

Altruistic Punishment and the Origin of Cooperation

James H. Fowler

Department of Political Science, University of California, San Diego

Published in

Proceedings of the National Academy of Sciences 102 (19): 7047-7049 (10 May 2005)

How did human cooperation evolve? Recent evidence 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 cycle of strategies 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 (*I*). 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.

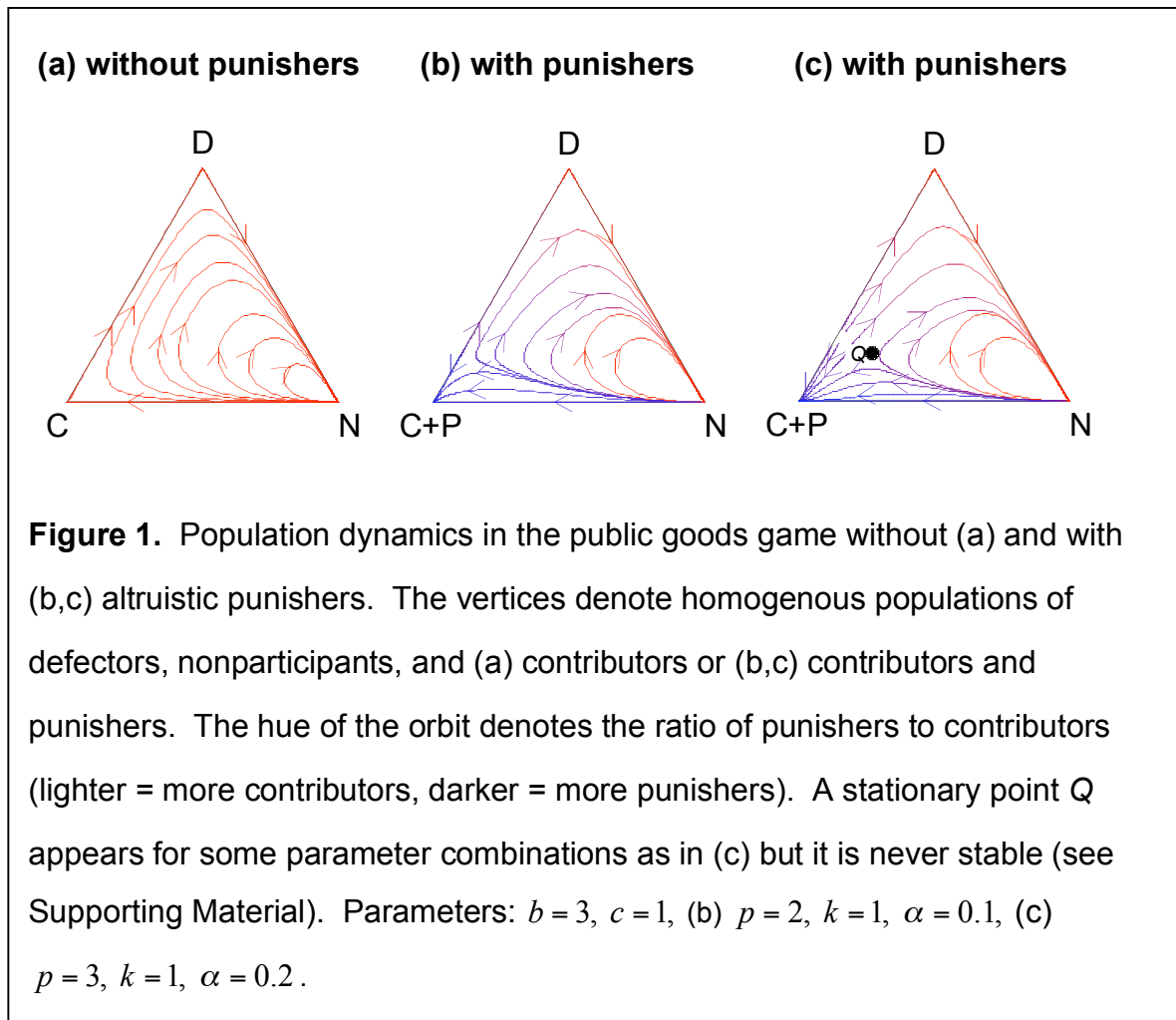
Punishment may yield a solution to the problem of cooperation. Laboratory (5, 6) and ethnographic (7, 8) evidence suggests that many people are willing to engage in *altruistic punishment*, paying a personal cost to punish free riders in public goods games. They do so even when interactions are completely anonymous and there are no reputation effects, and when the punisher is a third party who is unaffected by the free rider's actions (9). Altruistic punishment has also been shown to stimulate the reward center in the brain, suggesting that humans may have physically or developmentally evolved this behaviour (10). 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 punishers.

Previous efforts to show how altruistic punishment might have evolved typically rely on models of group selection rather than individual selection.(11-14) These models show that altruistic punishment is evolutionarily stable when it is common. However, they have difficulty explaining the emergence of punishment. 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. One recent model (15) attempts to solve this problem by allowing altruistic punishment and norm internalization to coevolve. This model shows that prosocial norms like altruistic punishment can emerge by "hitchhiking" on genes associated with norm internalization. However, it also shows that anti-social norms can emerge and relies on simulations of group selection to show that prosocial norms are more likely to evolve. How might altruistic punishment evolve in an individual selection context?

Suppose a large population has an opportunity to create a public good which is distributed equally to everyone in the population. Contributors (C) pay an individual cost c to increase the size of the public good by b . Defectors (D) do not contribute. If we let x_i denote the proportion of each type in the population, then the expected fitness π_i is $bx_C - c$ for contributors and bx_C for defectors. To analyze the dynamics of the population, suppose individuals occasionally compare their own performance with the performance of another randomly selected individual and then adopt the strategy with higher fitness. This process and a wide variety of imitation and genetic inheritance processes yield the standard replicator dynamics $\dot{x}_i = x_i(\pi_i - \bar{\pi})$, where $\bar{\pi} = \sum_i x_i \pi_i$ represents the average fitness level in the population. (16) Under this assumption, defectors will always take over the population because they always have a higher fitness than contributors.

So far we have assumed that behavioural types are restricted to the choice of whether or not to contribute to the public good. However, in many situations there is another choice. For example, individuals may face a choice between joining a hunting party and hunting on their own. The game they catch if they join the hunting party may be much larger than the game they can catch on their own. However, their expected share depends on the sum of the efforts of those who decide to join the party. If several defectors join, the expected share of the good will diminish and it may make more evolutionary sense to engage in other activities. We can think of those who decide not to join the party as nonparticipants (N).

As in recent work by other scholars (17-19), we will assume that nonparticipants neither pay a cost nor receive a benefit from the public good. Instead, they receive a fixed benefit σ for engaging in other activities. If we allow for this type in the population, then the expected payoffs are $bx_C/(1-x_N) - c$ for contributors, $bx_C/(1-x_N)$ for defectors, and σ for nonparticipants. Figure 1a shows that the resulting population



dynamics display a cycle. If contributors can produce a net benefit for the population that 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. Like the “moralists” in a previous model (13), punishers contribute to and benefit from the public

good, and engage in altruistic punishment with *both* defectors and nonpunishing contributors. Each punisher pays a cost k to incur a punishment p on the population of defectors and a cost αk to incur a punishment αp on the population of contributors who do not punish, where $0 < \alpha < 1$. Punishers ignore nonparticipants since they neither contribute to nor benefit from the public good. This changes the expected payoffs to $b(x_C + x_P)/(1 - x_N) - c - \alpha p x_P$ for contributors, $b(x_C + x_P)/(1 - x_N) - p x_P$ for defectors, σ for nonparticipants, and $b(x_C + x_P)/(1 - x_N) - c - k x_D - \alpha k x_C$ for punishers.

Figures 1(b,c) show the dynamics of a population with punishers. Although the 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 Supporting Material). 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 to the population of an individual contribution 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 ($p > c$). Moreover, these results all take place within the context of a single population, rather than between groups as in other models.(11-14) Nonparticipants do interact with participants in this model—they simply make the choice not to contribute to or benefit from the collective activity. When punishers invade the population, defectors are held at bay and the collective activity becomes much more lucrative. In the end, nonparticipants become participants because the defection problem is solved.

This model has certain features in common with models of good standing.(20, 21) For example, punishers in this model must be able to distinguish between defectors who are in ‘bad’ standing and cooperators who are in ‘good’ standing to determine who receives punishment. However, unlike previous models of good standing, the model

presented here also considers the possibility that some individuals will avoid a bad standing designation by not participating. This feature of the model prevents defectors from completely taking over the population since they are susceptible to nonparticipants.(17-19) Thus, while standing models have already been shown to have a cooperative equilibrium (20), these models also have a noncooperative equilibrium that does not occur in the model presented here.

Another difference between this model and models of good standing is that the mechanics of identifying who is or is not in good standing have not been fully modelled here. As a result, some may object that altruistic punishment cannot explain cooperation because of difficulties in monitoring—there may only be a small probability q of learning that other individuals failed to contribute or failed to punish other noncontributors. However, this probability can be easily incorporated into the model by substituting pq for p . Notice that the cooperative equilibrium is reachable as long as $pq > c$, suggesting that larger punishments may be able to offset any decrease in the probability of detection.

Several objections might be raised against this model. For example, the infrequency of punishment of nonpunishers observed in laboratory experiments (5, 10) might not be enough to keep nonpunishing cooperators from taking over the population. However, the model suggests that punishment of nonpunishing contributors can be arbitrarily small or infrequent since any $\alpha > 0$ gives punishers an advantage over contributors. Along these same lines, some may note that there is a second-order defection problem since a population of punishers with a given α can be invaded by punishers with a lower α . However, if punishers also punish anyone who does not punish nonpunishers enough (at least α) then they will be secure against such an invasion (see Supporting Material). Finally, some may worry that the option not to participate is merely a mathematical convenience to reach equilibrium. However,

models without nonparticipants implicitly assume that defection carries with it no opportunity cost. In many cases like the hunting example mentioned above, nonparticipants who rely on their own activities will outcompete defectors who rely on goods provided by others since the presence of defectors undermines the very provision of those goods. As a result, cooperation-enhancing strategies like altruistic punishment have an opportunity to evolve because they simultaneously acquire more benefits than nonparticipants and keep defectors at bay.

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 (11-15) has shown that punishment strategies can persist under these conditions, but it has relied on group selection to explain how such prosocial strategies might evolve. In contrast, this model demonstrates that both the origin and persistence of widespread cooperation is possible with voluntary, decentralized, anonymous 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 (17-19) 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 observed in the population dynamics.

Finally, the model questions a “folk theorem” result (13) 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* $b - c > \sigma$ to the population. This suggests that there are restrictions on what kinds of strategies punishment can actually enforce.

Supporting Material

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. In the case when $0 < x_p < 1$ and $0 < x_D < 1$, notice that

$$\frac{\partial \dot{x}_p}{\partial x_p} = \left(c - \alpha p + (p + k)(x_D + \alpha x_C - x_p) + \frac{bx_N}{1 - x_N} \right) x_p \text{ and}$$

$$\frac{\partial \dot{x}_D}{\partial x_D} = \left(-c - (p + k)(x_D + \alpha x_C - x_p) + p - \frac{bx_N}{1 - x_N} \right) x_D \text{ at any stationary point } \dot{\mathbf{x}} = 0$$

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 x_p and x_D are positive, both derivatives are nonpositive at a given point only if their sum is less than zero, but this is only true if $p(1 - \alpha) < 0$. Thus, any size positive punishment will mean 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

$b - c - \alpha k\varepsilon > b - c - \alpha p(1 - \varepsilon)$, or $\varepsilon < p/(k + p)$. This inequality is true for some positive ε as long as p and k are positive.

This evolutionary stable population also resists invasion by mutant punishers that punish defectors but not cooperators as long as the rule used by Punishers is to punish anyone who punishes nonpunishers by an amount less than α . For example, suppose the extreme case of a shirker (S) who cooperates and punishes only defectors and no one else. The shirker's payoff will be $b(x_C + x_P + x_S)/(1 - x_N) - c - kx_D - \alpha px_S$ and the punisher's payoff will change to $b(x_C + x_P + x_S)/(1 - x_N) - c - kx_D - \alpha k(x_C + x_S)$. Notice that punishers resist a fraction ε of shirkers if $b - c - \alpha k\varepsilon > b - c - \alpha p(1 - \varepsilon)$, or $\varepsilon < p/(k + p)$.

References

1. Sober, E. & Wilson, D. S. (1998) *Unto others : the evolution and psychology of unselfish behavior* (Harvard University Press, Cambridge, Mass.).
2. Hamilton, W. D. (1964) *Journal of Theoretical Biology* **7**, 1-52.
3. Axelrod, R. & Hamilton, W. D. (1981) *Science* **211**, 1390-1396.
4. Nowak, M. A. & Sigmund, K. (1998) *Nature* **393**, 573-577.
5. Fehr, E. & Gächter, S. (2002) *Nature* **415**, 137-140.
6. Fehr, E. & Gächter, S. (2000) *American Economic Review* **90**, 980-994.
7. Boehm, C. (1993) *Current Anthropology* **34**, 227-254.
8. Henrich, J. P. (2004) *Foundations of human sociality : economic experiments and ethnographic evidence from fifteen small-scale societies* (Oxford University Press, Oxford ; New York).
9. Fehr, E. & Fischbacher, U. (2004) *Evolution and Human Behavior* **25**, 63-87.
10. Quervain, D. J.-F. d., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A. & Fehr, E. (2004) *Science* **305**, 1254-8.
11. Bowles, S. & Gintis, H. (2004) *Theoretical Population Biology* **65**, 17-28.
12. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. (2003) *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3531-3535.
13. Boyd, R. & Richerson, P. J. (1992) *Ethology and Sociobiology* **13**, 171-195.
14. Gintis, H. (2000) *Journal of Theoretical Biology* **206**, 169-179.
15. Gintis, H. (2003) *Journal of Theoretical Biology* **220**, 407-418.
16. Weibull, J. W. (1995) *Evolutionary Game Theory* (MIT Press, Cambridge, Massachusetts).
17. Hauert, C., De Monte, S., Hofbauer, J. & Sigmund, K. (2002) *Science* **296**, 1129-1132.
18. Semmann, D., Krambeck, H. J. R. & Milinski, M. (2003) *Nature* **425**, 390-393.
19. Hauert, C., De Monte, S., Hofbauer, J. & Sigmund, K. (2002) *Journal of Theoretical Biology* **218**, 187-194.

20. Panchanathan, K. & Boyd, R. (2003) *Journal of Theoretical Biology* **224**, 115-126.
21. Sugden, R. (1986) *The Economics of Rights, Cooperation and Welfare* (Basil Blackwell, Oxford).