PROCEEDINGS OF THE ROYAL SOCIETY B

CrossMark click for updates

*Proc. R. Soc. B* (2012) **279**, 3716–3721
doi:10.1098/rspb.2012.0937
*Published online* 4 July 2012

# An economic experiment reveals that humans prefer pool punishment to maintain the commons

**Arne Traulsen[1,*], Torsten Röhl[1,2] and Manfred Milinski[2]**

[1]*Evolutionary Theory Group, and* [2]*Department of Evolutionary Ecology, Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Plön, Germany*

Punishment can stabilize costly cooperation and ensure the success of a common project that is threatened by free-riders. Punishment mechanisms can be classified into pool punishment, where the punishment act is carried out by a paid third party, (e.g. a police system or a sheriff), and peer punishment, where the punishment act is carried out by peers. Which punishment mechanism is preferred when both are concurrently available within a society? In an economic experiment, we show that the majority of subjects choose pool punishment, despite being costly even in the absence of defectors, when second-order free-riders, cooperators that do not punish, are also punished. Pool punishers are mutually enforcing their support for the punishment organization, stably trapping each other. Our experimental results show how organized punishment could have displaced individual punishment in human societies.

**Keywords:** evolution of cooperation; peer punishment; pool punishment

## 1. INTRODUCTION

It has been suggested that large-scale cooperation in humans is maintained because wrongdoers are punished [1,2], either by 'peer punishment' [3–12], where individuals decide to punish others in a dyadic way, or by 'pool punishment' [13–18], a kind of tax-paid organization to which punishment is outsourced. Intermediate punishment systems, where only some subjects are allowed to punish, have also been analysed [19,20]. Peer punishment is studied theoretically, experimentally and in naturally occurring environments [21] as a mechanism to stabilize cooperation in public goods games, social dilemmas in which the success of a common project is threatened by the individual temptation to free-ride on the contributions of others [22–26]. When stakes are low, we sometimes use peer punishment by personally reprimanding wrongdoers [27], a rare event in modern societies as we hardly ever observe commuters assaulting fare-dodgers or tax-payers affronting defrauders. In his *Leviathan*, Hobbes [28] suggested that the consent of people could lead to a central authority that punishes those who violate the laws of the society. At present times, these central authorities are institutions such as the police, to which punishment has been outsourced. How can such institutions emerge when they are initially inefficient [24]? In which situation is it better to rely on peer punishment and when does it pay to invest into pool punishment? When both options are available within a society, which one is preferred? We designed a behavioural experiment based on a public goods game to address these questions.

In a typical public goods game, all players can choose whether to cooperate and invest into a common pool or to defect and enjoy the benefit of the public good without investing. The invested sum is then multiplied by a constant factor and distributed among all participants. Since defecting free-riders earn more than cooperators, cooperation typically breaks down [29]. A possibility to overcome this is to give cooperators the option to peer punish and prosecute the free-riders, even if this is costly. However, there are several issues with this approach: in the short run, the efficiency cost due to peer punishment compensates or even overrides the gains from the public good [30–32], so only in the long run can peer punishment become worthwhile [8]. It works only if enough information is available [33] and the fine-to-fee ratio is high enough [34]. Counter punishment [35] or antisocial punishment [36] can lead to additional efficiency loss. Moreover, punishment itself is a second-order public good, and thus threatened by second-order free-riders who contribute to the public good, but do not punish [37–39]. Unless it can be coordinated [40], the initial emergence of a peer punishment system is problematic [41,42]. Peer punishment occurs individually, and after the public-goods interaction it resembles revenge.

Peer punishment is reactive (and may be emotional), whereas pool punishment requires planning. Pool punishers contribute 'taxes' to maintain a punishment system. Building up such a pool punishment system requires investments before free-riding occurs and it is costly to keep up the system even in the absence of wrongdoers [13,43]. It appears that the very nature of pool punishment is that a decision to support an organization which punishes defectors in case they show up has to be made before it is known that defectors are present. In this case, the costs to maintain the pool punishment organization (e.g. a sheriff or the police) must be independent of the later presence of defectors: it may appear to be low when many individuals are punished, but the amount is exactly the same when not a single person has to be punished. This is different from peer punishment,

* Author for correspondence (traulsen@evolbio.mpg.de).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2012.0937 or via http://rspb.royalsocietypublishing.org.

*Received* 24 April 2012
*Accepted* 14 June 2012

Table 1. Overview of the experimental design. In games 1 and 2, subjects gained experience with the two punishment mechanisms in isolation, both without and with second-order punishment. Only the results of game 3 are analysed further. Treatments (a) and (b) are controls.

| | rounds | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | second-order punishment | | | treatment (a) groups | | treatment (b) groups | | treatment (c) groups | |
| game | without | with | initial account | 3 × | 3 × | 3 × | 3 × | 4 × | 4 × |
| 1 | 5 | 5 | € 12 | peer | pool | peer | pool | peer | pool |
| 2 | 5 | 5 | € 12 | pool | peer | pool | peer | pool | peer |
| 3 | 10 | 15 | € 24 | peer | peer | pool | pool | peer and pool | peer and pool |

where each defector is punished in dyadic interactions. Thus, the cost of peer punishment is proportional to the number of people punished, but in contrast to pool punishment there is no cost when no one has to be punished. Peer punishers react to defectors directly, whereas pool punishers plan ahead and establish an organization for punishment. In its simplest form, pool punishment can be implemented by electing and paying an individual to perform the punishment [20]. Instead of allowing subjects to shape their own punishment organization, we implement the consequences of different punishment organizations that subjects can choose from. Putterman *et al.* [44] have tested voting for such formal sanction schemes for the group experimentally. They propose that one should investigate the choice between formal and informal sanctions when both are available. This approach was mathematically modelled by Sigmund *et al.* [16] and herein we test the predictions of that model experimentally.

Our basic assumptions are the same as those in the model of Sigmund *et al.* [16], which compares peer and pool punishment in a public goods game without and with second-order punishment (i.e. the punishment of those who cooperate, but do not punish). In addition to cooperators, defectors and the two forms of punishment, Sigmund *et al.* introduced loners, who abstain from the common enterprise entirely and rely on a small, but secure income [45,46]. The model compares peer and pool punishment alone incorporated into a public goods game as well as the combined availability of both forms of punishment. In summary, the model predicts that (i) the use of peer punishment is not greatly affected by the presence or absence of second-order punishment, which also has no effect on the efficiency (i.e. the average payoff of each individual in each round). (ii) Pool punishment is only used in the presence of second-order punishment, but it substantially decreases efficiency. (iii) If both punishment mechanisms are available, peer punishment is used more frequently in the absence of second-order punishment, but pool punishment prevails in the presence of second-order punishment, again with decreased efficiency. The predictions of this evolutionary model are tested experimentally herein.

## 2. METHODS

We have designed experimental public goods games with volunteers to study how peer and pool punishment alone (treatments (a) and (b)) and their combination (treatment (c)) are used in the absence or presence of second-order punishment. With second-order punishment, all those who cooperate but do not punish have to pay the same fine as the defectors (cf. [3] for a discussion of this assumption). In peer punishment, second-order punishment typically implied additional costs (because additional individuals have to be punished), whereas in pool punishment this was covered by a single tax. Groups within each treatment played three consecutive public goods games; the first and second games were used to familiarize the players with each punishment regime separately, while the third game was used to provide results (table 1). Individuals could make general decisions on whether to punish a certain action, but they did not have an opportunity to target a particular individual.

In treatments (a) and (b), we had six groups of five subjects; in treatment (c) we had eight groups of five subjects (see table 1). The groups remained the same throughout games 1–3. Individual decisions were made in a series of yes or no questions. In each round, the players first had to choose between being a loner (fixed gain of € 0.40) and taking part in a public goods (PG) game. Those subjects deciding for the PG game can contribute either € 0 or € 0.50 to the public pool from their initial endowment of € 12 (€ 24 in game 3; see table 1). The money in the pool was multiplied by 3.1 and redistributed to all PG players. In each of the three games either peer punishment, pool punishment or combined peer and pool punishment was added to the PG game. For peer punishment, the cost for punishing was € 0.50 per punished individual, whereas the cost for being punished was € 1.00 per punisher. In pool punishment, the cost for punishing was € 0.50 and the cost for being punished was € 1.00, as in the theory paper. In all cases, punishment is costly and thus leads to an efficiency loss. As in the mathematical model, the level of efficiency depends on the cost of punishment, which can be chosen as a parameter. Therefore, the experiment could have been designed in such a way that the stable pool punishment solution is also highly efficient, but this would have precluded distinguishing whether subjects prefer the stable to the efficient solution. In peer punishment, the decision to punish (pay € 0.50 per player who did not invest to impose a fine of € 1.00) was made by the individual after they had obtained the information on contributions. In pool punishment, the decision to pay taxes for the punishment organization (pay € 0.50 such that each player who did not invest must pay € 1.00 per tax payer) had to be made before the information on contributions was available. In the experiment, we have called this organization 'police', because the subjects can easily relate this to real life. When both forms of punishment were combined, the pool punishment decision had to be made before the information on contributions was available, while the peer punishment

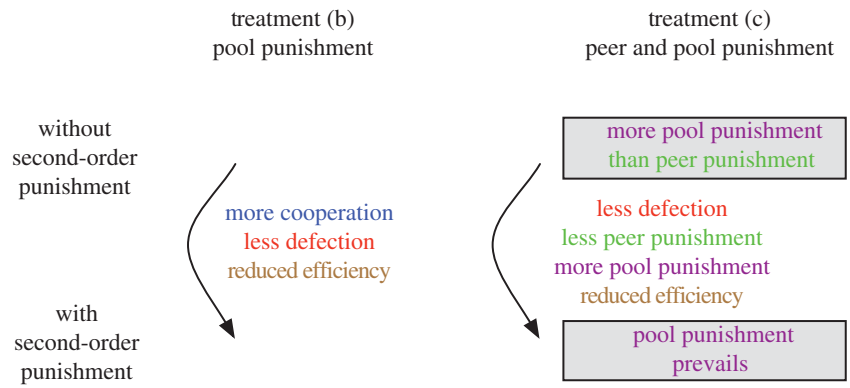


Figure 1. Overview over the relevant significant experimental results. In treatment (b), the introduction of second-order punishment led to a significant increase in the level of cooperation, a significant decrease in defection and a significant reduction in efficiency. In treatment (c), there is significantly more pool punishment than peer punishment in the absence of second-order punishment. If it is introduced, the level of defection and the use of peer punishment significantly decrease. The use of pool punishment significantly increases; the efficiency is reduced significantly. With second-order punishment, pool punishment prevails (i.e. the number of players using pool punishment is significantly different from 50%). Within treatment (a), the changes after the introduction of second-order punishment were not significant. See §3 for details.

decisions thereafter, however without knowledge about pool punishment decisions.

See electronic supplementary material for further details. For data requests, please contact the corresponding author.

## 3. RESULTS

In all treatments of our experiment (table 1), we found that the majority of players cooperate, both with and without second-order punishment. This is also expected from the model. An overview of all significant results is presented in figure 1.

Treatment (a) considered peer punishment incorporated into the public goods game, where the subjects can postpone the punishment decision to the end of each round (see figure 2*a*). We observed no significant differences in the level of cooperation or in the efficiency between the absence and presence of second-order punishment, which corroborates the model's prediction. As in the model, the absence or presence of second-order punishment did not lead to significant differences in the frequency of defectors or loners. We observed a large fraction of cooperators that do not punish and there is little need to punish. The model's prediction is that peer punishment 'prevails', which means that the majority of the population adopts that strategy after some time. This can be analysed by testing whether the frequency of a strategy is above 50 per cent (i.e. whether it prevails). To answer this question, we focused on the last 10 rounds of game 3, when subjects had sufficient time to settle on a strategy. On average, the majority of the subjects used peer punishment in only 27 per cent of the rounds, which is below but not significantly different from a base value of 50 per cent (Wilcoxon one-sample test: $n = 6$ groups; we treat whole groups as statistical units and use two-tailed tests throughout, $Z = -1.444$, $p = 0.1486$). Here, the model's prediction was not supported, but if there is no defector, there is no reason to punish in the experiment see (figure 2*a*). In fact, choosing punishment in such a case does not have any effect.

Treatment (b) considered pool punishment incorporated into the public goods game (see figure 2*b*). In the absence of second-order punishment, the punishment strategy was used rarely. Without second-order punishment, there were no significant differences in the level of cooperation (Mann–Whitney *U*-test: $n_1 = n_2 = 6$, $Z = -0.241$, $p = 0.8095$) or the level of defection (Mann–Whitney *U*-test: $n_1 = n_2 = 6$, $Z = -0.641$, $p = 0.261$) between pool and peer punishment. However, once second-order punishment was added, the majority of players seemed to invest into pool punishment. This significantly increased the level of cooperation (Wilcoxon matched-pairs signed-rank test: $n = 6$, $Z = -1.992$, $p = 0.046$), and escalated the use of pool punishment (Wilcoxon matched-pairs signed-rank test: $n = 6$, $Z = -1.992$, $p = 0.046$; see figure 2*b*). The introduction of second-order punishment did not seem to influence the decisions to act as a loner, but it suppressed the number of defectors (Wilcoxon matched-pairs signed-rank test: $n = 6$, $Z = -2.207$, $p = 0.027$). However, the suppression of defection did not pay out, as second-order punishment substantially reduced the net average payoff in euros (i.e. efficiency) per individual compared with the situation without second-order punishment (Wilcoxon matched-pairs signed-rank test: $n = 6$, $Z = -1.992$, $p = 0.046$). The introduction of second-order punishment led to a loss in efficiency of about a third. Despite the increase in the use of pool punishment, it did not significantly prevail in the last 10 rounds of game 3: on average in 83 per cent of these rounds, the majority of subjects chose pool punishment, but this is not significantly different from the base value of 50 per cent (Wilcoxon one-sample test: $n = 6$, $Z = -1.633$, $p = 0.1025$).

In treatment (c), both forms of punishment were combined. As in the other treatments, the level of cooperation was high, with no significant differences between the absence and the presence of second-order punishment (see figure 2*c*). In the absence of second-order punishment the level of cooperation was significantly higher than in pool punishment alone (Mann–Whitney *U*-test: $n_1 = 6$, $n_2 = 8$, $Z = -2.144$, $p = 0.032$) and the level of defection was lower (Mann–Whitney *U*-test: $n_1 = 6$, $n_2 = 8$, $Z = -2.591$, $p = 0.010$). This effect must have resulted from the interaction between peer and pool punishment. However, the level of pool punishment was still slightly higher than the level of peer punishment

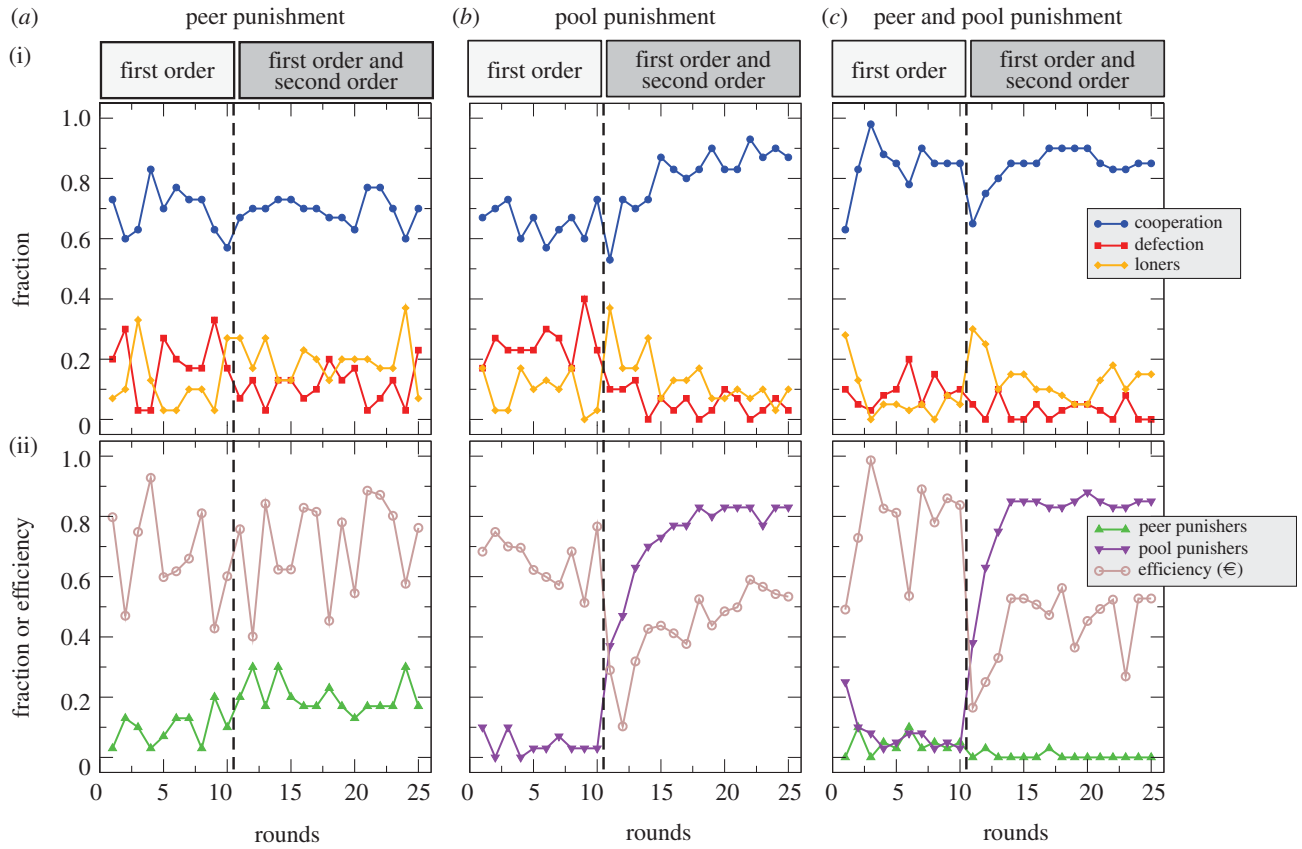


Figure 2. Dynamics of decision-making in game 3, which had 25 rounds. (i) Fraction of decisions to cooperate, defect or act as a loner in the three treatments (see table 1 for details). (ii) Efficiency (i.e. the average payoff in euros) per individual per round, and the fraction of players using punishment (both quantities happened to be of a similar range). The vertical lines mark the introduction of second-order punishment after round 10, which has a large impact in the experiments with pool punishment, where it reduces the efficiency significantly. In treatments (*a*) and (*c*), the average level of cooperation is slightly affected by second-order punishment only, but in treatment (*b*), second-order punishment significantly increases the level of cooperation. In isolation, pool punishment is used much more frequently than peer punishment in the presence of second-order punishment. If we combine both forms of punishment, pool punishment clearly prevails and little peer punishment is used (averages over six groups in peer punishment and pool punishment, eight groups in peer and pool punishment).

(Wilcoxon matched-pairs signed-rank test: $n = 8$, $Z = -2.047$, $p = 0.043$). Once second-order punishment was introduced, the use of pool punishment increased (Wilcoxon matched-pairs signed-rank test: $n = 8$, $Z = -2.521$, $p = 0.012$; see figure 2*c*). The fraction of defectors decreased when second-order punishment was added (Wilcoxon matched-pairs signed-rank test: $n = 8$, $Z = -2.371$, $p = 0.018$). In the presence of two punishment mechanisms, the incentive to cooperate seems to be high: at the time when players decided about cooperation or defection, it was unknown if pool punishers had already committed to punish defectors. But even if no one did, there was still the option of peer punishment later. Again, second-order punishment led to a substantial loss in efficiency compared with the situation without second-order punishment (Wilcoxon matched-pairs signed-rank test: $n = 8$, $Z = -2.521$, $p = 0.012$). The level of peer punishment decreased further to almost zero (Wilcoxon matched-pairs signed-rank test: $n = 8$, $Z = -2.366$, $p = 0.018$), but this was not significantly different from the level in peer punishment alone (figure 2*a*; Mann–Whitney *U*-test: $n_1 = 6$, $n_2 = 8$, $Z = -0.784$, $p = 0.4733$).

In the last 10 rounds of game 3, we did not observe a single instance where at least three players used peer punishment. Pool punishment clearly prevailed: on average, in 87.5 per cent of the last 10 rounds, the majority of subjects chose pool punishment, which is significantly higher than 50 per cent (Wilcoxon one-sample test: $n = 8$, $Z = -2.121$, $p = 0.034$). When second-order punishment was added and pool punishment had been established, it was very difficult to escape contributing to pool punishment. For the group, cooperation without punishment would be a more profitable option, but it is very difficult to achieve.

## 4. DISCUSSION

So far, the vast majority of theoretical and experimental studies on enforcement of cooperation in public goods games has been based on peer punishment. It has little effect on efficiency, because the destructive consequences of punishment occur rarely when the game is repeated for enough rounds—typically the threat of possible punishment suffices. In our case, the maximum average payoff was € 0.79 ± 0.25, which is below the theoretical optimum of € 1.05 occurring when all players cooperate but no one punishes. When only pool punishment was available, the level of cooperation was not significantly different from peer punishment without second-order punishment. Pool punishment was rarely used in this case, as expected. With second-order punishment (i.e. the punishment of

cooperators who do not punish), pool punishment, however, improved the stability of cooperation, but led to a loss of efficiency (i.e. a decrease in the average payoff per individual and round) approximately € $0.51 \pm 0.48$. In fact, the pool punishment system was so costly that it would have been beneficial to abandon it in favour of peer punishment. In this study, every defector faced the same fine under pool punishment due to the absence of individual differences in defection.

In order to test its predictions, we had to follow the mathematical model and implemented a pool punishment mechanism that was less efficient than other more sophisticated approaches. For example, one could punish only the largest deviator in the way to give her/him a precise incentive to cooperate more, as Andreoni & Gee [17] proposed. Another approach is to assume that a small number of punishers is sufficient to achieve an optimal punishment effciency [40,47], which is similar to the volunteer's dilemma [48,49]. When both peer and pool punishment are available within the social group, both punishment options were used at a low level in the absence of second-order punishment, but their interaction significantly enhanced cooperation. In the presence of second-order punishment, pool punishers dominated and prevailed, corroborating with the model's prediction. Since thereafter any other strategy has a lower payoff, pool punishers mutually enforced each other not to deviate, and thus the situation was stable. However, 'efficiency is traded for stability' ([16], p. 861); stable cooperation comes at a price that reflects the fact that taxes for the organizational punishment had to be paid even in the absence of defectors. Similar to the theoretical study that motivated our experiment [16], it turned out that second-order punishment is crucial for the maintenance of pool punishment.

The major result of the corresponding theoretical model [16] that peer and pool punishment can evolve by individual selection alone was supported by our experimental findings: our players have democratically built up a pool punishment organization within their group and have forgone the opportunity to decide individually who is to be punished, as predicted. Pool punishment seemed to be a safe haven, but it came at a significant loss of efficiency. Following Hobbes, the goal of the establishment of a central authority is not to achieve the best for all, but to prevent the worst for all in a stable society.

We are grateful to C. Hauert and B. Rockenbach for helpful discussions and M. Abou Chakra for comments and proofreading. We thank P. M. Altrock, H. Brendelberger and B. Werner for support in performing the experiment, and the students of the University of Kiel for their participation.

## REFERENCES

1 Sigmund, K. 2007 Punish or perish? Retaliation and collaboration among humans. *Trends Ecol. Evol.* **22**, 593–600. (doi:10.1016/j.tree.2007.06.012)
2 Shinada, M. & Yamagishi, T. 2008 Bringing back Leviathan into social dilemmas. In *New issues and paradigms in research on social dilemmas*. (eds A. Biel, D. Eek, T. Garling & M. Gustafson), pp. 93–123. Berlin, Germany: Springer.
3 Yamagishi, T. & Takahashi, N. 1994 Evolution of norms without metanorms. In *Social dilemmas and cooperation* (eds U. Schulz, W. Albers & U. Mueller), pp. 311–326. Berlin, Germany: Springer.
4 Sigmund, K. 2010 *The calculus of selfishness*. Princeton, NJ: Princeton University Press.
5 Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
6 Gurerk, O., Irlenbusch, B. & Rockenbach, B. 2006 The competitive advantage of sanctioning institutions. *Science* **312**, 108–111. (doi:10.1126/science.1123633)
7 Rockenbach, B. & Milinski, M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723. (doi:10.1038/nature05229)
8 Gächter, S., Renner, E. & Sefton, M. 2008 The long-run benefits of punishment. *Science* **322**, 1510. (doi:10.1126/science.1164744)
9 Helbing, D., Szolnoki, A., Perc, M. & Szabo, G. 2010 Punish, but not too hard: how costly punishment spreads in the spatial public goods game. *New J. Phys.* **12**, 083005. (doi:10.1088/1367-2630/12/8/083005)
10 Guala, F. 2012 Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35**, 1–59. (doi:10.1017/S0140525X11000069)
11 Hilbe, C. & Traulsen, A. 2012 Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scient. Rep.* **2**, 458. (doi:10.1038/srep00458)
12 Raihani, N. J., Thornton, A. & Bshary, R. 2012 Punishment and cooperation in nature. *Trends Ecol. Evol.* **27**, 288–295. (doi:10.1016/j.tree.2011.12.004)
13 Yamagishi, T. 1986 The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116. (doi:10.1037/0022-3514.51.1.110)
14 McCusker, C. & Carnevale, P. J. 1995 Framing in resource dilemmas: loss aversion and the moderating effects of sanctions. *Organ. Behav. Hum. Decision Process.* **61**, 190–201. (doi:10.1006/obhd.1995.1015)
15 Rapoport, A. & Au, W. T. 2001 Bonus and penalty in common pool resource dilemmas under uncertainty. *Organ. Behav. Hum. Decision Process.* **85**, 135–165. (doi:10.1006/obhd.2000.2935)
16 Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. 2010 Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863. (doi:10.1038/nature09203)
17 Andreoni, J. & Gee, L. 2011 *The hired gun mechanism*. Working paper 17032. Cambridge, MA: NBER.
18 Perc, M. 2012 Sustainable institutionalized punishment requires elimination of second-order free-riders. *Scient. Rep.* **2**, 344. (doi:10.1038/srep00344)
19 O'Gorman, R., Henrich, J. & Van Vugt, M. 2009 Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proc. R. Soc. B* **276**, 323–329. (doi:10.1098/rspb.2008.1082)
20 Baldassarri, D. & Grossman, G. 2011 Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl Acad. Sci. USA* **108**, 11 023–11 027. (doi:10.1073/pnas.1105456108)
21 Mathew, S. & Boyd, R. 2011 Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl Acad. Sci. USA* **108**, 11 375–11 380. (doi:10.1073/pnas.1105604108)
22 Gordon, H. S. 1954 The economic theory of a common-property resource: the fishery. *J. Polit. Econ.* **62**, 124–142. (doi:10.1086/257497)
23 Hardin, G. 1968 The tragedy of the commons. *Science* **162**, 1243–1248. (doi:10.1126/science.162.3859.1243)
24 Ostrom, E. 1990 *Governing the commons: the evolution of institutions for collective action*. Cambridge, UK: Cambridge University Press.

25 Ledyard, J. O. 1995 Public goods: a survey of experimental research. In *Handbook of experimental economics* (eds J. H. Kagel & A. E. Roth), pp. 111–194. Princeton, NJ: Princeton Univeristy Press.
26 Chaudhuri, A. 2011 Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Exp. Econ.* **14**, 47–83. (doi:10.1007/s10683-010-9257-1)
27 Balafoutas, L. & Nikiforakis, N. 2011 *Norm enforcement in the city: a natural field experiment.* Working paper, University of Melbourne, Melbourne, Australia.
28 Hobbes, T. 1909 *Leviathan*. Oxford, UK: Oxford University Press.
29 Fischbacher, U. & Gaechter, S. 2010 Social preferences, beliefs, and the dynamics of free riding in public goods. *Am. Econ. Rev.* **100**, 541–556. (doi:10.1257/aer.100.1.541)
30 Egas, M. & Riedl, A. 2008 The economics of altruistic punishment and the maintenance of cooperation. *Proc. R. Soc. B* **275**, 871–878. (doi:10.1098/rspb.2007.1558)
31 Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. 2008 Winners don't punish. *Nature* **452**, 348–351. (doi:10.1038/nature06723)
32 Milinski, M. & Rockenbach, B. 2008 Punisher pays. *Nature* **452**, 297. (doi:10.1038/452297a)
33 Bornstein, G. & Weisel, O. 2010 Punishment, cooperation, and cheater detection in 'noisy' social exchange. *Games* **1**, 18–33. (doi:10.3390/g1010018)
34 Nikiforakis, N. & Normann, H. T. 2008 A comparative statics analysis of punishment in public-good experiments. *Exp. Econ.* **11**, 358–369. (doi:10.1007/s10683-007-9171-3)
35 Nikiforakis, N. 2008 Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Pub. Econ.* **92**, 91–112. (doi:10.1016/j.jpubeco.2007.04.008)
36 Herrmann, B., Thöni, C. & Gächter, S. 2008 Antisocial punishment across societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)
37 Milinski, M., Semmann, D. & Krambeck, H. J. 2002 Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426. (doi:10.1038/415424a)
38 Panchanathan, K. & Boyd, R. 2004 Indirect reciprocity can stabilize cooperation without the second-order free-rider problem. *Nature* **432**, 499–502. (doi:10.1038/nature02978)
39 Cinyabuguma, M., Page, T. & Putterman, L. 2006 Can second-order punishment deter perverse punishment? *Exp. Econ.* **9**, 265–279. (doi:10.1007/s10683-006-9127-z)
40 Boyd, R., Gintis, H. & Bowles, S. 2010 Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620. (doi:10.1126/science.1183665)
41 Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A. & Sigmund, K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
42 Mathew, S. & Boyd, R. 2009 When does optional participation allow the evolution of cooperation? *Proc. R. Soc. B* **276**, 1167–1174. (doi:10.1098/rspb.2008.1623)
43 Kosfeld, M., Okada, A. & Riedl, A. 2009 Institution formation in public goods games. *Am. Econ. Rev.* **99**, 1335–1355. (doi:10.1257/aer.99.4.1335)
44 Putterman, L., Tyran, J. R. & Kamei, K. 2011 Public goods and voting on formal sanction schemes. *J. Pub. Econ.* **96**, 1213–1222. (doi:10.1016/j.jpubeco.2011.05.001)
45 Hauert, C., De Monte, S., Hofbauer, J. & Sigmund, K. 2002 Volunteering as Red Queen mechanism for cooperation in public goods games. *Science* **296**, 1129–1132. (doi:10.1126/science.1070582)
46 Semmann, D., Krambeck, H. J. & Milinski, M. 2003 Volunteering leads to rock-paper-scissors dynamics in a public goods game. *Nature* **425**, 390–393. (doi:10.1038/nature01986)
47 Raihani, N. J. & Bshary, R. 2011 The evolution of punishment in *n*-player public goods games: a volunteer's dilemma. *Evolution* **65**, 2725–2728. (doi:10.1111/j.1558-5646.2011.01383.x)
48 Dieckmann, A. 1985 Volunteer's dilemma. *J. Conflict Resol.* **29**, 605–610. (doi:10.1177/0022002785029004003)
49 Archetti, M. 2011 A strategy to increase cooperation in the volunteer's dilemma: reducing vigilance improves alarm calls. *Evolution* **65**, 885–892. (doi:10.1111/j.1558-5646.2010.01176.x)