PROCEEDINGS OF THE ROYAL SOCIETY B

FirstCite® e-publishing

Proc. R. Soc. B
doi:10.1098/rspb.2008.1082
Published online

Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation

Rick O'Gorman[1,*], Joseph Henrich[2] and Mark Van Vugt[3]

[1]*Psychology Group, Sheffield Hallam University, Collegiate Crescent Campus, Sheffield S10 2BP, UK*
[2]*Psychology and Economics Departments, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4*
[3]*Psychology Department, University of Kent, Canterbury CT2 7NP, UK*

Much of human cooperation remains an evolutionary riddle. Unlike other animals, people frequently cooperate with non-relatives in large groups. Evolutionary models of large-scale cooperation require not just incentives for cooperation, but also a credible disincentive for free riding. Various theoretical solutions have been proposed and experimentally explored, including reputation monitoring and diffuse punishment. Here, we empirically examine an alternative theoretical proposal: responsibility for punishment can be borne by one specific individual. This experiment shows that allowing a single individual to punish increases cooperation to the same level as allowing each group member to punish and results in greater group profits. These results suggest a potential key function of leadership in human groups and provides further evidence supporting that humans will readily and knowingly behave altruistically.

Keywords: cooperation; free riding; punishment; altruism; leadership

1. INTRODUCTION

In recent years, there has been a spate of papers providing evidence for various mechanisms to coax cooperation out of groups of individuals (Rockenbach & Milinski 2006; Sigmund 2007). It is to state the obvious that humans can cooperate readily in extraordinary numbers (Smirnov *et al.* 2007) and that this cooperation often provides public goods, despite the risk of free riding (Andreoni 1988; Fehr & Fischbacher 2003). Much of the recent empirical work on the puzzling aspects of human cooperation have focused on testing evolutionary models of diffuse or altruistic punishment (Boyd & Richerson 1992; Henrich & Boyd 2001; Boyd *et al.* 2003), in which many individuals share the burden of punishing non-cooperators (Sober & Wilson 1998; Fehr & Gächter 2002; Fehr *et al.* 2002; Fehr & Fischbacher 2003).

However, since recent work has shown a lack of motivation for costly punishment in some otherwise cooperative societies (Henrich *et al.* 2006)—perhaps because the solutions have not addressed the problem of second-order free riding—and a possible taste for counter-vailing anti-social punishment (Herrmann *et al.* 2008), it seems plausible that different mechanisms may stabilize cooperation in different ways in different populations. We explore a solution to *n*-person co-operation in which a designated individual is responsible for punishment, contrasting with prior research in this area (but see Carpenter 2007). Over the course of human evolution, individuals in groups capable of motivating cooperation would have gained an adaptive advantage. Observed hunter-gatherer groups adopt various mech-anisms to ensure cooperation, and leadership is one such mechanism that both integrates with humanity's primate heritage and offers a mechanism for groups to coordinate activity (Brown 1991; Boehm 1999; Van Vugt 2006). Models in economics (Hirshleifer & Rasmusen 1989) and evolutionary biology (Boyd & Richerson 1992) indicate that evolution can favour a single punisher per social group and that the actions of this one punisher can efficiently galvanize group cooperation. This solution is particularly interesting since it lacks the second-order free rider problem—which has been the central focus of much theoretical effort—and it avoids the problem of uncoordi-nated over punishment.

Our experimental findings confirm that (i) when placed in the sole punisher role, individuals will punish sufficiently to sustain cooperation, (ii) others will respond by increasing cooperative contributions and (iii) a single punisher can sustain levels of cooperation comparable with that maintained by diffuse punishment (Yamagishi 1986; Ostrom *et al.* 1994; Fehr & Gächter 2002) and at more profitable levels, since punishing efforts are not unnecess-arily duplicated. Such findings suggest that in the smaller scale societies that have dominated human evolutionary history (as well as in the smaller groups of contemporary societies), the single punisher solution may have been an important means of maintaining cooperation. In such groups, single punishers may even be a superior mech-anism, compared with diffuse punishment systems.

2. MATERIAL AND METHODS

(a) *Participants*

One hundred and thirty six participants (35% male) who were undergraduate students from the University of Kent at Canterbury were recruited from across the campus by way of

* Author for correspondence (rogorman@alumni.binghamton.edu).

All authors contributed equally to this work.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2008.1082 or via http://journals.royalsociety.org.

Received 4 August 2008
Accepted 5 September 2008

a job advertisement service. Six experimental sessions took place with 20–24 participants per session. The sessions lasted approximately 1 hour and the average earning for participants was UK£5.47. Each monetary unit (MU) earned during the session equated to UK£0.01.

(b) *Design and procedure*

Initially, participants were informed, by way of a projected presentation, of the procedure of the experiment (including that assignment to groups was random and occurred each round, interactions were anonymous, the amount of endowment, how it could be invested, how pay-offs were allocated and how they would be paid), examples of different contribution patterns and the corresponding pay-offs, and the use of the computer software. Participants were not informed at the beginning of the first segment that there were two segments. After the presentation of the instructions, participants were tested on their understanding of the pay-off procedure. All participants showed satisfactory comprehension.

For those in punishment conditions, further instruction was provided prior to the commencement of the second segment while those continuing with a second control condition received a brief refresher. Instructions relating to the making of deductions did not make any suggestion as to how such deductions could be used, or whether they should be used. Participants were simply informed that such deductions would be possible for the second segment and it was explained how to make such deductions, should participants wish to use such a facility. If a participant queried the purpose, then he or she was simply told that it was an option that would be available and it was up to him or her how it could be used.

We used a modified methodology (Fehr & Gächter 2002) of a public goods experiment that had real monetary earnings at stake run on networked PCs using z-Tree software (Fischbacher 2007). All participants completed a two-segment experiment with an initial no-punishing control segment followed by a second segment of either a further control condition (no-punishment), a condition with punishment permitted for all group members (all-punishment), or a condition with only one individual permitted to punish (one-punishment); therefore, all participants acted as their own controls and partook in only one of three conditions. We did not counterbalance as, firstly, Fehr & Gächter (2002) showed that there was no order effect for not punishing versus punishing and, secondly, our focus was on comparison between the two punishment conditions.

In all conditions, participants played the same public goods game: assigned to groups of four, participants were allocated an endowment of 20 MUs, of which they could invest any amount into a group fund and retain the remainder. Each MU invested in the group fund yielded a pay-off of 0.5 MU to each group member, irrespective of who invested. That is, each MU invested in the group fund was doubled and then divided into four equal shares. Thus, participants would always be better off contributing nothing to the group fund as the return was less than the investment. However, if every member invested their full endowment, then each member would earn 40 MUs, a profit of 20.

In each round, groups were randomly formed so that participants never knew with whom they were interacting ('stranger protocol' in the economics literature), thus controlling for reputation and reciprocity effects. All interactions were anonymous. Investment decisions were made simultaneously, after which information was provided on the investments of other group members. In the second-segment punishment conditions, individuals could simultaneously make deductions from each other by paying a fee, drawn from their earnings for that round, up to a fee maximum of 10 MUs per punished member (the deduction was equivalent to three times the fee). For the one-punishment condition, one member per group was randomly selected after each investment phase to make deductions, whereas in the all-punishment condition, all individuals could make deductions. We conducted the public goods game for six rounds in each condition, so that participants played a total of 12 rounds over two segments to avoid one-shot effects and to examine participants' behaviour over a series of games. With each participant acting as his or her own control and with a fixed order, we could compare between conditions.

During the experiment, participants received no information other than that of the contributions made by each of the other group members to the group fund and, in the punishing conditions after punishing occurred, of the level of deductions made from their own account only. Participants were located in a large computer laboratory and were spaced apart such that no one could see another participant's screen. After completing the public goods games, participants completed online and paper questionnaires to assess their attitudinal and emotional responses to the experiment and their interactions in the games, group identity and a number of other measures not reported here.

3. RESULTS

The average contribution made by participants across all sessions and rounds was 8.28 MUs (s.d. = 6.55). For analysis, we used generalized estimating equations (GEEs), available in SPSS v. 15, which uses robust (Huber–White) errors to correct for lack of independence in the data. Because participants interacted with each other within sessions, this represents a conservative approach to analysis (we also performed a non-parametric analysis, which yielded qualitatively similar results; however, GEE allows for more powerful analysis and is what we report here). We present our analysis firstly of the contribution data, then of the profit data and finally of the punishing data.

(a) *Analysis of contributions*

GEEs for the contribution data used a first-order autoregressive working correlation matrix, due to correlations between adjacent rounds' contributions and a normal/identity link. The data from segment 2 of the study concerns our primary hypothesis that the one-punishment condition would increase contributions above the control condition (see figure 1 for mean contributions and 95% confidence intervals for the three conditions over the six rounds). We examined the effects of condition, round and sex on contributions. There is a main effect on segment 2 contributions (Wald $\chi^2_2 = 10.41$, $p = 0.005$). Regression values (derived from the GEE model; see the electronic supplementary material S1) show that both all-punishment and one-punishment differ significantly from the control group (all-punishment versus control: $B = 6.20$, s.e. = 1.32, $p < 0.001$; one-punishment versus control: $B = 5.47$, s.e. = 1.19, $p < 0.001$) while all-punishment and

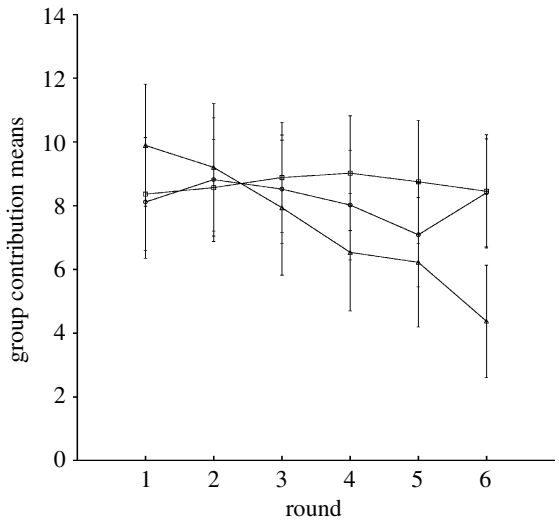

Figure 1. Mean contributions of MUs to the group fund by participants in segment 2 with 95% confidence intervals indicated by error bars. Circles, all-punishment; squares, one-punishment; triangles, no-punishment.

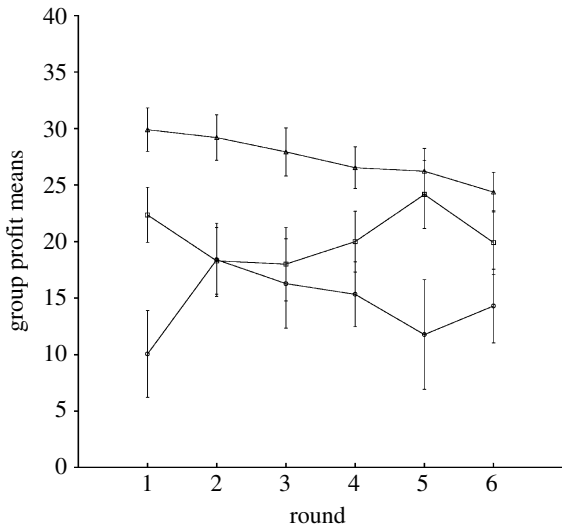

Figure 2. Mean profits (MUs) for participants in segment 2 with 95% confidence intervals indicated by error bars. Circles, all-punishment; squares, one-punishment; triangles, no-punishment.

one-punishment do not appear to significantly differ ($B=-0.73$, s.e.$=1.32$, $p=0.579$; obtained by switching the reference category from control to all-punishment).

Additionally, a main effect for round approaches significance (Wald $\chi^2_5=10.74$, $p=0.057$) and there is an interaction between manipulation and rounds (Wald $\chi^2_{10}=25.29$, $p=0.005$), reflecting the decrease in contributions in the no-punishment condition in contrast to the more stable contributions in the other two conditions (figure 1). Contributions in the control condition decreased significantly across the six rounds in segment 2 (rounds regressed on contributions with robust errors, $B=-1.08$, s.e.$=0.22$, $p<0.001$), whereas contributions in the two punishing conditions remained relatively constant (all-punishment $B=-0.12$, s.e.$=0.21$, $p=0.570$; one-punishment $B=0.03$; s.e.$=0.19$, $p=0.868$). There is no effect for sex, nor is there an interaction ($p>0.370$). Our findings support the hypothesis that, under these conditions, a single individual operating as the sole punisher in a group can improve contributions relative to a control condition without punishment and matches the effect produced by allowing everyone to punish.

We should note that there are differences between conditions in contributions in segment 1 (Wald $\chi^2_2=19.59$, $p<0.001$), possibly due to participants attending with an understanding of the experiment, but these initial differences disappear after the six rounds (round 1: Wald $\chi^2_2=16.72$, $p<0.001$; round 6: Wald $\chi^2_2=1.80$, $p=0.408$). However, this change is not reflected in a significant interaction (Wald $\chi^2_{10}=4.68$, $p=0.912$), though there is a main effect for round (Wald $\chi^2_5=33.09$, $p<0.001$). Finally, there is no difference due to sex of participant (Wald $\chi^2_1=2.91$, $p=0.088$), nor did sex interact with either condition or round ($p>0.711$). The lack of a significant difference between conditions in round 6 of segment 1 suggests that any initial differences in contribution levels between conditions had been eliminated by the end of segment 1, but to control for differences in baseline contribution dispositions, we used participants' average contributions in segment 1 as a covariate in the analysis of segment 2 data.

(b) *Analysis of profits*

Differences in segment 1 profits due to condition and round necessarily follow contribution differences in the same study segment and so, not surprisingly, are significant (condition: Wald $\chi^2_2=25.68$, $p<0.001$; round: Wald $\chi^2_5=33.68$, $p<0.001$), though there is no effect for sex or interactions. As above, for analysis of the profit data from segment 2 (see figure 2 for mean profits and 95% confidence intervals), we use segment 1 contributions as a covariate. There is a main effect for condition (Wald $\chi^2_2=144.79$, $p<0.001$) and an interaction between condition and round (Wald $\chi^2_{10}=42.10$, $p<0.001$), though no main effects for round or sex, nor are there interaction effects. Regression values (as earlier, derived from the GEE model, see the electronic supplementary material S2) show that all three conditions differ (all-punishment versus control: $B=-11.55$, s.e.$=2.03$, $p<0.001$; one-punishment versus control: $B=-5.48$, s.e.$=1.79$, $p=0.002$; all-punishment versus one-punishment: $B=6.07$, s.e.$=2.32$, $p=0.009$; the latter is again obtained by switching the reference category from control to all-punishment).

The lower mean values for the punishment conditions are primarily due to the cost of punishing and deductions, relative to the control condition. However, it is worth noting that, whereas the slopes of the punishment conditions appear stable relative to rounds (rounds regressed on contributions with robust errors, all-punishment $B=0.01$, s.e.$=0.46$, $p=0.980$; one-punishment $B=0.21$; s.e.$=0.34$, $p=0.542$), the control condition's slope is not ($B=-1.08$, s.e.$=0.23$, $p<0.001$), suggesting that both punishment conditions would be likely to be more profitable than the control condition in the long run. Importantly, the one-punishment condition has an advantage over the all-punishment condition due to lower total costs incurred by group members and this is reflected in its higher profit levels (figure 2).

(c) *Analysis of punishment*

Looking at punishing, overall participants in the all-punishment condition punished on 38.9 per cent of

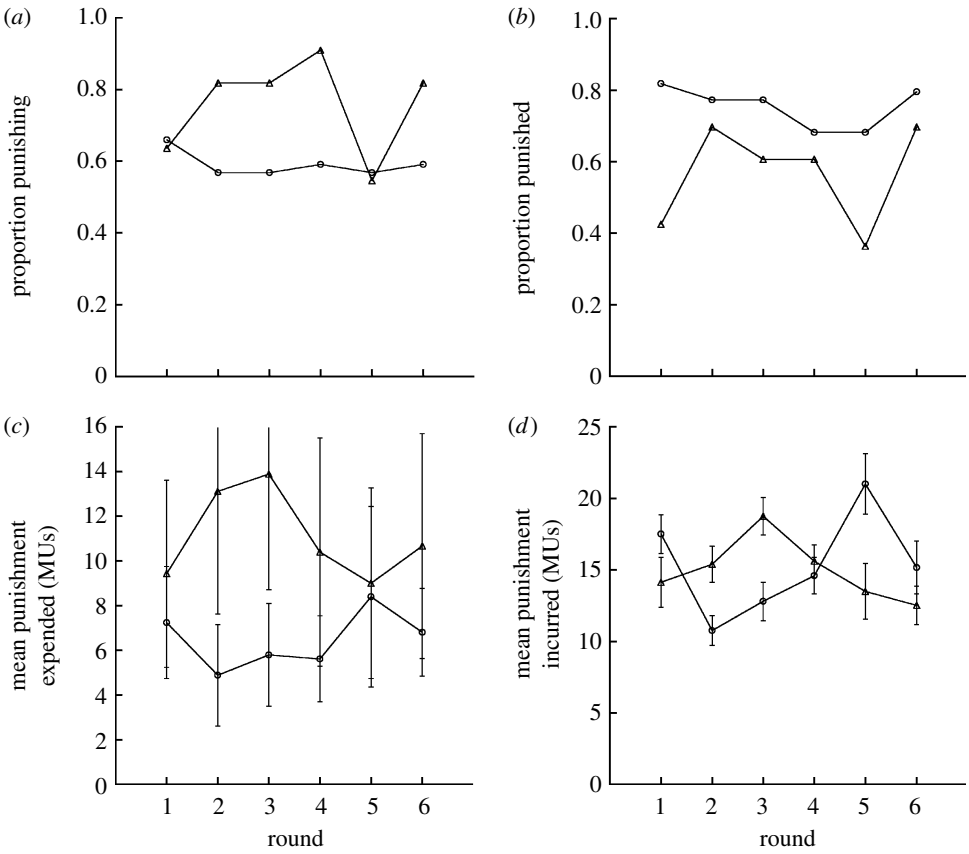

Figure 3. (*a*) There were more punishers in the one-punishment condition who punished at least once per round than in the all-punishment condition, (*b*) although punishers in the one-punishment condition did not punish as many group members. (*c*) One-punishment punishers did, however, expend greater resources to punish, (*d*) resulting in a similar level of penalties being incurred within each punishment condition when considered over the six rounds in segment 2 (circles, all-punishment; triangles, one-punishment).

opportunities to do so whereas punishers in the one-punishment condition did so on 56.6 per cent of opportunities. Per round, the proportion of participants who punished in the one-punishment condition (i.e. punished at least once) was greater than the proportion in the all-punishment condition (figure 3*a*). Fewer participants were punished in the one-punishment condition (figure 3*b*) but that condition's punishers made greater deductions (figure 3*c*), although the total incurred punishments were not consistently harsher in either condition (figure 3*d*).

When we examined possible factors that affected punishment behaviour, we found that punishers in the all-punishment condition and the one-punishment condition appear to be influenced by the same factors. Using a GEE approach (using a gamma/log link, due to the positively skewed data), we examined separately for the two punishment conditions the relationship of punishment levels with the contribution of potential punishers and targets, and rounds. In addition, we also split our analysis between cases where the potential punisher's contribution was less than the target's contribution, versus when it was equal to, or greater than, the target's. Examining the all-punishment condition (see table 1 for GEE regression parameter estimates), when the punisher's contribution was greater, we found that higher punishment is associated with lower contributions by the target ($B=-0.086$, s.e.$=0.01$, $p<0.001$), while higher punisher contributions are also associated with higher punishment ($B=0.039$, s.e.$=0.01$, $p=0.001$). However, when the receiver's contribution was greater, higher punisher contributions are associated with lower punishments. In the one-punishment condition, the pattern was similar irrespective of whether the punisher's or target's contribution was greater. In both cases, we found that higher punishment is associated with lower contributions by the target.

Thus, it appears that while participants in diffuse punishment situations attend to both their own contribution and that of the target, perhaps using their own contributions to guide their decision on whether to punish, those in the solitary punisher condition attend only to the contributions of the target, possibly focused solely on whether contributions are maximally beneficial for the group, in which case any deviation from a full contribution represents an undesirable shortfall. Figure 4 shows that for both punishment conditions, lower target contributions are associated with higher punishment, although this pattern is clearer for those in the all-punishment condition (figure 4*a*). However, it is worth drawing attention to the finding for the all-punishment data that lower punisher contributions are associated with higher punishment when the target has made a higher contribution. This resonates with prior findings which suggest that some participants may be acting spitefully (Herrmann *et al.* 2008). However, such interactions place together such cases where both have contributed very little (one point versus two points), moderate amounts (10 versus 11), or substantial amounts (19 versus 20). We would not expect the behaviour in these

Table 1. GEE parameter estimates for regression of imposed punishment on punisher's and target's contributions and round in study segment 2. (Round 6 was the reference category for the round factor. The results were analysed separately for all-punishment and one-punishment conditions, and for when the punisher's contribution was equal to or greater than the target's contribution and when the target's contribution was greater.)

parameter	*B*	s.e.	Wald χ^2	sig.	*B*	s.e.	Wald χ^2	sig.
	all-punishment/punisher's contribution greater				*all-punishment/target's contribution greater*			
intercept	0.974	0.1865	27.264	0.000	0.898	0.2381	14.211	0.000
punisher's contribution	0.039	0.0119	10.989	0.001	−0.071	0.0265	7.107	0.008
target's contribution	−0.086	0.0129	44.568	0.000	−0.001	0.0120	0.003	0.954
round 1	−0.054	0.1691	0.100	0.751	0.241	0.2055	1.379	0.240
round 2	−0.340	0.1102	9.521	0.002	0.028	0.2515	0.012	0.911
round 3	−0.215	0.1678	1.638	0.201	0.071	0.1553	0.211	0.646
round 4	−0.238	0.1204	3.910	0.048	0.021	0.1873	0.013	0.910
round 5	−0.087	0.1694	0.264	0.607	0.269	0.2068	1.692	0.193
	one-punishment/punisher's contribution greater				*one-punishment/target's contribution greater*			
intercept	0.683	0.1998	11.692	0.001	0.753	0.3831	3.861	0.049
punisher's contribution	−0.003	0.0124	0.063	0.802	0.015	0.0157	0.929	0.335
target's contribution	−0.021	0.0103	4.232	0.040	−0.025	0.0128	3.827	0.050
round 1	0.038	0.2335	0.026	0.872	−0.366	0.3119	1.378	0.240
round 2	0.088	0.2577	0.116	0.733	0.112	0.3676	0.093	0.761
round 3	0.190	0.2586	0.539	0.463	0.087	0.3446	0.064	0.801
round 4	0.175	0.2521	0.483	0.487	−0.095	0.4381	0.047	0.829
round 5	−0.386	0.1725	5.017	0.025	0.071	0.3798	0.035	0.852

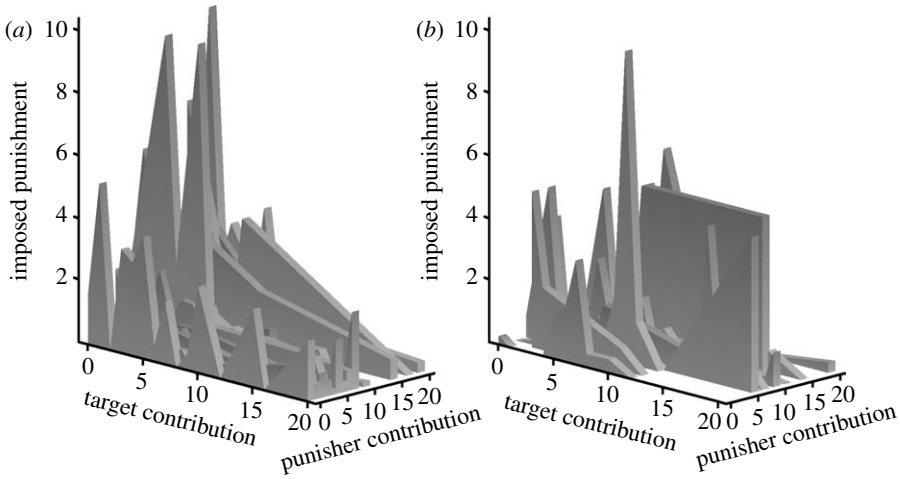

Figure 4. Punishers tended to apply greater deductions for values that deviated more from higher levels of possible contributions, though this effect is stronger in the (*a*) all-punishment than (*b*) one-punishment condition.

three examples to be similar, yet in all cases the target has contributed one point more. Thus, we need further study to understand the strategies being used by participants. Clearly, the strategies are not as straightforward as might be anticipated.

4. DISCUSSION

Individual contributions were significantly higher when punishment was available as an option, with participants responding as effectively to a single individual as to all group members making deductions. Our results suggest that a single punisher successfully enhances and stabilizes group contributions, while doing so more profitably than in the all-punishment condition. As punishment costs are lost to the system, punishments by a single punisher are more coordinated and thus reduce inefficient losses. It is important to note that the success of punishing in this study (in either punishment condition) is facilitated by the 1 : 3 ratio of the cost of punishing for the punisher to the cost for the target. While this is a common ratio in this methodology, studies have shown that lower ratios tend not to produce punishing behaviour sufficient to sustain cooperation (Yamagishi 1986; Burnham & Johnson 2005; Nikiforakis & Normann in press). However, we do not view this as an unnecessary stumbling block. Asymmetrical impacts of punishment can be readily achieved in the real world, for example, through the use of a weapon or social support.

The pattern of punishment for contribution levels suggests that lower contributions tend to incur greater levels of punishment. As Carpenter & Matthews (2008) argue, it appears that punishers are more focused on actual contribution levels rather than deviations from the group (or session) mean, *per se*. However, actual strategies in anonymous games inevitably are likely to be complex,

reflecting the fact that individuals vary in their cooperative intent (Van Lange 1999) and thus how they respond to both being able to 'punish' and being 'punished'. Further in-depth examination of participants' strategies, motives and goals is needed.

Somewhat unexpectedly, more participants in the one-punishment condition punished more often and more harshly than those in the all-punishment condition, incurring greater personal costs. There are a number of possible explanations for this result. The first is that the study protocol with a single designated punisher may have simply enhanced an experimental demand characteristic for participants to punish, with the one-punishment condition having increased compliance due to the reduction of diffusion of responsibility (Latané & Darley 1970), although participants would have had to incur a real cost to comply (in contrast to many studies), and some researchers dispute whether participants do indeed respond to such demands (Berkowitz & Troccoli 1986). The second is that participants may indeed have not felt their actions were anonymous and, cued by the presence of other participants, acted in a manner that they found appropriate to enhance, or at least maintain, their reputation as positive group members (Haley & Fessler 2005; Hagen & Hammerstein 2006). In this case, participants would have to view the incurred real costs as necessary for the (imagined) gain to reputation, and to view punishing and acting altruistically as a reputation-enhancing behaviour rather than signalling aggression or vulnerability to exploitation. Finally, humans may altruistically punish for the benefit of their group (Fehr & Gächter 2002), at least under certain conditions. Although participants were ostensibly anonymous, even if they were not convinced of this state, designated punishers in the one-punishment condition knew that they alone would carry the costs of acting to the benefit of the group (assuming that they saw punishing as such).

These findings may have an important implication for the study of cooperation and the functions of leadership in humans. As noted earlier, large-scale cooperation in human groups (beyond the hunter-gatherer level) represents an evolutionary puzzle. Diffuse punishment does not fully solve this issue owing to the iterated problem of second-order free riders. A system with a single designated punisher can potentially avoid this problem because there is clearer accountability. In human groups, leaders often fulfil the role of designated punishers (Heizer 1978; Krackle 1978; Diamond 1997). Moreover, some form of leadership, even if only ephemeral (Steward 1938; Johnson & Earle 2000), is a human universal and readily emerges in ad hoc laboratory groups (Van Vugt 2006).

Of course, such a leadership role is potentially costly to the individual who occupies it. There is both the energy budget of punishing and the incumbent costs of self-defence by the target or retaliation. Why would individuals take on this role? There may be compensatory benefits for acting as a leader. Some individuals more readily fulfil this role than others based on heritable differences in personality (Hogan *et al.* 1994). In human societies, leaders acquire status and prestige (Van Vugt 2006), which may translate into increased reproductive success (Henrich & Gil-White 2001; Fieder *et al.* 2005). Alternatively, group-level selection could facilitate leadership emergence, either by genetic or cultural mechanisms (Sober & Wilson 1998; Richerson & Boyd 2004). Groups often favour altruists for the leader role (Milinski *et al.* 2002; Hardy & Van Vugt 2006). Competition between rival groups results in selective pressure for its adoption culturally or its evolution, genetically. If all participants can punish each other, such situations risk deteriorating into retaliatory actions that do not just reduce benefits from joint activities but damage the integrity of the group (Denant-Boemont *et al.* 2007; Nikiforakis 2008). A designated punisher avoids these risks.

The issue of anonymity and the consequential inability of punished individuals to retaliate represent a constraint on our argument that we provide evidence for leadership to function as a constraint on free riding. If retaliation were possible, a single punishing individual would be less costly to retaliate against than a set of punishers. However, in this study, we seek only to demonstrate that leadership could fulfil such a function successfully. In reality, a leader is not just one individual but represents the pinnacle of a social structure. Thus, although responsibility may lie with one individual to act, such actions nonetheless, by virtue of the role, carry the support of the group, or at least a majority. Additionally, the actual form of punishment varies substantially, and indeed a leader may not need to be the individual to actually impose the punishment, as caricatured by Mafia films and as is very familiar to anyone working in an organization with punishment capabilities.

In the present study, the random selection of punishers in the one-punishment condition served as a means to impose the role on individuals to control for other confounds. Nonetheless, future studies would do well to attend to more realistic exploration of the role of leaders as punishers. One interesting follow-up would be to examine a series of experimental rounds, allowing participants either to experience different regimes (no punishing, diffuse punishing, single punisher) or gain information on the performance of different regimes, and choose which system to play under. This could further demonstrate the willingness (or not) of individuals to operate under a designated punisher (leader) system. Related research documenting cross-cultural variation in costly punishing (Henrich *et al.* 2006) suggests our findings may be constrained and it would be worthwhile to consider whether punishing through leadership is a cultural universal. The potential impact of retaliation also warrants consideration.

In smaller scale human societies, prestigious leaders can galvanize the trace of larger scale cooperation (Johnson 2003). At least in some circumstances, individuals respond as effectively to a single punishing individual as they do to a more general punitive environment without obvious negative reactions. Consistent with the existing theoretical work (Boyd & Richerson 1992), our research suggests that human psychology may have evolved to recognize situations in which a single motivated leader can enforce cooperation (Van Vugt *et al.* 2008).

This research complied with ethical standards for the treatment of participants and received institutional ethical approval.

We acknowledge the support by the British Academy for this work. We would like to thank anonymous reviewers for their helpful comments.

Author contributions. The authors declare that they have no competing financial interests.

REFERENCES

Andreoni, J. 1988 Why free ride? Strategies and learning in public good experiments. *J. Public Econ.* **37**, 291–304. (doi:10.1016/0047-2727(88)90043-6)

Berkowitz, L. & Troccoli, B. T. 1986 An examination of the assumptions in the demand characteristics thesis: with special reference to the Velten mood induction procedure. *Motiv. Emot.* **10**, 337–349. (doi:10.1007/BF00992108)

Boehm, C. 1999 *Hierarchy in the forest*. Cambridge, MA: Harvard University Press.

Boyd, R. & Richerson, P. J. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)

Brown, D. 1991 *Human universals*. Boston, MA: McGraw-Hill.

Burnham, T. C. & Johnson, D. D. P. 2005 The biological and evolutionary logic of human cooperation. *Anal. Kritik* **27**, 113–135.

Carpenter, J. P. 2007 Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. *Games Econ. Behav.* **60**, 31–51. (doi:10.1016/j.geb.2006.08.011)

Carpenter, J. P. & Matthews, P. H. 2008 What norms trigger punishment. Working paper obtained from http://community.middlebury.edu/~jcarpent/papers.html on 20 June 2008.

Denant-Boemont, L., Masclet, D. & Noussair, C. N. 2007 Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theor.* **33**, 145–167. (doi:10.1007/s00199-007-0212-0)

Diamond, J. 1997 *Guns, germs, and steel: the fates of human societies*. New York, NY: Norton.

Fehr, E. & Fischbacher, U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)

Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)

Fehr, E., Fischbacher, U. & Gächter, S. 2002 Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13**, 1–25. (doi:10.1007/s12110-002-1012-7)

Fieder, M., Huber, S., Bookstein, F., Iber, K., Schäfer, K., Wallner, B. & Winckler, G. 2005 Status and reproduction in humans: new evidence for the validity of evolutionary explanations on basis of a university sample. *Ethology* **111**, 940–950. (doi:10.1111/j.1439-0310.2005.01129.x)

Fischbacher, U. 2007 z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10**, 171–178. (doi:10.1007/s10683-006-9159-4)

Hagen, E. H. & Hammerstein, P. 2006 Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theor. Popul. Biol.* **69**, 339–348. (doi:10.1016/j.tpb.2005.09.005)

Haley, K. J. & Fessler, D. M. T. 2005 Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* **26**, 245–256. (doi:10.1016/j.evolhumbehav.2005.01.002)

Hardy, C. & Van Vugt, M. 2006 Nice guys finish first: the competitive altruism hypothesis. *Pers. Soc. Psychol. Bull.* **32**, 1402–1413. (doi:10.1177/0146167206291006)

Heizer, R. (ed.) 1978 *Handbook of North American Indians: California, 8*. Washington, DC: Smithsonian Institution.

Henrich, J. & Boyd, R. 2001 Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* **208**, 79–89. (doi:10.1006/jtbi.2000.2202)

Henrich, J. & Gil-White, F. J. 2001 The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22**, 165–196. (doi:10.1016/S1090-5138(00)00071-4)

Henrich, J. *et al.* 2006 Costly punishment across human societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)

Herrmann, B., Thöni, C. & Gächter, S. 2008 Antisocial punishment across societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)

Hirshleifer, D. & Rasmusen, E. 1989 Cooperation in a repeated Prisoners' Dilemma with ostracism. *J. Econ. Behav. Organ.* **12**, 87–106. (doi:10.1016/0167-2681(89)90078-4)

Hogan, R., Curphy, G. J. & Hogan, J. 1994 What we know about leadership. *Am. Psychol.* **49**, 493–504. (doi:10.1037/0003-066X.49.6.493)

Johnson, A. 2003 *Families of the forest: Matsigenka Indians of the Peruvian Amazon*. Berkeley, CA: University of California.

Johnson, A. & Earle, T. 2000 *The evolution of human societies*. Stanford, CA: Stanford University Press.

Krackle, W. H. 1978 *Force and persuasion: leadership in an Amazonian Society*. Chicago, IL: University of Chicago Press.

Latané, B. & Darley, J. M. 1970 *The unresponsive bystander: why doesn't he help?* New York, NY: Appleton-Crofts.

Milinski, M., Semmann, D. & Krambeck, H. 2002 Donors to charity gain in both indirect reciprocity and political reputation. *Proc. R. Soc. B* **269**, 881–883. (doi:10.1098/rspb.2002.1964)

Nikiforakis, N. 2008 Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Public Econ.* **92**, 91–112. (doi:10.1016/j.jpubeco.2007.04.008)

Nikiforakis, N. & Normann, H.-T. In press. A comparative statics analysis of punishment in public goods experiments. *Exp. Econ*. (doi:10.1007/s10683-007-9171-3)

Ostrom, E. R., Gardner, R. & Walker, J. M. 1994 *Rules, games, and common-pool resources*. Ann Arbor, MI: University of Michigan Press.

Richerson, P. J. & Boyd, R. 2004 *Not by genes alone: how culture transformed human evolution*. Chicago, IL: University of Chicago Press.

Rockenbach, B. & Milinski, M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723. (doi:10.1038/nature05229)

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)

Smirnov, O., Arrow, H., Kennett, D. & Orbell, J. 2007 Ancestral war and the evolutionary origins of "heroism". *J. Polit.* **69**, 927–940. (doi:10.1111/j.1468-2508.2007.00599.x)

Sober, E. & Wilson, D. S. 1998 *Unto others: the evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Steward, J. 1938 *Basin–plateau aboriginal sociopolitical groups*. Washington, DC: Bureau of American Ethnology.

Van Lange, P. A. M. 1999 The pursuit of joint outcomes and equality in outcomes: an integrative model of social value orientation. *J. Pers. Soc. Psychol.* **77**, 337–349. (doi:10.1037/0022-3514.77.2.337)

Van Vugt, M. 2006 Evolutionary origins of leadership and followership. *Pers. Soc. Psychol. Rev.* **10**, 354–371. (doi:10.1207/s15327957pspr1004_5)

Van Vugt, M., Hogan, R. & Kaiser, R. 2008 Leadership, followership, and evolution: some lessons from the past. *Am. Psychol.* **63**, 182–196. (doi:10.1037/0003-066X.63.3.182)

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)