Via Freedom to Coercion: The Emergence of Costly Punishment

Published in final edited form as:
Science. 2007 June 29; 316(5833): 1905–1907.

Via freedom to coercion: the emergence of costly punishment

Christoph Hauert[1], **Arne Traulsen**[1], **Hannelore Brandt**[2], **Martin A. Nowak**[1], and **Karl Sigmund**[3,4,*]

1 *Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, One Brattle Square, Cambridge, MA 02138, USA*

2 *Vienna University of Economics and Business Administration A-1090 Vienna, Austria*

3 *Faculty of Mathematics, University of Vienna A-1090 Vienna, Austria*

4 *International Institute for Applied Systems Analysis A-2361 Laxenburg, Austria*

Abstract

In human societies, cooperative behaviour in joint enterprises is often enforced through institutions that impose sanctions on defectors. Many experiments on so-called public goods games have shown that in the absence of such institutions, individuals are willing to punish defectors, even at a cost to themselves. Theoretical models confirm that social norms prescribing the punishment of uncooperative behaviour are stable: once established, they prevent dissident minorities from spreading. But how can such costly punishing behaviour gain a foothold in the population? A surprisingly simple model shows that if individuals have the option to stand aside and abstain from the joint endeavour, this paves the way for the emergence and establishment of cooperative behaviour based on the punishment of defectors. Paradoxically, the freedom to withdraw from the common enterprise leads to enforcement of social norms. Joint enterprises which are compulsory rather than voluntary are less likely to lead to cooperation.

Keywords

evolutionary game theory; public goods games; cooperation; altruistic punishment; voluntary interactions

An impressive body of evidence shows that many humans are willing to pay a personal cost in order to punish wrong-doers (1–8). In particular, punishment is an effective mechanism to ensure cooperation in public goods interactions (9–11). All human populations seem willing to use costly punishment to varying degrees, and their willingness to punish correlates with the propensity for altruistic contributions (12). This raises an evolutionary problem: in joint enterprises, free-riding individuals who do not contribute, but exploit the efforts of others, fare better than those who pay the cost of contributing. If successful behaviour spreads, for instance through imitation, these defectors will eventually take over. Punishment reduces the defectors' payoff, and thus may solve the social dilemma. But since punishment is costly, it also reduces the punishers' payoff. This raises a 'second order social dilemma': costly punishment seems to be an altruistic act, since individuals who contribute, but do not punish, are better off than the punishers.

*Corresponding author: Karl Sigmund, Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria, e-mail: karl.sigmund@univie.ac.at, phone: +43 (0)1 4277 506 12, fax: +43 (0)1 4277 9 506.

The emergence of costly punishing behaviour is acknowledged to be a major puzzle in the evolution of cooperation. ”We seem to have replaced the problem of explaining cooperation with that of explaining altruistic punishment” (13).

This puzzle can be solved in situations where individuals can decide whether to take part in the joint enterprise or not. We consider four strategies. The *non-participants* (individuals who, by default, do not join the public enterprise) rely on some activity whose payoff is independent of the other players' behavior. Those who participate include *defectors*, who do not contribute but exploit the contributions of the others; *cooperators*, who contribute, but do not punish; and *punishers*, who not only contribute to the commonwealth, but also punish the defectors. We show that in such a model, punishers will invade and predominate. However, in the absence of the option to abstain from the joint enterprise, punishers are often unable to invade, and the population is dominated by defectors. This means that if participation in the joint enterprise is voluntary, cooperation-enforcing behaviour emerges. If participation is obligatory, then the defectors are more likely to win.

This intriguing result was originally presented by Fowler (14) but his argument was based on a model which lacked an explicit micro-economical foundation. It assumes (a) that single cooperators can play the public goods game alone, which neglects the fact that contributing to a joint effort is a risky investment, whose return depends on what other players are doing; (b) that cooperators will be punished, even in the absence of defectors, which neglects the fact that the cooperators' unwillingness to punish cannot be observed in that case. Correcting for this leads to a dynamics which is structurally unstable for infinitely large populations and hence inconclusive (15). It is thus necessary to tackle the stochastic dynamics of finite populations.

We consider a well-mixed population of constant size M whose members live on a small, but fixed income σ. In this situation, N individuals are randomly selected and offered the option to participate instead in a risky, but potentially profitable public goods game. Those who participate can decide whether or not to contribute an investment at a cost c to themselves. All individual contributions are added up and multiplied with a factor $r > 1$. This amount is then divided equally among all participants of the public good game. After this interaction, each contributor can impose a fine β upon each defector, at a personal cost γ for each fine. By x we denote the total number of cooperators, by y that of defectors, by z that of the non-participants, and by w the number of punishers. Thus $M = x + y + z + w$.

Among the random sample of size N, there will be N_x cooperators, N_y defectors, N_z non-participants and N_w punishers. These are random variables distributed according to a multivariate distribution which describes sampling without replacement. Each non-participant receives a constant payoff σ. The group of those willing to participate in the public goods game has size $S := N_x + N_y + N_w$. If $S > 1$, each participant of the public goods game obtains an income $r(N_x + N_w)c/S$. The payoff for the contributors (i.e. the cooperators and the punishers) is reduced by c. The payoff for the defectors is reduced by βN_w, and the payoff for punishers by γN_y. The social enterprise is risky in the sense that if all defect, the payoff is below that of the non-participants; it is promising in the sense that if all cooperate, the payoff is larger than that of the non-participants. This means that $0 < \sigma < (r - 1)c$. This assumption offers players a non-trivial choice: to stick with a safe, self-sufficient income, or to speculate on a joint effort whose outcome is uncertain because it depends on the decisions of others. (If $S = 1$ then the public goods game does not take place. In this case a single player who volunteers for the joint effort receives the default payoff σ.)

We next specify how strategies propagate within the population. We only need to assume that players can imitate each other, and are more likely to imitate those with a higher payoff. This can be done in various ways (see (16) and (17)). For simplicity, let us assume here that players

can update their strategy from time to time by imitating a player chosen with a probability which is linearly increasing with that player's payoff. In addition, we shall assume that with a small probability μ, a player can switch to another strategy irrespective of its payoff (we refer to this as 'mutation' without implying a genetic cause: it simply corresponds to blindly experimenting with the alternatives).

The analysis of the corresponding stochastic dynamics is greatly simplified in the limiting case. $\mu \to 0$ The population consists almost always of one or two types at most. Indeed, for $\mu = 0$ the four monomorphic states are absorbing: if all individuals use the same strategy, imitation will not introduce any change. For sufficiently small μ the fate of a mutant (i.e. its elimination or fixation) is settled before the next mutant appears (18). This allows to calculate the probability that the population is in the vicinity of a pure state (i.e. composed almost exclusively of one type) (17). Computer simulations show that the approximation also holds for larger mutation rates (on the order of $1/M$).

The outcome is striking: in the limit of rare mutations, the system spends most of the time in the homogeneous state with punishers only, irrespective of the initial composition of the population. For large populations ($M = 1000$ can be considered large for most of our prehistory) and small mutation rates, the system spends most of the time in or near the punisher state (Figs. 1a and 2a, as well as Fig. S1). The outcome is robust with respect to changes in $\sigma$and r (Fig. S1).

The situation is very different in the traditional case of a public goods game where participation is compulsory. If only cooperators and defectors are present, defectors obviously win. Adding the punishers as a third strategy does not change the qualitative outcome: In the limit of rare mutations, the system spends most of the time in or near the state with defectors only. For the same parameter values as before, the state is time dominated by defectors, and there is hardly any economic benefit from the interaction (Fig. 1b and 2b, and Fig. S2).

Volunteering in the absence of punishment leads to a more cooperative outcome than for the obligatory game, but not to the fixation of the cooperative state (Fig. 3a). Instead, the system exhibits a strong tendency to cycle (from cooperation to defection to non-participation and back to cooperation), due to a rock-paper-scissors mechanism (19–21). If there are many defectors, it does not pay to participate in the joint enterprise, but if most players refuse to participate, then the typical group size can become sufficiently small such that the social dilemma disappears: cooperators earn on average more than defectors (and non-participants). However, this is a fleeting state only: quickly cooperators spread, group size increases, the social dilemma returns and the cycle continues.

The gist of the analysis for small mutation rates is captured in Fig. 2. The effect of substantial mutation rates can only be handled by numerical simulations ((17) and (22)). In the absence of punishers, defectors do worst, whereas non-participants and cooperators perform comparably well. In the compulsory game, punishers do not prevail, except for large mutation rates, in which case mutational drift supplying defectors keeps the punishers active and prevents them from being undermined by cooperators. If all four types are admitted, punishers prevail.

This result remains unaffected if we assume that the punishers are also punishing the cooperators (who are not punishing defectors, and thus can be viewed as second-order defectors). It is well-known that any norm that includes the rule to punish those who deviate is evolutionarily stable: once established, it cannot be displaced by an invading minority of dissidents (9). But how can such punishing behaviour gain a foothold in the population? The trait has to be rare, initially, and thus will incur huge costs by ceaselessly punishing. To model this situation it seems plausible to assume that for this second type of punishment, fines and costs are reduced by a factor α, with $0 \leq \alpha \leq 1$ (14). Thus the payoff for cooperators is reduced

by $\alpha\,\beta N_w$, and that for punishers by $\alpha\,\gamma N_x$, provided that $N_y > 0$ (if there are no defectors in the group, non-punishing behaviour will go unnoticed). As it turns out, whether cooperators who fail to punish are punished or not plays a surprisingly small role. The parameter α has little influence on the dynamics (17). The reason is that for small μ, the three types of punishers, cooperators and defectors rarely co-exist: hence punishers cannot hold cooperators accountable for not punishing defectors. Interestingly, experimental evidence for the punishment of non-punishers (i.e. for non-vanishing α) seems to be lacking (23).

We could also assume that punishers penalise non-participants, with a fine $\delta\beta$ and the cost to the punisher $\delta\gamma$ (with $0 \leq \delta \leq 1$). Although this further stabilizes punishment once it is established, it also hinders the emergence of punishment (see Fig. 3b, and (17)). It follows that resorting to stricter forms of social coercion may not be an efficient way to increase cooperation: second order punishment ($\alpha > 0$) barely affects the outcome whereas punishing non-participants ($\delta > 0$) can even lead to contrary effects. The system responds to an increase in compulsion with a decrease in cooperation.

When punishers are common, individual level selection against them is weak (since only little punishment occurs), and may be overcome by selection among groups (10). Several other models confirm that the punishment of defectors is stable, if it is the prevalent norm. This happens for example if some degree of conformism in the population is assumed (11): individuals preferentially copy what is frequent. Similarly, cooperation in the public goods game can also be stabilised through additional rounds of pairwise interactions based on indirect reciprocity: in this case, players can reward contributors (24,25). But in each case, the emergence of the pro-social norm remains an open problem (26,27).

Our model, in contrast, shows that even when initially rare, punishing behaviour can be advantageous, and is likely to become fixed. We consider the most challenging scenario, namely a single well-mixed population whose members imitate preferentially what fares better, not what is more common. Once established, group selection, conformism, and reputation effects may maintain pro-social norms and promote their spreading. Eventually, institutions for punishing free-riders may arise, or genetic predispositions to punish dissidents.

Recent experiments show that if players can choose between joining a public goods game either with or without punishment, they prefer the former (28). The interpretation seems clear: whoever freely accepts that defection may be punished is unlikely to be a defector. For contributors, it is thus less risky to join such a group. Players voluntarily commit themselves to sanctioning rules. This voluntary submission is not immediate, however: in the majority of cases it requires a few preliminary rounds. Many players appear to have initial reservations against the possibility of sanctions and need a learning phase. In another series of experiments, it has been shown that a threat of punishment can decrease the level of cooperation in trust games (29). Experimental evidence for costly punishment can also be found in the ultimatum game (rejecting an unfair offer is costly to both players) (2) and in indirect reciprocity (by not helping defectors, players reduce their own chances of being helped) (30). If punishment is combined with rewarding through indirect reciprocity, punishment is focussed on the worst offenders, and is otherwise strongly reduced in favour of rewarding contributors (31). In all these investigations, and in the experiments on voluntary public goods games without punishment (21), there is ample evidence that players can adapt their strategy from one round to the next, as a reaction to the current state of the population. Our model is based on this aptitude for social learning.

In our framework, the joint effort represents an innovation, a new type of interaction which improves the payoff of participants if it succeeds, but costs dear if it fails. Abstaining from such a risky enterprise does not mean living a hermit’s life. It means collecting mushrooms

instead of participating in a collective hunt; remaining at home in lieu of joining a raiding party; dispersing in the woods rather than erecting a stronghold against an invader; growing potatoes on one's plot of land instead of handing it over to a commons likely to be ruined by overgrazing.

Our model predicts that if the joint enterprise is optional, cooperation backed by punishment is more likely than if the joint enterprise is obligatory. Sometimes, public goods cannot be opted out of: the preservation of our climate is one example (32). In that case, participation is obligatory – and defection widespread.

Reports from present-day hunter-gatherer societies often stress their egalitarian and 'democratic' features: individuals have a great deal of freedom (33). This creates favourable conditions for voluntary participation. On the other hand, ostracism was probably an early form of severe punishment. There seems to be a smooth transition between choosing not to take part in a joint enterprise, and being excluded. Together, these two alternatives may explain the emergence of rule-enforcing institutions promoting pro-social behaviour, following Hardin's recipe for overcoming the *tragedy of the commons*: mutual coercion, mutually agreed upon (34).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

A.T. is supported by the "Deutsche Akademie der Naturforscher Leopold-ina (Grant No. BMBF-LPD 9901/8134). C.H. & M.A.N. are supported by the John Templeton Foundation and the NSF/NIH joint program in mathematical biology (NIH grant R01GM078986). The Program for Evolutionary Dynamics (PED) at HarvardUniversity is sponsored by Jeffrey Epstein.

References and Notes

1. Fehr E, Gächter S. Nature 2002;415:137. [PubMed: 11805825]
2. Fehr E, Fischbacher U. Nature 2003;425:785. [PubMed: 14574401]
3. Hammerstein, P., editor. Genetic and cultural evolution of cooperation. MIT Press; Cambridge MA: 2003.
4. Price ME, Cosmides L, Tooby J. Evol Hum Behav 2002;23:203.
5. Gintis, H.; Bowles, S.; Boyd, R.; Fehr, E., editors. Moral sentiments and material interests: the foundations of cooperation in economic life. MIT Press; Cambridge MA: 2005.
6. de Quervain DJF, et al. Science 2004;305:1254. [PubMed: 15333831]
7. Camerer C, Fehr E. Science 2006;311:47. [PubMed: 16400140]
8. Nakamaru M, Iwasa Y. J Theor Biol 2006;240:475. [PubMed: 16325865]
9. Boyd R, Richerson PJ. Ethology and Sociobiology 1992;13:171.
10. Boyd R, Gintis H, Bowles S, Richerson P. Proc Natl Acad Sci USA 2003;100:3531. [PubMed: 12631700]
11. Henrich J, Boyd R. J theor Biol 2001;208:79. [PubMed: 11162054]
12. Henrich J, et al. Science 2006;312:1767. [PubMed: 16794075]
13. Colman A. Nature 2006;440:744.
14. Fowler JH. Proc Natl Acad Sci USA 2005;102:7047. [PubMed: 15857950]
15. Brandt H, Hauert C, Sigmund K. Proc Natl Acad Sci USA 2006;103:495. [PubMed: 16387857]
16. Nowak MA, Sasaki A, Taylor C, Fudenberg D. Nature 2004;428:646. [PubMed: 15071593]
17. For an analytic treatment we refer to the Supporting Online Material.
18. Fudenberg D, Imhof LA. J Econ Theory 2006;131:251.
19. Hauert C, De Monte S, Hofbauer J, Sigmund K. Science 2002;296:1129. [PubMed: 12004134]

20. Hauert C, De Monte S, Hofbauer J, Sigmund K. J theor Biol 2002;218:187. [PubMed: 12381291]
21. Semmann D, Krambeck HJ, Milinski M. Nature 2003;425:390. [PubMed: 14508487]
22. Complementing interactive online simulations are provided at http://homepage.univie.ac.at/hannelore.brandt/publicgoods/ and the VirtualLabs at http://www.univie.ac.at/virtuallabs.
23. Kiyonari T, Barclay P, Wilson M, Daly M. 2007to appear
24. Milinski M, Semmann D, Krambeck HJ. Nature 2002;415:424. [PubMed: 11807552]
25. Panchanathan K, Boyd R. Nature 2004;432:499. [PubMed: 15565153]
26. Fowler JH. Nature 2005;437:E8. [PubMed: 16177738]
27. Panchanathan K, Boyd R. Nature 2005;437:E8. [PubMed: 16177738]
28. Gürerk O, Irlenbush B, Rockenbach B. Science 2006;312:108. [PubMed: 16601192]
29. Fehr E, Rockenbach B. Nature 2003;422:137. [PubMed: 12634778]
30. Wedekind C, Milinski M. Science 2000;288:850. [PubMed: 10797005]
31. Rockenbach B, Milinski M. Nature 2006;444:718. [PubMed: 17151660]
32. Milinski M, Semmann D, Krambeck HJ, Marotzke M. Proc Natl Acad Sci USA 2006;103:3994. [PubMed: 16537474]
33. Johnson, AW.; Earle, T. The Evolution of Human Societies: from Foraging Group to Agrarian State. Stanford UP, Stanford CA: 1987.
34. Hardin G. Science 1968;162:1243.

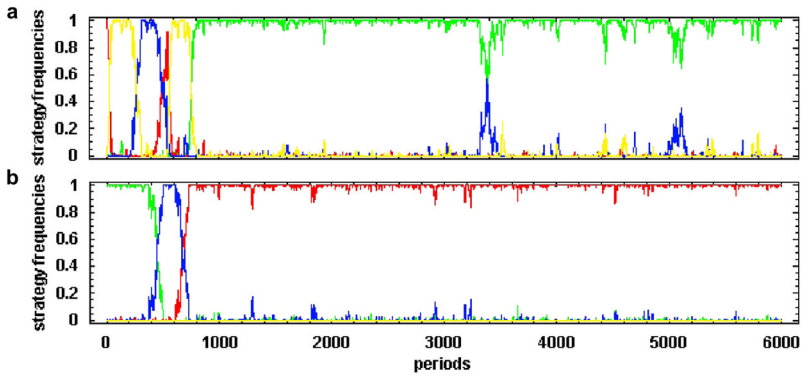

Figure 1.

Punishment and abstaining in joint effort games. (a) Simulations of finite populations consisting of four types of players show that after some initial oscillations, punishers usually dominate the population. In longer runs, their regime can occasionally break down, because cooperators invade by neutral drift, but after another series of oscillations punishers will emerge again. The transient oscillations generally display a rock-paper-scissors-like succession of cooperators, defectors and non-participants. When non-participants are frequent, groups are small, and punishing therefore is less costly, so that punishers have a chance to invade. (b) If participation is compulsory (no non-participants), defectors take over in the long run, even if the population consisted initially of punishers. Parameter values are $M = 100$, $N = 5$, $r = 3$, $\sigma = 1$, $\gamma = 0.3$, $\beta = 1$, $c = 1$, $\mu = 0.01.c$

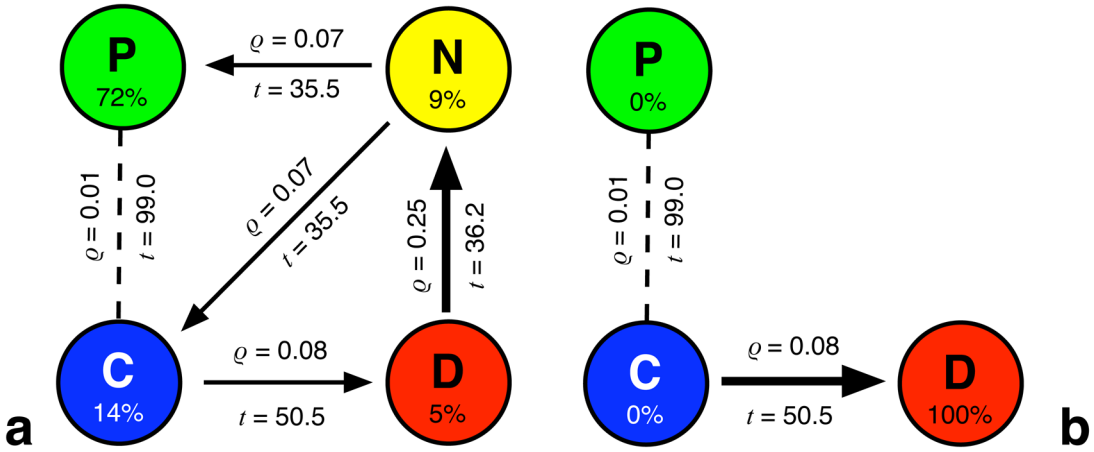

Figure 2.

Stationary probability distributions, transition probabilities and fixation times can be computed analytically for sufficiently small mutation rates, if we assume that players update their strategies according to some specified rule. (Here, we use a Moran process with selection strength $s = 0.249$, see (17)). The dynamics is reduced to transitions between homogeneous population states consisting entirely of cooperators (C), defectors (D), non-participants (N) or punishers (P). The transition probabilities ρ denote the probabilities that a single mutant takes over, the conditional fixation time t indicates the average number of periods required for a single mutant to reach fixation, provided that the mutant takes over. **a** voluntary participation in the joint effort game with punishment, parameter values $N = 5$, $r = 3$, $\sigma = 1$, $\gamma = 0.3$, $\beta = 1$, $c = 1$, $M = 100$. **b** compulsory participation in a joint effort game with punishment, for the same parameter values.

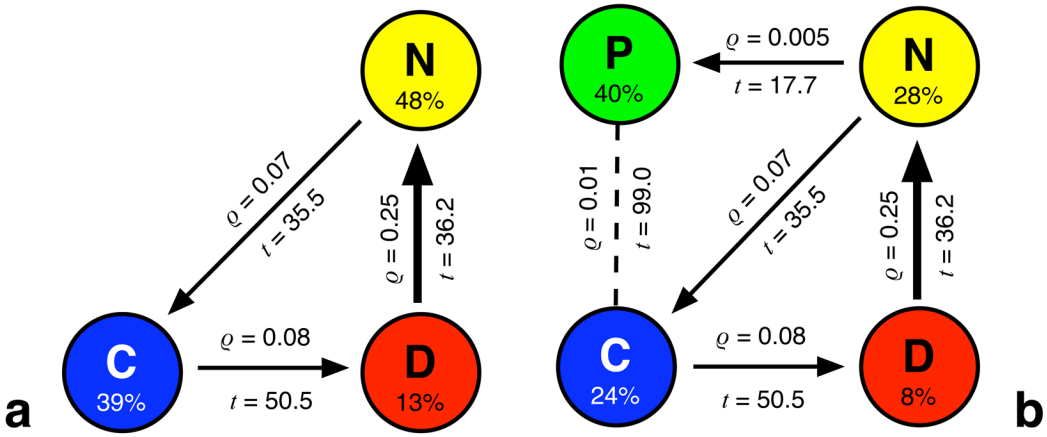

Figure 3.

Punishment is best directed at defectors only. **a** Same as in Fig. 2a, but without punishers. The three remaining strategies supersede each other in a rock-paper-scissors type of cycle. **b** Same as in Fig. 2a, but assuming that punishers equally punish the non-participants. This makes it more difficult for punishers to dominate.