Altruistic Punishment and the Origin of Cooperation

James H. Fowler*

October 4, 2004

** Department of Political Science, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA*

How did human cooperation evolve? Recent evidence from anonymous public goods experiments shows that many people are willing to engage in *altruistic punishment*, voluntarily paying a cost to punish noncooperators. While this behaviour helps to explain how cooperation can persist, it creates an important puzzle. If altruistic punishment provides benefits to nonpunishers and is costly to punishers, then how could it evolve? Drawing on recent insights from voluntary public goods games, I present a simple evolutionary model in which altruistic punishers can enter and will always come to dominate a population of contributors, defectors, and nonparticipants. The model suggests that the rock-paper-scissors cycle in voluntary public goods games does not persist in the presence of punishment strategies. It also suggests that punishment can only enforce payoff-improving strategies, contrary to a widely-cited "folk theorem" result that suggests punishment can allow the evolution of *any* strategy.

Human beings frequently cooperate with genetically unrelated strangers whom they will never meet again, even when such cooperation is individually costly[1]. This behaviour is puzzling since natural selection works against those willing to engage in costly cooperation and in favour of those who 'free ride' on their efforts. Several theories have been advanced to explain the persistence of cooperative behaviour such as the theory of kin selection[2], and theories of direct[3] and indirect[4] reciprocity. However,

none of these theories can explain cooperation between unrelated individuals when interactions are not repeated and reputation effects are absent.

Punishment may yield a solution to the problem of cooperation. Laboratory[5,6] and ethnographic[7] evidence suggests that many people are willing to engage in *altruistic punishment*, paying a personal cost to punish free riders in public goods games. They do so even when interactions are completely anonymous and there are no reputation effects, and when the punisher is a third-party who is unaffected by the free rider's actions[8]. Altruistic punishment has also been shown to stimulate the reward center in the brain, suggesting that humans may have physically or developmentally evolved this behavior[9]. But this is equally puzzling since natural selection should work against those who engage in costly punishment and in favour of those who free ride on the cooperative benefits generated by their efforts.

Previous efforts to show how altruistic punishment might have evolved rely on models of group selection rather than individual selection[10-13]. Only one of these models[12] demonstrates that altruistic punishment is evolutionarily stable when it is common. More importantly, these authors point out that when punishers first enter a population there are few punishers and many free riders, so the cost of punishing is very large relative to the cost of being punished. Thus, they cannot explain how such behaviour could emerge when it is rare in a single large group. How can we explain the origin of altruistic punishment, and with it, the origin of cooperation in humans?

Suppose a population has an opportunity to create a public good. Contributors (C) benefit from the public good and pay an individual cost c to increase the size of the public good by b. Defectors (D) also benefit from the public good, but they do not pay a cost to cooperate. In recent work, scholars have considered a third behavioural type, the nonparticipant.[14-16] Nonparticipants (N) neither pay a cost nor receive a benefit

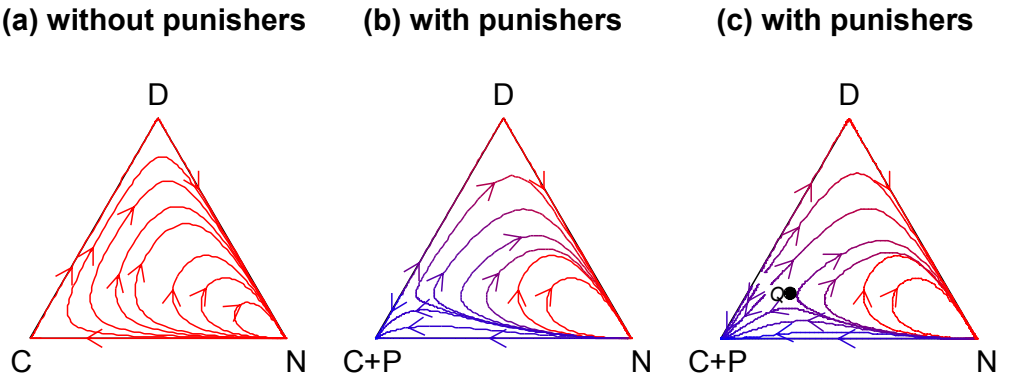

Figure 1. Population dynamics in the public goods game without (a) and with (b,c) altruistic punishers. The vertices denote homogenous populations of defectors, nonparticipants, and (a) contributors or (b,c) contributors and punishers. The hue of the orbit denotes the ratio of punishers to contributors (more red = more contributors, more blue = more punishers). Orbits indicate the replicator dynamics $\dot{x}_i = x_i(\pi_i - \bar{\pi})$. A stationary point *Q* appears for some parameter combinations as in (c) but it is never stable (see Methods). Parameters: $b = 3,\ c = 1,$ (b) $p = 2,\ k = 1,\ \alpha = 0.1,$ (c) $p = 3,\ k = 1,\ \alpha = 0.2$.

from the public good. Instead, they receive a fixed benefit σ for engaging in other activities. If we let x_i denote the proportion of each type in the population, then the expected payoffs π_i are $bx_C/(1-x_N) - c$ for contributors, $bx_C/(1-x_N)$ for defectors, and σ for nonparticipants.

To analyze the dynamics of the population, note that a wide variety of imitation and genetic inheritance processes yield the standard replicator dynamics[17] $\dot{x}_i = x_i(\pi_i - \bar{\pi})$. Figure 1a shows that under this assumption a population of contributors, defectors, and nonparticipants displays a cycle similar to the game of rock-paper-scissors. If the net benefit of a contribution to the public good exceeds the payoff from other activities, $b - c > \sigma$, then a mutant cooperator can invade a population of

nonparticipants and even take over the whole population. However, cooperation is short-lived because the growth of the population of contributors creates an environment in which defectors can benefit from the public good without paying for it. As cooperation collapses, the public good shrinks, and nonparticipants once again take over the population since they receive a small fixed payoff.

Suppose a fourth type, the altruistic punisher (P), enters the population. Each punisher contributes to and benefits from the public good, but also engages in altruistic punishment with a randomly chosen individual. Like the "moralists" in a previous model[12], the punisher will punish both defectors and nonpunishing contributors (note that punishment of both contributors and those who do not intend to defect has been observed in laboratory experiments[5,9]) . If the punisher is matched with a defector then she pays a cost k to incur a punishment p on the defector for not contributing. If she is matched with a (nonpunishing) contributor then she pays a cost αk to incur a punishment αp on the contributor for not punishing, where $0 < \alpha < 1$. Punishers ignore nonparticipants since they neither contribute to nor benefit from the public good. This changes the expected payoffs to $b(x_C + x_P)/(1 - x_N) - c - apx_P$ for contributors, $b(x_C + x_P)/(1 - x_N) - px_P$ for defectors, σ for nonparticipants, and $b(x_C + x_P)/(1 - x_N) - c - kx_D - \alpha kx_C$ for punishers.

Figures 1(b,c) show the dynamics of a population with punishers. Although the rock-paper-scissors cycle continues, there is now a significant region where the population tends towards toward all punishers. Moreover, a single punisher can invade a population of nonparticipants, and the unique evolutionarily stable population is composed entirely of punishers (see Methods). These results are robust to large populations and a wide range of parameters—the only restrictions are that the parameters must all be positive, the net benefit of the public good must exceed the

payoff from nonparticipation ($b - c > \sigma$), and the effect of punishment must be larger than the cost of contributing to the public good ($p > c$).

Some may object that altruistic punishment cannot explain cooperation because of difficulties in monitoring—there may only be a small chance of learning that another individual failed to contribute or failed to punish other noncontributors. However, it is important to note that *p* may represent the *expected* impact of punishment reflecting both the actual cost of the penalty and the probability of being identified by a given individual. The cooperative equilibrium is still attainable as long as $E(p)>c$. Moreover, the punishment of nonpunishing contributors can be arbitrarily small or infrequent since any $\alpha > 0$ gives punishers an advantage over contributors.

To conclude, this model has several important implications. First, it shows how altruistic punishment can emerge in a population where there is both an incentive not to contribute and an incentive not to punish noncontributors. Past work[12] has shown that punishment strategies can persist under these conditions, but it has not explained how such strategies might evolve. In contrast, this model demonstrates that both the origin and persistence of widespread cooperation is possible with voluntary, decentralized enforcement, even in very large populations under a broad range of conditions.

Second, the model suggests that the cycle of cooperation, defection, and nonparticipation recently identified by scholars[14-16] is important for understanding the *origin* of cooperation, but may not be useful for understanding its *persistence*. Once altruistic punishment evolves, the cycle should disappear and cease to be an important mechanism in the population.

Finally, the model questions a "folk theorem" result[12] which indicates that punishment strategies can enforce any other strategy, even those that yield a payoff

disadvantage. Notice that when participation is optional, punishers can only evolve and persist if they yield a payoff *advantage* $b - c > \sigma$. This suggests that there are restrictions on what kinds of strategies punishment can actually enforce.

Methods

Proof that a population with all Punishers is the unique evolutionarily stable population

A given population is evolutionarily stable if it cannot be invaded by an arbitrarily small mutation. Consider a population of contributors, defectors, nonparticipants, and punishers with payoffs as described in the text. When $0 < x_P < 1$ and $0 < x_D < 1$, notice that $\frac{\partial \dot{x}_N}{\partial x_N} = c\left(c + (p+k)(\alpha x_C + x_D - x_P) + k + \frac{bx_N}{1-x_N} \right)$ and $\frac{\partial \dot{x}_D}{\partial x_D} = -c - (p+k)(\alpha x_C + x_D - x_P) + p + \frac{bx_N}{1-x_N}$ at any stationary point $\dot{\mathbf{x}} = 0$ when x_C and x_N are held constant. If either of these derivatives is positive, it means a mutant can invade the population. Given that c is positive, in order for both derivatives to be nonpositive at a given point it must be true that $p + k + \frac{2bx_N}{1-x_N} < 0$, but this is always false since p, b, and k are positive and $0 \le x_N \le 1$. Thus, an opportunity always exists either for a single defector or for a single punisher to invade the population.

Now consider the case when $x_D = x_P = 0$. Without punishment and defection, contributors gain an average payoff of b which is always larger than the nonparticipants' payoff of σ under the assumption $b - c > \sigma$. Thus the only stationary point is the population $x_C = 1$, but this point is not stable because a single defector can invade with payoff bx_C compared to the contributors payoff of $bx_C - c$.

In the case $x_D = 1$ the population is not stable since a single nonparticipant can invade with payoff σ compared to the defector's payoff of 0 in the absence of any contributors or punishers.

The remaining case $x_P = 1$ is the only evolutionarily stable population. Punisher payoffs are always larger than nonparticipant payoffs since $b - c > \sigma$. Punishers resist invasion by a fraction ε of defectors if $b(1-\varepsilon) - c - k\varepsilon > b(1-\varepsilon) - p(1-\varepsilon)$, or $\varepsilon < (p-c)/(k+p)$. This inequality is true for some positive ε as long as $p > c$. Finally, punishers resist invasion by a fraction ε of contributors if $b(1-\varepsilon) - c - \alpha k\varepsilon > b(1-\varepsilon) - c - \alpha p(1-\varepsilon)$, or $\varepsilon < p/(k+p)$. This inequality is true for some positive ε as long as p and k are positive.

1. Sober, E. & Wilson, D. S. *Unto others : the evolution and psychology of unselfish behavior* (Harvard University Press, Cambridge, Mass., 1998).
2. Hamilton, W. D. The Genetic Evolution of Social Behavior I and II. *Journal of Theoretical Biology* **7**, 1-52 (1964).
3. Axelrod, R. & Hamilton, W. D. The Evolution of Cooperation. *Science* **211**, 1390-1396 (1981).
4. Nowak, M. A., Sigmund, K. & Source:. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573-577 (1998).
5. Fehr, E. & Gachter, S. Altruistic punishment in humans. *Nature* **415**, 137-140 (2002).
6. Fehr, E. & Gachter, S. Cooperation and punishment in public goods experiments. *American Economic Review* **90**, 980-994 (2000).
7. Boehm, C. Egalitarian Behavior and Reverse Dominance Hierarchy. *Current Anthropology* **34**, 227-254 (1993).
8. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evolution and Human Behavior* **25**, 63-87 (2004).
9. Quervain, D. J.-F. d. et al. The Neural Basis of Altruistic Punishment. *Science* **305**, 1254-8 (2004).
10. Bowles, S. & Gintis, H. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology* **65**, 17-28 (2004).
11. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3531-3535 (2003).
12. Boyd, R. & Richerson, P. J. Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology* **13**, 171-195 (1992).
13. Gintis, H. Strong reciprocity and human sociality. *Journal of Theoretical Biology* **206**, 169-179 (2000).
14. Hauert, C., De Monte, S., Hofbauer, J. & Sigmund, K. Volunteering as Red Queen mechanism for cooperation in public goods games. *Science* **296**, 1129-1132 (2002).
15. Semmann, D., Krambeck, H. J. R. & Milinski, M. Volunteering leads to rock-paper-scissors dynamics in a public goods game. *Nature* **425**, 390-393 (2003).

16. Hauert, C., De Monte, S., Hofbauer, J. & Sigmund, K. Replicator dynamics for optional public good games. *Journal of Theoretical Biology* **218**, 187-194 (2002).
17. Weibull, J. W. *Evolutionary Game Theory* (MIT Press, Cambridge, Massachusetts, 1995).

Acknowledgments The author thanks … for helpful comments.

Competing Interests Statement The author declares that he has no competing financial interest.

Correspondence and requests for materials should be addressed to J.F. (e-mail: jhfowler@ucdavis.edu).