$  is small ( $|\varepsilon| \ll e$ ). If we assume that an individual's average payoff,  $b$ , is proportional to her average genetic fitness, then we can ask whether prosocial mutants will spread. The expected fitnesses for the two types,  $F$  and  $F_m$  ("m" for mutant), and the difference between

them,  $\Delta F$ , are as follows (assuming  $i > 0$ ):§

$$\begin{aligned} F &\approx (1 - e)(B - C - eN(\phi + \rho(1 - e)(i + 1))), \\ F_m &\approx B(1 - e) - C(1 - e') \\ &\quad - N(e\phi + e'\rho(1 - e)(i + 1)), \end{aligned} \quad (16)$$

$$\Delta F = F_m - F\varepsilon(N\rho(i + 1) - C).$$

When  $\Delta F$  is positive, prosocial genes can invade. If  $C < (1 - e)N\rho + (eN)^i\phi$  (*Stability Condition 1*), then  $C$  is always less than  $N\rho(1 - e)(i + 1)$ , and prosocial genes are always favored. Once at fixation, these prosocial genes cannot be invaded by more error prone, antisocial, individuals.

In *Stability Condition 2*, where  $C > (1 - e)N\rho + (eN)^i\phi$ , prosocial genes are favored (for  $i > 0$ ) when

$$1 + \frac{(Ne)^i\phi}{N\rho(1 - e)} < \frac{C}{N\rho(1 - e)} < i + 1, \quad (17)$$

which is a wide range, since the smallest possible value of  $i$  is 1. However, there exists a range of conditions in which culturally evolved cooperation is stable, but prosocial genes cannot invade—in fact, anti-social genes (genes favoring more mistakes) may invade. This occurs when (for  $i > 0$ )

$$\underbrace{(i + 1)}_{\text{Noprosocial}} < \frac{C}{N\rho(1 - e)} < \underbrace{\frac{(1 - \alpha)}{1 - 2\alpha}}_{\text{Stability}}. \quad (18)$$

When condition (18) holds, cultural transmission will stabilize cooperation, but prosocial genes will not be able to invade—instead, anti-social genes will be favored (i.e.  $\varepsilon$  is negative). Note, however, that the *minimum* value of  $\alpha$  for this condition to exist requires  $\alpha > 0.333$ , which occurs when  $i = 1$ . Generally, we believe  $\alpha$  is much smaller than this, but we will await the verdict of future empirical work. Interestingly, this anti-social invasion is likely to occur in the groups

§ If conformist transmission alone can stabilize cooperation without any punishment ( $i = 0$ ), then  $\Delta F < 0$ , and prosocial genes will never spread.

most favored by cultural group selection—i.e. those who maximize group payoff by minimizing punishment costs (and *i*), without destabilizing cooperation. Unfortunately, anti-social invasion will decrease average payoffs, and may eventually destabilize cooperation. Further work on this gene–culture interaction will require coevolutionary models that combine both cultural and genetic evolutionary processes (perhaps using quantitative traits), and particularly the cultural group selection process we have described above.

As we have begun to model it here, prosocial genes are not strongly selected against in non-cooperative populations because error making, in terms of mistaken cooperation and punishment, only occurs when individuals adopt prosocial traits—defectors do not mistakenly cooperate. So, if the world is a mix of cooperative and non-cooperative populations, prosocial genes will be favored in a wide range of circumstances in cooperative populations and will be comparatively neutral in non-cooperative populations. It is possible that incorporating defector errors, in the form of mistaken cooperation or punishment, may affect this prediction. Furthermore, cooperation may not be a dispositional trait of individuals, but rather a specific behavior or value tied only to certain cultural domains. Some cultural groups, for example, may cooperate in fishing, and house-building, but not warfare. Other groups may cooperate in warfare, and fishing, but not house-building. Such culturally transmitted traits would have the form “cooperate in fishing”, “cooperate in house-building”, and “do not cooperate in warfare”, rather than the more dispositional approach of simply “cooperate” vs. “do not cooperate”. If this is the case, then the migration and spread of prosocial genes becomes more difficult. As prosocial genes spread among groups with different stable cooperative domains, individuals with such genes would be more likely to mistakenly cooperate in non-cooperative cultural domains. For example, in cultures where people cooperate in fishing, but not warfare, individuals with prosocial genes may be more likely to mistakenly cooperate in warfare (and pay the cost), as well as less likely to mistakenly defect in cooperative fishing. We intend to pursue those avenues in subsequent papers.

## Conclusion

We have done three things in this paper. First, we have shown that, if humans possess a psychological bias towards copying the majority, as well as a bias towards imitating the successful, then cultural evolutionary processes will stabilize cooperation and punishment for some finite number of punishment stages. Second, we discussed how, once cooperation is stable, a particular form of cultural group selection is likely to spread these group-beneficial cultural traits through human populations. And finally, we have demonstrated that prosocial genes, which cannot otherwise spread, can invade in the wake of these cultural evolutionary processes, under a wide range of conditions.

The authors would like to thank Natalie Smith, Herbert Gintis and the anonymous reviewers for their assistance and suggestions in preparing this paper.

## REFERENCES

- ASCH, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In: *Groups, Leadership and Men* (Guetzkow, H., eds), pp. 39–76. Pittsburgh: Carnegie.
- BARON, R., VANDELLO, J. & BRUNSMAN, B. (1996). The forgotten variable in conformity research: impact of task importance on social influence. *J. Pers. Soc. Psychol.* **71**, 915–927.
- BOWLES, S. (2000). Individual interactions, group conflicts and the evolution of preferences. In: *Social Dynamics* (Durlauf, S. & Young, P., eds). Washington, DC: Brookings Institution.
- BOYD, R. & RICHERSON, P. J. (1985). *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.
- BOYD, R. & RICHERSON, P. J. (1988). The evolution of reciprocity in sizable groups. *J. theor. Biol.* **132**, 337–356.
- BOYD, R. & RICHERSON, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195.
- BOYD, R. & RICHERSON, P. J. (2000). Norms and bounded rationality. In: *The Adaptive Tool Box* (Gigerenzer, G. & Selten, R., eds). Cambridge, MA: MIT Press.
- BROWN, J. L. (1983). Cooperation—a biologist's dilemma. *Adv. Stud. Behav.* **13**, 1–37.
- CAMPBELL, J. D. & FAIREY, P. J. (1989). Informational and normative routes to conformity: the effect of faction size as a function of norm extremity and attention to the stimulus. *J. Pers. Soc. Psychol.* **57**, 457–468.
- FUDENBERG, D. & MASKIN, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533–554.
- HAMILTON, W. D. (1964). The genetical evolution of social behavior. *J. theor. Biol.* **7**, 1–52.

- HARRIS, J. R. (1998). *The Nurture Assumption: Why Children Turn out the Way they do*. New York: Touchstone.
- HENRICH, J. (1999). Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change and much of sociocultural evolution. Working Paper, University of Michigan: webuser.bus.umich.edu/henrich.
- HENRICH, J. & BOYD, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evol. Hum. Behav.* **19**, 215–242.
- HENRICH, J. & GIL-WHITE, F. (2000). The evolution of prestige. Working Paper, University of Michigan: webuser. <www.bus.umich.edu/henrich>
- HIRSHLEIFER, D. & RASMUSEN, E. (1989). Cooperation in the repeated prisoner's dilemma with ostracism. *J. Econ. Behav. Organ.* **12**, 87–106.
- INSKO, C. A., SMITH, R. H., ALICKE, M. D., WADE, J. & TAYLOR, S. (1985). Conformity and group size: the concern with being right and the concern with being liked. *Pers. Soc. Psychol. Bull.* **11**, 41–50.
- KAPLAN, H. & HILL, K. (1985). *Curr. Anthropol.* **26**, 223–245.
- MCADAMS, R. H. (1997). The origin, development, and regulation of norms. *Mich. Law Rev.* **96**, 338.
- MUELLER, D. (1989). *Public Choice II*. Cambridge: Cambridge University Press.
- RICHERSON, P. J. & BOYD, R. (1998). The evolution of ultrasociality. In: *Indoctrinability, Ideology and Warfare* (Eibl-Eibesfeldt, I. & Salter, F. K., eds), pp. 71–96. New York: Berghahn Books.
- SEELEY, T. D. (1995). *The Wisdom of the Hive*. Cambridge: Harvard University Press.
- SKINNER, J. & SLEMROD, J. (1985). An economic perspective on tax evasion. *Natl. Tax J.* **38**, 345–353.
- SMITH, J. M. & BELL, P. A. (1994). Conformity as a determinant of behavior in a resource dilemma. *J. Soc. Psychol.* **134**, 191–200.
- SOBER, E. & WILSON, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
- TAKAHASI, K. (1999). Theoretical aspects of the mode of transmission in cultural inheritance. *Theor. Popul. Biol.* **55**, 208–225.
- TRIVERS, R. L. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57.
- WIT, J. (1999). Social learning in a common interest voting game. *Game. Econ. Behav.* **26**, 131–156.

## APPENDIX

For all  $i$ :

$$\Delta p_i = p_i(1 - p_i)[(1 - \alpha)\beta(\Delta b_i) + \alpha(2p_i - 1)].$$

Difference in payoff for  $i = 0$ :

$$\Delta b_0 = b_C - b_D = (1 - e)(Np_1(1 - e)\rho - C).$$

Difference in payoffs for  $i > 0$ :

$$\Delta b_i = b_{P_i} - b_{NP_i} = -(1 - e)N(\phi(1 - p_{i-1}(1 - e)) \times \prod_{j=0}^{i-2} (1 - p_{j-2}^N(1 - e)^N) - p_{i+1}(1 - e)\rho),$$

where

$$(1 - e)^N = 1 + \sum_{j=1}^N \frac{(-1)^j N! e^j}{j!(N - j)!} \approx 1 - Ne.$$

Thus,

$$\Delta b_i = b_{P_i} - b_{NP_i} \cong -(1 - e)N(\phi(1 - p_{i-1}(1 - e)) \times \prod_{j=0}^{i-2} (1 - p_{j-2}^N(1 - Ne) - p_{i+1}(1 - e)\rho).$$

Eigenvalues for the system of  $i + 1$  equations with punishment up to the  $i$ -th stage

$$\lambda_0 = -\alpha + (1 - \alpha)(1 - e)\beta(C - (1 - e)N\rho),$$

$$\lambda_j = -\alpha + (1 - \alpha)(1 - e)\beta((eN)^j \phi - \rho N(1 - e)),$$

$$0 < j < i,$$

$$\lambda_i = -\alpha + (1 - \alpha)(1 - e)\beta(eN)^i \phi.$$

When the dominant eigenvalue (that with the largest value) is less than zero, the system is locally stable at point  $(p_0, p_1, \dots, p_{i+1}) = (1, 1, \dots, 0)$ .