High cooperation and welfare despite – and because of – the threat of antisocial punishments and feuds

This is preprint version of https://doi.org/10.1037/xge0001004. Please see the final version for citation purposes

Author Information

DAVID S. GORDON[1], & MIKAEL PUURTINEN[2]

[1] Department of Psychology, University of Chester, Chester, CH1 4BJ, United Kingdom

[2]Centre of Excellence in Biological Interactions, Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, P.O. Box 35, FI-40014, Finland

Correspondence Address:

David S Gordon, Department of Psychology, University of Chester, Chester, CH1 4BJ, United Kingdom

email: david.gordon@chester.ac.uk

Funding source:

Kone Foundation (grant 31-233) and Academy of Finland (grant 258385)

Abstract

Cooperation can be difficult to sustain when there is temptation to free-ride on efforts of others. Punishment can stabilize cooperation, but it is usually costly to both the punisher and the punished. In addition, antisocial use of punishment – punishment of co-operators, counter-punishment and feuds – can reduce overall welfare. The current study investigated if powerful individuals – individuals who can punish more effectively or who are immune from punishment – police the antisocial use of punishment, thus reducing the welfare-harming consequences of punishment. To create ample opportunities for anti-social punishment, our modified Public Goods game implemented fixed groups, fixed participant identifiers, two punishment stages, and full information about participant actions. Participants who were immune or had lower punishment cost punished low contributors more often, and immune participants also punished those who punished co-operators. Intriguingly, we found that whenever all participants could punish each other – regardless of the cost of delivering punishment or asymmetry in the cost – cooperation and net earnings reached very high levels. However, participants who were immune from punishment cooperated at a markedly low level, reducing welfare in the group. The results show that in an environment with repeated interactions, plenty of information, and everyone being accountable, even inefficient punishment can maintain high cooperation and welfare

Keywords: cooperation, punishment, policing, welfare, anti-social punishment

Introduction

The scale of cooperation among unrelated humans presents a puzzle for scientists across disciplines. In many cooperative ventures, individuals face a temptation to free-ride on the efforts of others, collecting the benefits of cooperation without contributing to its costs (Fehr & Fischbacher, 2003). Yet, cooperation is sustained in human societies. One suggested solution to the problem of free-riding is peer punishment directed at those who do not contribute to the public good (Acheson, 1988; Ostrom, Walker, & Gardner, 1992). Experimental evidence shows that non-centralized peer punishment can indeed stabilize cooperation in situations where interactions are not repeated, and that cooperation unravels without a sanctioning mechanism (Fehr & Gächter, 2000; Yamagishi, 1988).

However, despite its positive effect on cooperation, peer punishment often fails to improve participant and overall welfare. In experiments, punishment is typically implemented as a monetary cost, with the punisher paying a fee to deliver the punishment. Research suggests that for punishment to stabilize cooperation, it has to be relatively cheap for the punisher (Burns & Visser, 2006; Egas & Riedl, 2008; Nikiforakis & Normann, 2008; Reuben & Riedl, 2013).Yet, due to the costs of delivering and receiving punishment, net earnings at the group level are often less or equal to situations where punishment is not available (Chaudhuri, 2011; Grimalda, Pondorfer, & Tracer, 2016). Furthermore, in experiments where the identity of punishers is concealed, co-operators are often punished in 'blind revenge' and this also undermines cooperation (Chaudhuri, 2011; Sylwester, Herrmann, & Bryson, 2013).

When the experimental set-up allows for direct retaliation, punishment frequently leads to counter-punishments and feuds, and consequently to reduced cooperation and welfare (e.g. Nikiforakis & Engelmann, 2011; Nikiforakis, Noussair, & Wilkening, 2012). Fear of retaliation may be why direct punishment of non-cooperation does not occur outside the laboratory (Balafoutas, Nikiforakis, & Rockenbach, 2014, 2016; Berger & Hevenstone, 2016; Hill, Barton, & Hurtado, 2009; Palmstierna,

Frangou, Wallette, & Dunbar, 2017; Sigmund, 2007; Tarling & Morris, 2010) to the extent seen in many laboratory experiments (Guala, 2012). Further evidence for fear of retaliation hindering punishment comes from experiments showing that participants pay to hide their punishment behaviour (Rockenbach & Milinski, 2011) and, in dyadic interactions, punish only when the partner can be avoided in the future (Bone, Wallace, Bshary, & Raihani, 2015). The threat of retaliation thus makes punishment costly even if the act of punishment itself is cheap (Dreber & Rand, 2012; Masclet, 2003; Masclet, Noussair, Tucker, & Villeval, 2003).

One suggested solution to the problem of anti-social use of punishment is to concentrate 'punishment power' into a single participant. Theoretical studies and experiments have shown that punishment can promote cooperation and welfare if punishment is available (or cheap) for only one (O'Gorman, Henrich, & Van Vugt, 2009; Przepiorka & Diekmann, 2013) or a few individuals (de Weerd & Verbrugge, 2011; Frank, 1996; Gross, Méder, Okamoto-Barth, & Riedl, 2016; Nikiforakis, Normann, & Wallace, 2009). Such asymmetries in punishment power are something we should expect to occur naturally outside of a lab setting; whether due to personal formidability, social status or utility, or the extent of social or kin alliances (for an overview, see Phillips, 2018). Asymmetries in punishment power can allow a participant to expend fewer resources in order to punish and can also reduce – or remove – the threat of retaliatory punishments (Clutton-Brock & Parker, 1995; Gordon & Lea, 2016; Singh & Boomsma, 2015). However, typically in experiments the powerful individual is immune from punishment because others are not able to punish at all; the effect of asymmetry in immunity from punishment *per se* has not been studied experimentally. Importantly, having one participant immune from punishment preserves the ability for other group members to punish each other, and to shoulder the range of associated costs.

While immunity from punishment resolves the problem of retaliation for the 'powerful', the question remains whether the powerful will use their position in a prosocial or self-serving manner. In experiments, participants who punish non-cooperation also tend to cooperate at a high level (e.g.

Barclay, 2006), and participants placed in a position of relative strength within experiments tend to punish non-cooperation while also acting cooperatively themselves (Diekmann & Przepiorka, 2016; Gross et al., 2016; O'Gorman et al., 2009). Immunity from punishment may also promote pro-social 'policing' – punishing those who punish co-operators or retaliate against 'deserved' punishment. Yet, studies using symmetric groups have either found no evidence of policing, or policing was limited to situations where direct relation was not allowed (e.g. Cinyabuguma, Page, & Putterman, 2006; Denant-Boemont, Masclet, & Noussair, 2007; Kamei & Putterman, 2015). Asymmetry in punishment power might thus allow for policing by removing, or restricting, the threat of retaliation by others.

On the other hand, theoretical studies suggest that freedom from punishment might lead to exploitative or coercive behaviour (Dasgupta, 2011; Eldakar, Kammeyer, Nagabandi, & Gallup, 2018; for an experimental example, see Leibbrandt & López-Pérez, 2011). This is more in line with the non-human animal and social psychological literature, where powerful individuals tend to us their relative freedom from punishment to act selfishly (e.g. Clutton-Brock & Parker, 1995; Piff, Stancato, Côté, Mendoza-Denton, & Keltner, 2012). Thus, whether asymmetry in immunity from punishment leads to prosocial or selfish behaviour remains an open question.

To recap, previous literature has shown that peer punishment often fails to promote cooperation and welfare when there is a threat of counter-punishments and feuds. We set out to investigate if powerful individuals – individuals who can punish more effectively or who are immune from punishment – police the antisocial use of punishment. Such policing could lessen the threat of retaliations in the goup, and lead to higher welfare for all. On the other hand, powerful individuals could also use their position selfisly, jeoperdizing cooperation and overall welfare. We designed an experiment where we manipulated immunity from punishment and cost of punishment, and the within-group symmetry/asymmetry in these traits (see Table 1 for the experimental treatments) to investigate the behavioral and welfare consequences of asymmetries in immunity and punishment costs.

To make the threat of retaliation and feuds salient, participants interacted in fixed groups with fixed identities, had complete information on the behaviour of other group members, and could punish and counter-punish all other group members (with the exception of the control treatment with no punishment, and the 'immune' member in one particular treatment). Participants could thus use punishment as they wished: to punish free riding, to counter-punish (possibly across game rounds; feuding), or to punish others for their (antisocial) punishment activity (*i.e.* to police). Thus, the use of punishment was more flexible and, importantly for our purposes, the motive more identifiable than in many other experiments (see, Denant-Boemont et al., 2007; Guala, 2012; Sylwester et al., 2013).

Methods

Participants

Two hundred participants were recruited to the experiments through the paid-participant recruitment database at the University of Jyväskylä (107 females). Mean age of participants was 26. Fourteen experimental sessions were conducted with 16 or 12 participants in each session. Mean duration of a session was 70 minutes. The mean payment received by participants was €17.50.

General procedure

In experimental sessions, participants were each seated in a visually isolated experimental cubicle that contained a computer terminal and written instructions. The instructions covered the entire game structure. The instructions were also given verbally and participants were asked to raise their hand if there was anything they did not understand. Before starting, participants had to answer a series of questions regarding the game mechanics; the study did not begin until all participants had answered correctly. To avoid any end-round effect, participants were not told how many rounds

would be played. Following the session, participants completed a post-game survey (see Supplementary material). The total points earned by each participant were converted to Euros at an exchange rate of 40 points to 1€. Participants were paid in private following completion of the session. All study material was presented in Finnish. The experiment was programmed with zTree (Fischbacher, 2007).

Experimental treatments

There were five treatments in the experiment (see Table 1). Each participant participated in one treatment only. All treatments consisted of 15 game rounds. In all treatments, participants were randomly assigned a permanent participant identifier of either A, B, C or D. The groups and in-game identities were fixed for the whole experiment. Participants in treatments with asymmetric roles were aware of the differences in the opportunities between members. All decisions and outcomes were visible to group members.

Each game round started with a contribution stage, where participants were first allocated 20 points. Next, each participant decided how many of those points (0-20) they contribute to a group project. The total amount of points contributed to the project was then doubled and divided equally amongst the four group members, i.e. a return of 0.5 points for each point contributed. After this, contributions and earnings of all group members were shown to all participants. In the 'No Punishment' treatment, this concluded a game round. In treatments with punishment, participants next entered the first punishment stage.

Table 1. The five experimental treatments. The 'No Punishment' treatment functions as a control to treatments with punishment. In symmetric treatments, all participants have the same options available, whereas in the asymmetric treatments there is one 'High Power' (HP) and three 'Low

Power' (LP) participants in each group. The cost of delivering punishment varied according to the treatment and participant role, but the impact on the target was always the same (see text for further details).

Treatment		Cost:impact of punishment (in points)	Can individual(s) be punished?
No Punishment (NP)		-	no
Symmetric and Free (SF)		0:1	yes
Symmetric and Costly (SC)		1:1	yes
Asymmetry in Cost (AC)	HP	0:1	yes
	LP	1:1	yes
Asymmetry in Immunity (AI)	HP	0:1	no
	LP	0:1	yes

Number of groups per condition: NP=9; SF=10; SC=8; AC=11; AI=12

In the first punishment stage, participants were presented with the contributions and current round earnings for each group member, and were given 20 'deduction tokens'. Deduction tokens had no value in themselves (i.e. they could not be converted to points and Euros), and unused deduction tokens were not carried forward to future stages or rounds. Participants were told they could freely assign anything from zero to 20 deduction tokens (in total) to other group members (excluding the 'High Power' participant in the 'Asymmetry in Immunity' treatment). An assigned deduction token always removed one point from the target, but the cost of assigning a deduction token varied according to the treatment and role of the participant (see Table 1). Once all participants had made

their decisions regarding the use of deduction tokens, they were told how many deduction tokens they had received (if any) from others in their group.

Participants then entered the second punishment stage. Participants saw the full punishment activity in the first punishment stage (i.e. who assigned deduction tokens to whom, and how many), as well as each participant's contribution and current earnings from the round. Participants could again assign an overall maximum of 20 deduction tokens to group members at the same cost as in the first punishment stage. After decisions were made, participants were informed how many deduction tokens they received from each of the other group members at the second punishment stage.

Participants then saw full data for the round: each group members' contribution, earnings, sanctions assigned and received at both punishment stages. Participants were then shown a final screen that broke down their own total for that round (contribution, deduction tokens allocated and deduction tokens received), and displayed their total earnings for the session. As with other studies using multiple punishment stages (e.g. Kamei & Putterman, 2015), participants were aware that finishing the round with negative points would result in zero being added to their overall score. However, in only a single instance did a participant finish a round with negative points. The final screen also reminded participants that they would stay in the same group with the same identifier.

After the final round, participants were directed to a brief questionnaire regarding their behaviour in the session. The two open-ended questions were "Briefly describe your contributions to the group project. Why did you contribute as you did?" and "Briefly explain why you gave (or did not give) deduction tokens to other players?". The results are reported in Supplementary Information.

Punishment mechanism and rationale

In laboratory experiments, the punishment mechanism typically involves a direct cost to punishment. In the current study, we used a cost-free mechanism in a number of treatments. Often, the act of punishment itself is not costly (e.g. verbal or even physical discipline does not necessarily require much resources), but (the threat of) retaliation and other repercussions can make punishment costly (Dreber & Rand, 2012; Gordon & Lea, 2016; Masclet, 2003; Masclet et al., 2003; Nikiforakis & Engelmann, 2011). Thus, the only in-game cost to punishment in the treatments where punishment was cost-free was the response it provoked in the target and other group members. Further, in our experiment a participant could not remove the threat of retaliation by depriving the target of funds by excessive punishment. This is analogous to feuding in human societies, where the threat of retaliation from kin and social allies exists, even if the initial target is incapacitated (e.g. Dunbar, Clark, & Hurst, 1995; Palmstierna et al., 2017). In treatments where punishment was not cost-free, we deliberately implemented a high 1:1 cost:impact ratio to facilitate interpretation of results. In previous studies, such high costs of punishment have not been conducive to cooperation (e.g. Egas & Riedl, 2008), so this treatment was included to investigate whether our game environment would produce similar results. In the 'Asymmetry in Cost' treatment, varying the cost:impact ratio (0:1 for the 'High Power' and 1:1 for the 'Low Power' participants) allowed us to vary punishment power by giving one member greater punishment effectiveness, while still being vulnerable to punishment.

Statistical Analysis

For analysis of group-level contributions and earnings, we used mean group contributions and earnings per round as observations (i.e. the mean of four participants in each group). The full

longitudinal analyses were conducted with Generalized Estimating Equations (GEE) models; these models allowed us to adjust for repeated observations and are less affected by assumptions of distribution (Tang, He, & Tu, 2012). In the analysis, game rounds were coded from -14 to 0 allow interpretation of treatment parameters in the models and to have tests of treatment effects at the final game round, where behaviours had approximately stabilized (the participants did not know how many rounds they would be playing, so no there was no end-game effects). Analysis of contributions and earnings used the 'No Punishment' treatment as the comparison treatment. Pairwise comparisons of the estimated treatment parameters were carried out at Round 15, adjusting for multiple comparisons with sequential Bonferroni correction (Holm, 1979).

To assess the behaviour of the 'High Power' (HP) and 'Low Power' (LP) participants, additional analyses were conducted on the asymmetrical treatments only. The contributions and earnings of the HP and LP participants were separated out for each treatment (LP data representing a mean of the three participants in that role). The cooperation and earning data was analysed in a GEE model to test for effects of asymmetry type and role in the group.

The definition of each type of punishment is given in Table 2. A bout of punishment was defined as any non-zero allocation of deduction tokens by a participant to another participant. Intensity of punishment was the amount of deduction tokens allocated per bout of punishment. Due to the low number of punishment bouts, we analysed total count of bouts of punishment per group. The analysis of punishments at the group- level was conducted using Kruskal-Wallis test, and with Mann-Whitney U-tests for pair-wise comparisons between individual treatments. Correction for multiple comparisons was applied using a sequential Bonferroni (Holm, 1979).

Comparisons of punishment behaviour between 'High Power' and 'Low Power' participants were carried out using Mann-Whitney U-tests. As we assumed a priori that the form of asymmetry and role within group would affect participant punishment behaviour, uncorrected p-values are reported for pre-planned comparisons between 'High' and 'Low Power' participants within treatments and between participants in the same role in the two asymmetric treatments. All analyses were conducted using SPSS 26.

Behaviour	Definition
Pro-social punishments	
Punishment of non-cooperation	Stage-1 punishment directed at participants who contributed less than the actor
Policing	Stage-2 punishment directed against participants who punished a cooperator at Stage 1, excluding cases where the participant had punished the actor
Anti-social punishments	
Punishment of cooperators	Stage-1 punishment directed at participants who contributed more than the actor
Counter-punishment	Stage-2 punishment directed against a participant who punished the actor at Stage 1.
Feuding	Actor and another participant punished one another reciprocally over at least four punishment opportunities

Punishment was divided into two categories: 'pro-social punishments' – the use of punishment that, in principle, should result in increased group cooperation and 'anti-social punishment' – the use of punishment that in principle should have a negative impact on group cooperation.

Results

Treatment effects on cooperation and earnings

The trends of mean contributions are shown in Figure 1A. Contributions at the end of the game session differed significantly between treatments (Wald $\chi^2 4=54.83$, $p<0.001$). The estimate for the effect of Round was also significant (Wald $\chi^2 1=29.38$, $p<0.001$; B=-0.363, s.e.=0.12, $p<0.002$). There was a significant interaction between Treatment and Round on contribution levels (Wald $\chi^2 4=34.76$, $p<0.001$). As shown in Figure 1A, contributions increased in treatments where punishment was available compared to the No Punishment treatment (SF, B=0.73, s.e.=0.12, $p<0.001$; SC, B=0.69, s.e.=0.14, $p<0.001$; AC, B=0.76, s.e.=0.15, $p<0.001$; AR, B=0.63, s.e.=0.12, $p=0.002$).

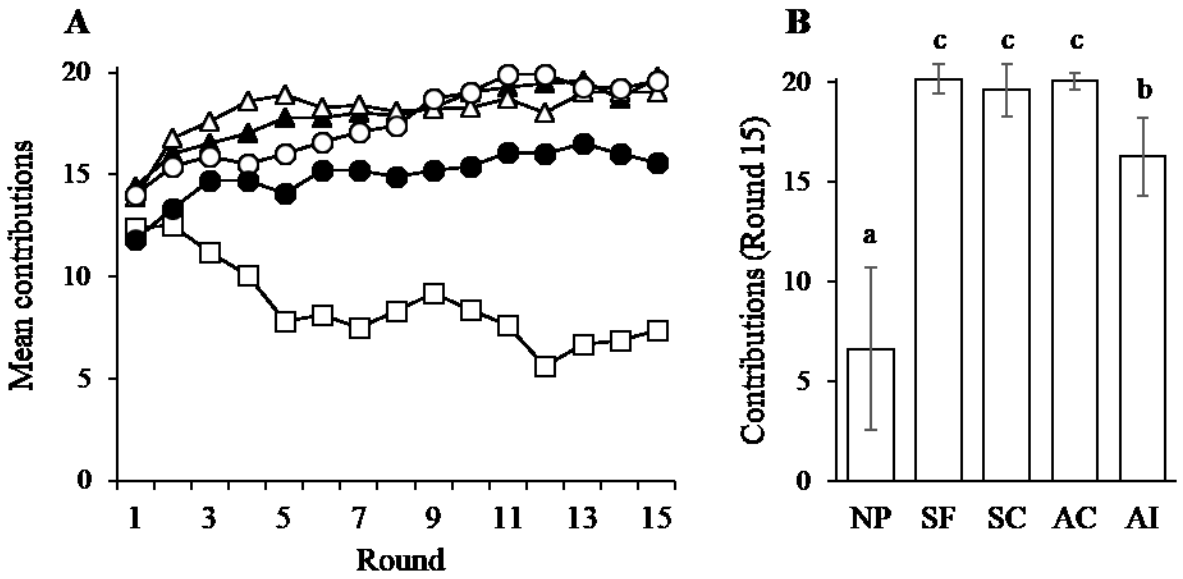

Figure 1. Contributions. a) Mean contributions in each treatment over rounds. □ = 'No Punishment'; ▲ = 'Symmetric and Free'; △ = 'Symmetric and Costly'; ○ = 'Asymmetry in Cost; ● = 'Asymmetry in Immunity'. b) Model estimates for mean contributions at Round 15. Columns that have no letters in common differ from each other significantly after controlling for multiple comparisons ($p<0.05$). Error bar = 95% Wald CI

Pairwise comparisons between estimated treatment -means were done at Round 15 (where behaviours had approximately stabilized), applying sequential Bonferroni correction to adjust significance levels. As shown in Figure 1B, contributions were significantly lower in the 'No Punishment' than in all treatments with punishment. Further, contributions were lower in the 'Asymmetry in Immunity' treatment than in the other punishment treatments.

The trends of mean earning are shown in Figure 2A. Earnings at the end of the game session differed significantly between treatments (Wald $\chi^2 4=66.17$, $p<0.001$). The estimate for the effect of Round was also significant (Wald $\chi^2 1=41.01$, $p<0.001$; B=-0.37, s.e.=0.12, p=0.002). T here was a significant interaction between Treatment and Round on contribution levels (Wald $\chi^2 4=34.76$, $p<0.001$). As shown in Figure 2A, contributions increased in treatments where punishment was available compared to the 'No Punishment' treatment (SF, B=1.25, s.e.=0.22, $p<0.001$; SC, B=0.83, s.e.=0.18, $p<0.001$; AC; B=1.15, s.e.=0.22, $p<0.001$; AR, B=0.88, s.e.=0.19, p=0.002).

Pairwise comparisons between estimated treatment means were again made at Round 15, with a

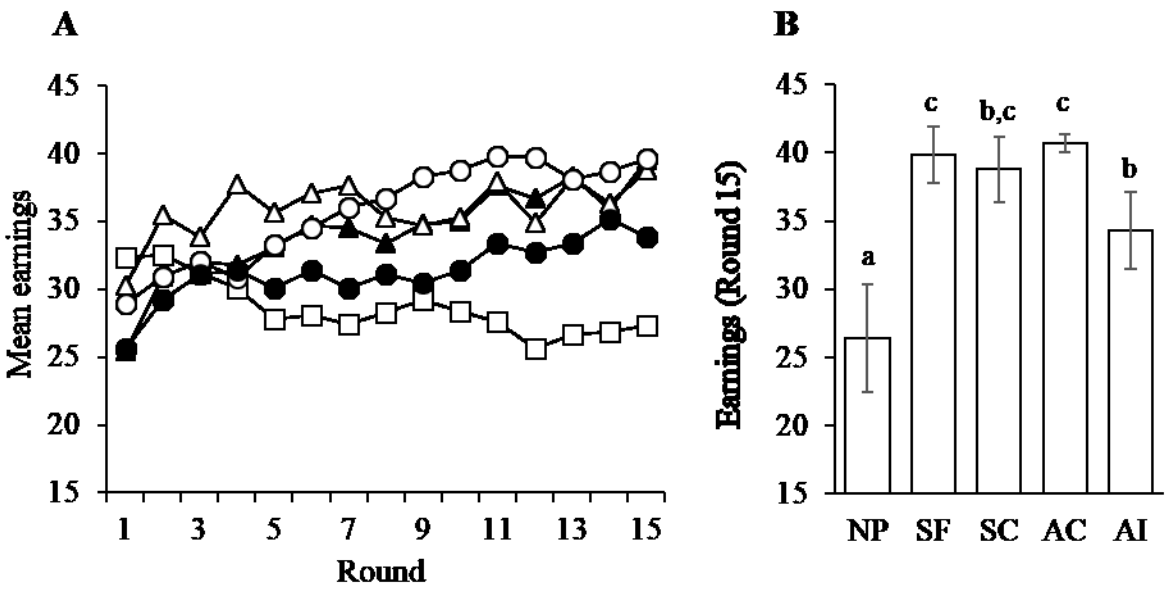

Figure 2. Earnings. A) Mean earnings in each treatment over rounds □ = 'No Punishment' (NP); ▲ = 'Symmetric and Free' (SF); △ = 'Symmetric and Costly' (SC); ○ = 'Asymmetry in Cost (AC); ● = 'Asymmetry in Immunity' (AI). B) Model estimates for mean earnings at Round 15. Columns that have no letters in common differ from each other significantly after controlling for multiple comparisons ($p<0.05$). Bar = 95% Wald CI.

sequential Bonferroni correction applied. As shown in Figure 2B, earnings in the 'No Punishment' treatment were significantly less than in treatments with punishment opportunity. Among punishment treatments, earnings were lower in 'Asymmetry in Immunity' than the other treatments (significant for SF and AC, non-significant for SC).

Effects of 'High' and 'Low Power' roles on cooperation and earnings

The trends of mean contributions of 'High Power' and 'Low Power' participants in the two asymmetric treatments are shown in Figure 3A. Contributions at the end of the game session differed significantly between the participants in different roles/treatments (Wald $\chi^2 3=21.50$, $p<0.001$). The estimate for the effect of Round was also significant (Wald $\chi^2 1=32.27$, $p<0.001$; B=0.001, s.e.=0.12, p=0.99). There was a significant interaction between roles/treatments and Round on contribution levels (Wald $\chi^2 3=8.82$, $p<0.001$). As shown in Figure 3A, compared to the HP participants in AI treatment, contributions of other participants increased over the rounds (LP-AC, B=0.39, s.e.=0.15, p=0.009; HP-AC, B=0.43, s.e.=0.17, p=0.013; LP-AI, B=0.34, s.e.=0. 13, p=0.011).

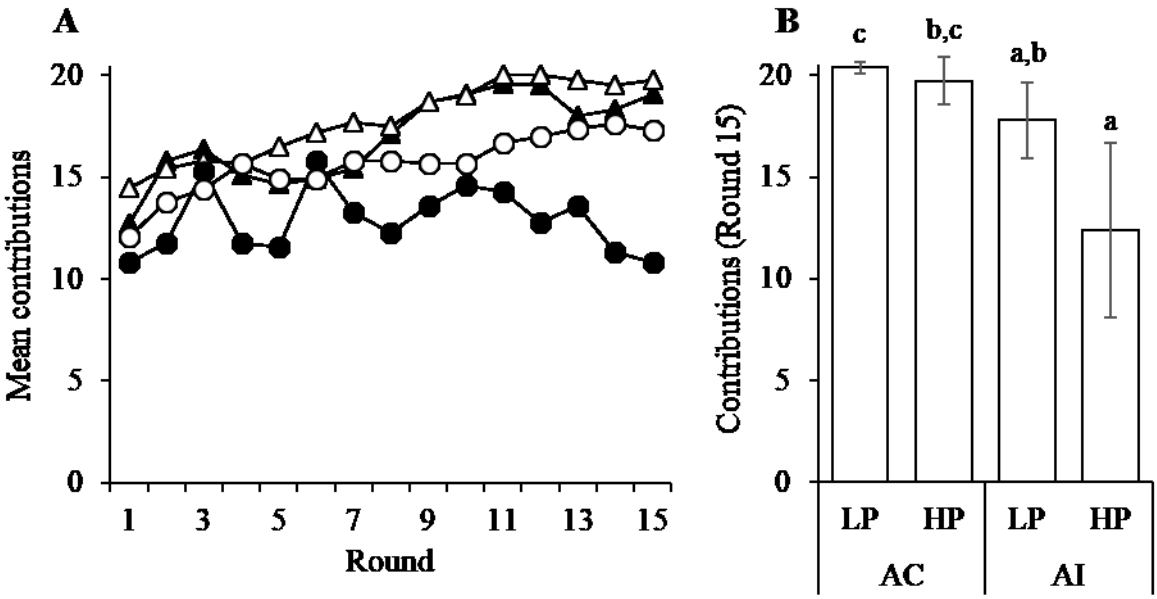

Figure 3. Contributions of the 'High Power' (HP) and 'Low Power' participants in the 'Asymmetry in Cost' (AC) and 'Asymmetry in Immunity' (AI) treatments. a) Mean contributions over rounds. △ = LP-AC; ▲ = HP-AC; ○ = LP-AI; ● = HP-AI. b) Model estimates for mean contributions at Round 15. Columns that have no letters in common differ from each other significantly after controlling for multiple comparisons ($p<0.05$). Bar = 95% Wald CI

Pairwise comparisons of estimated means were made at Round 15, with a sequential Bonferroni correction applied. As shown in Figure 3B, at Round 15 the HP-AI participants contributed significantly less than the HP participants in the AC treatment. The LP individuals in the AI treatment also contributed significantly less than their counterparts in the AC treatment.

The trends of mean earnings are shown in Figure 4A. Earnings at the end of the game session differed significantly between participants in different roles/treatments (Wald $\chi^2 3=20.92$, $p<0.001$). The estimate for the effect of Round was also significant (Wald $\chi^2 1=57.46$, $p<0.001$; $B=0.51$, s.e.=0.11, $p<0.001$). There was no significant interaction between Treatment and Round on earnings levels (Wald $\chi^2 3=2.19$, $p=0.53$).

Pairwise comparisons of estimated means were made at Round 15, with a sequential Bonferroni correction applied. As shown in Figure 4B, at Round 15, in the AI treatment the LP participants earned significantly less than the HP participants, also less than both HP and LP participants in AC treatment.

Treatment effects on punishment behaviour

Punishment of non-cooperation

Table 3 provides an overview of punishment behaviour in the different treatments (see Supplementary Information (A) for trends in punishment behaviour). The number of bouts of

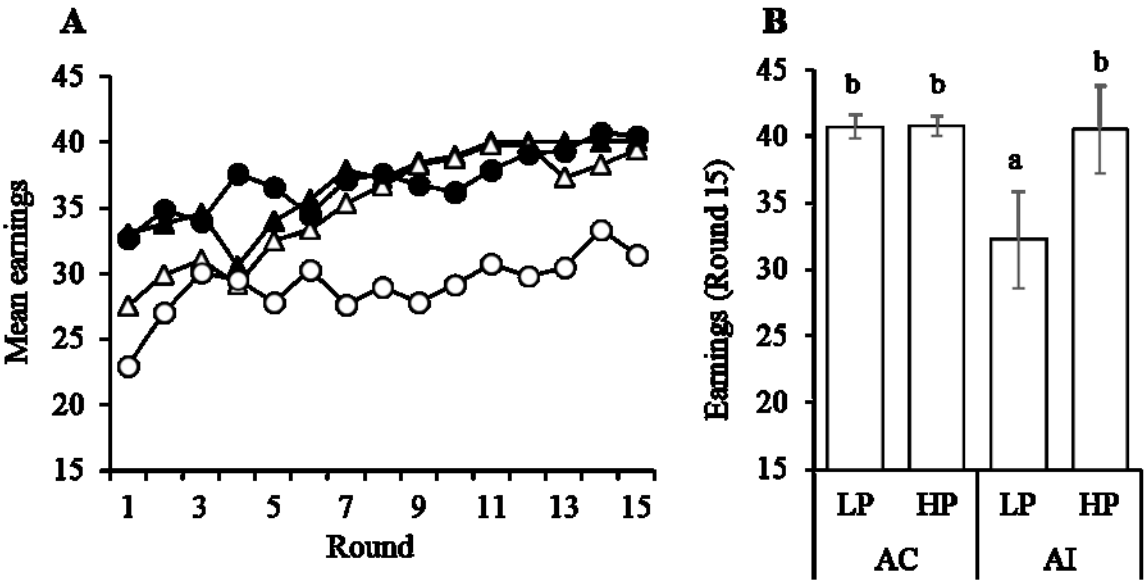

Figure 4. Earnings of the 'High Power' (HP) and 'Low Power' participants in the 'Asymmetry in Cost' (AC) and 'Asymmetry in Immunity' (AI) treatments. A) Mean contributions over rounds. △ = LP-AC; ▲ = HP-AC; ○ = LP-AI; ● = HP-AI. B) Model estimates for mean earnings at Round 15. Columns that have no letters in common differ from each other significantly after controlling for multiple comparisons ($p<0.001$). Bar = 95% Wald CI.

punishment of non-cooperation (i.e. Stage-1 punishment targeted at a participant who contributed less to the group project than the punisher) did not differ between treatments (K- W H=2.47, df=3, p=0.48; Figure 5A), nor did the amount allocated per punishment bout (K-W H=2.84, df=3, p=0.41; Figure 5E).

Table 4. Punishment across different treatments

Treatment	Total bouts				Amount allocated per bout			
	Punishment of non-cooperation	Punishment of cooperators	Counter-punishment	Policing	Punishment of non-cooperation	Punishment of cooperators	Counter-punishment	Policing
No Punishment	-	-	-	-	-	-	-	-
Symmetric and Free (SF)	107	42	91	2	4.9	5.9	7.9	11.5
Symmetric and Costly (SC)	48	25	23	0	1.9	1.8	2.6	0
Asymmetry in Cost (AC)	84	32	31	1	2.7	4.3	5.9	10

Asymmetry in Immunity (AI)	154	94	84	22	3.9	4.5	7.7	2.9
Overall	393	193	229	25	4.8	4.4	7	8.2

Punishment of cooperation

The number of bouts of punishment of co-operators (i.e. punishment at first punishment stage targeted at a participant who had contributed more than the punisher) did not differ between treatments (K-W H=3.40, df=3, p=0.33; Figure 5B). The amount allocated per bout differed between treatments (K-W H=8.33, df=3, p=0.04; Figure 5F), but after correction for multiple comparisons, there were no significant pairwise differences.

Counter-punishment

The number of bouts of counter-punishment (i.e. punishment at second punishment stage directed at a participant who punished the actor at the first stage) did not differ between treatments (K-W H =7.01, df=3, p=0.07; Figure 5C), neither did the severity of the counter-punishment (K-W H=1.40, df=3, p=0.71; Figure 5G).

Policing

Policing, i.e. second stage punishment directed at a participant who punished a co-operator at the first stage, excluding cases of counter-punishment, occurred very rarely (see Table 3). The number of bouts of policing differed significantly between treatments (K-W H=12.7, df=3, p=0.005; Figure 5D), but the amount spent per bout did not (K-W H=1.2, df=3, p=0.55; Figure 5H).

Effect of 'High' and 'Low Power' roles on punishment behaviour

Punishment of non-cooperation

There was no difference within the 'High Power' and the 'Low Power' participants between the asymmetric treatments in the number of punishment bouts or in allocation per bout of punishment. Within the 'Asymmetry in Cost' treatment, the 'High Power' participants engaged in more bouts of punishment ($U=6.5$, $p=0.02$, Figure 5I) and allocated more per bout ($U=5.0$, $p=0.011$, Figure 5M) than the 'Low Power' participants. Within the 'Asymmetry in Immunity' treatment, the 'High Power' participants did not punish more often than the 'Low Power' participants did, but they allocated more per bout ($U=8.5$, $p=0.003$, Figure 5M).

Punishment of co-operators

Punishment of co-operators did not differ between the 'High Power' and 'Low Power' participants among the asymmetric treatments, or between the roles within treatments (Figure 5J and 5N).

Counter-punishment

The 'Low Power' participants did not differ among the asymmetric treatments either in the number of bouts of counter-punishment or in allocation per bout. As the 'High Power' participants in the 'Asymmetry in Immunity' treatment could not be punished, any comparisons involving these participants were not carried out. Within the 'Asymmetry in Cost' treatment, the 'High Power' and 'Low Power' participants did not differ in the number of bouts of counter-punishment (Figure 5K), but the 'High Power' participants allocated more per bout ($U=4.0$, $p=0.03$; Figure 5O).

Policing

As there was only one instance on policing in the ‘Asymmetry in Cost’ treatment, comparisons between ‘High Power’ and ‘Low Power’ roles is not feasible in this treatment. In the ‘Asymmetry in Immunity’ treatment, the ‘High Power’ participants engaged in more bouts of policing than the ‘Low Power’ participants ($U=39.0$, $p=0.033$; Figure 5O), but did not differ in the allocation per bout.

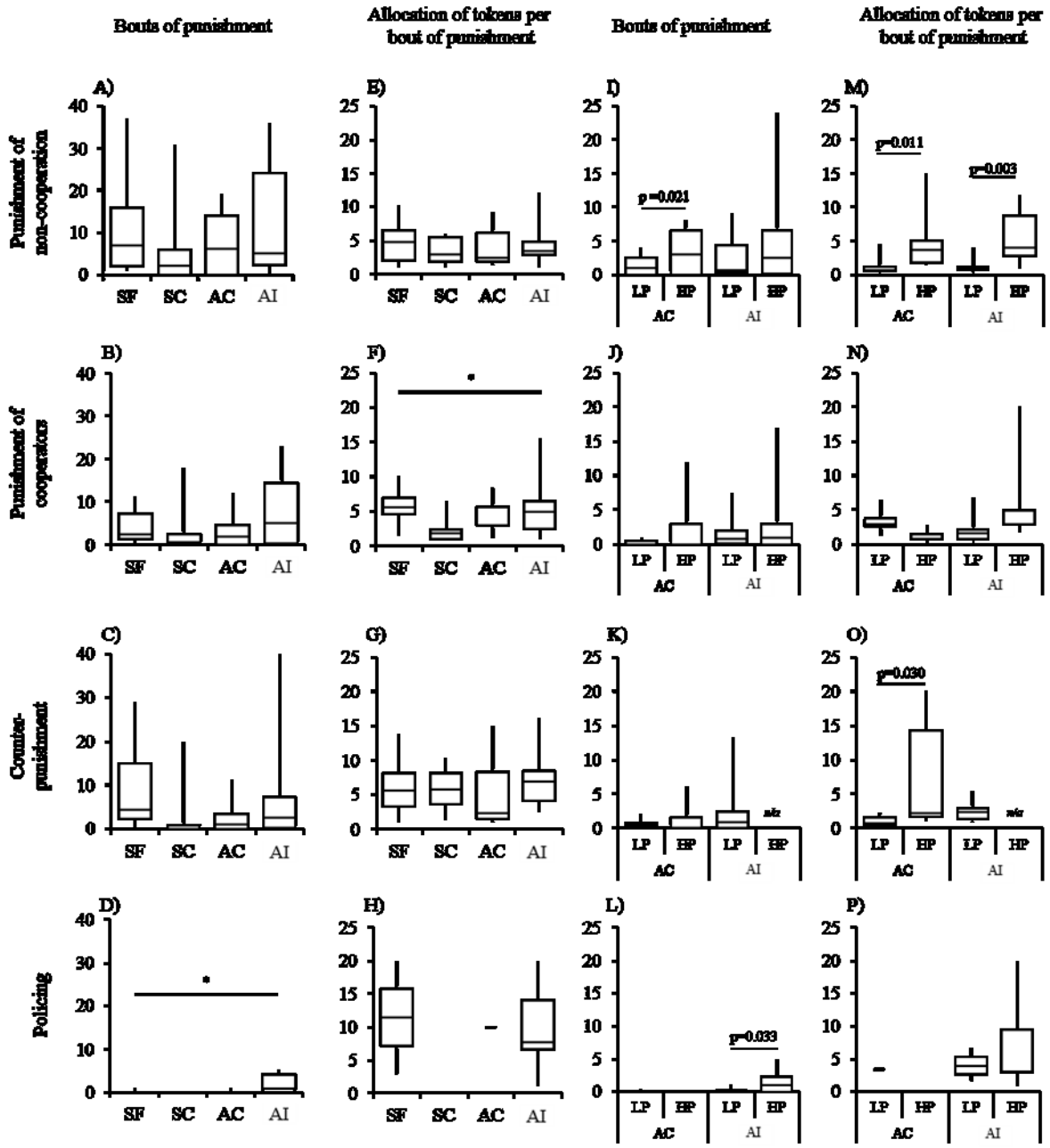

Figure 5. Punishment behaviour across treatments and roles. Panels A-H: Punishment in different treatments. The bar indicates significant overall difference among treatments; *p<0.05, **p<0.01. Panels I-P: Punishment by 'Low Power' (LP) and 'High Power' (HP) participants in the asymmetric treatments. Significant uncorrected p-values shown. From each group, the data from three LP participants was averaged before analysis. Bar = median, boxes = quartiles, poles = range

Feuding

There were 32 feuds (i.e. reciprocal punishments taking place between two actors that lasted for least four consecutive punishment opportunities) in the whole study. Most of the feuds were short, lasting for four punishment bouts (the minimum to be recognized as a feud). The longest feud lasted for 20 punishment bouts. Table 4 shows the distribution of feuds across treatments. While Table 4 suggests that more feuds occurred in the 'Asymmetry in Immunity' and 'Symmetric and Free' treatments, feuds were confined to single groups, and no consistent differences between treatments were detected. Policing of feuds was defined as punishment of feuding participants (meeting the criteria above) by an uninvolved group member, if that punishment could not be explained as either retaliation or punishment for non-cooperation. This differed from policing, as the 'policed' participant must have been involved in a feud. There were no instances where a group-member not involved in a feud punished one or both of the feuding parties.

Table 4. Feuding by treatments

	Number of feuds	Mean allocation per feud	Longest feud (number of reciprocal punishments)
Symmetric and Free (SF)	14	31.3	10
Symmetric and Costly (SC)	1	7.0	4
Asymmetry in Cost (AC)	7	27.6	6
Asymmetry in Immunity (AI)	10	26.6	20

Discussion

The aim of this study was to investigate whether asymmetry in punishment cost and immunity from punishment affect cooperation and welfare in groups where retaliations and feuds are a real threat.

In light of many previous studies, our results were somewhat surprising. We found that whenever all participants could punish each other – regardless of the cost of delivering punishment or asymmetry in the cost – cooperation and net earnings reached very high levels and little actual punishment took place. However, in the treatment where the powerful participant was immune from punishment, the immune participants cooperated at a markedly low level. Looking at punishment behaviour, we found that participants in a powerful position punished non-cooperation only slightly more often than the other participants did, but when they did, they punished more severely. Policing (punishing antisocial use of punishment) was limited to participants who were protected from retaliation by immunity. Overall, the results suggest that (the threat of) peer punishment can maintain cooperation in small groups with ample behavioural information, even with very ineffective punishment (1:1 cost:impact ratio). Although the individuals who were immune from punishment policed anti-social use of punishment to some degree, immunity lead to overall lower cooperation and group welfare.

While previous studies suggest that anti-social punishments are common (Pleasant & Barclay, 2018; Sylwester et al., 2013) and lead to lower cooperation (Chaudhuri, 2011; Cinyabuguma et al., 2006; Dreber & Rand, 2012; Hauser, Nowak, & Rand, 2014; Rand, Armao, Nakamaru, & Ohtsuki, 2010) and to long running feuds (Denant-Boemont et al., 2007; Nikiforakis, 2008; Nikiforakis et al., 2012), we found that cooperation was high and the use of anti-social punishments low in all treatments with punishment opportunities. Indeed, our experiment did not find support for one of the most consistent findings within the experimental literature: that ineffective punishment (high cost:impact ratio) results in lower cooperation (Chaudhuri, 2011; Egas & Riedl, 2008; Kroupa, 2014; Nikiforakis & Normann, 2008).

We suggest that the high level of cooperation and the low level of punishment observed in our experimental treatments are likely due to the same features that make feuding and antisocial use of punishment possible. The visibility of all behaviours and the prospect of long-term repercussions from behaviours deemed reprehensible could have deterred both free-riding and anti-social use of punishment. In studies that that have reported deleterious effects of anti-social punishment, the behaviour is usually hidden from other group members by confining retaliation within the dyad (Nikiforakis et al., 2012) or the round (Cinyabuguma et al., 2006; Denant-Boemont et al., 2007). There is evidence that individuals wish to hide their use of punishment (Rockenbach & Milinski, 2011), and in small-scale society even pro-social punishment is viewed as a necessary evil (Wiessner, 2005). Furthermore, anti-social punishments seem to be more sensitive to costs than pro-social ones (Sylwester et al., 2013), so it is possible the 'cost' of being observed to punish anti-socially was too high for it to occur (which may include feelings of shame or guilt, see Hopfensitz & Reuben, 2009). Our results echo those of Kamei & Putterman (2015), who conclude that full information and availability of higher-order punishment opportunities increases cooperation and efficiency; we find this to be case also with very ineffective (1:1 cost:impact) punishment. As long as others can see and (potentially) react to anti-social use of punishment, it might be much less of a problem to cooperation than is often assumed.

Further, even ineffective punishment was sufficient to buttress cooperation in an environment where punishment by multiple participants over a long timescale was a possibility (see also, Singh & Boomsma, 2015). While surprising given the laboratory studies on the necessity of effective punishment (e.g. Egas & Riedl, 2008), this is similar to small-scale societies where direct physical punishment is rare, but pro-social behaviour is maintained by more subtle means of gossip, threat of losing social ties, and public ridicule (Kroupa, 2014; von Rueden, Gurven, & Kaplan, 2008; Wiessner, 2005). The utility of ineffective punishment is illustrated by the behaviour of the 'High Power'

participants in the asymmetric treatments. In the 'Asymmetry in Cost' treatment, the 'High Power' participants (who could be punished, but at a high cost:impact) showed high levels of cooperation, even though they could retaliate with no monetary cost. Even when power asymmetries exist, powerful participants still wish to avoid retaliation where possible (see, Barclay & Raihani, 2016; Bone et al., 2015, and see SI-B). Thus participants, powerful or not, behaved cooperatively as long as there was some threat of punishment.

The results add to a wider debate on the role of punishment in the evolution of cooperation. Other factors such as partner choice (Barclay & Raihani, 2016) and reputation (Grimalda et al., 2016; Rockenbach & Milinski, 2011; Santos, Rankin, & Wedekind, 2011) have been shown to be equal if not greater drivers of cooperative behaviour than punishment. Therefore the human capacity (and concern) for reputation and coalitions might have played a greater role in group cooperation than the desire to inflict dyadic punishments (see, Boehm, 2012; Fessler & Holbrook, 2013; Gavrilets, 2015; Gavrilets, Duenez-Guzman, & Vose, 2008; Kroupa, 2014; von Rueden, Redhead, O'Gorman, Kaplan, & Gurven, 2019). Nevertheless, punishment behaviour does confer a reputation on the punisher (e.g. Gordon & Lea, 2016; Gordon, Madden, & Lea, 2014; Raihani & Bshary, 2015b), with the type of reputation earned depending on the type of punishment (see Raihani & Bshary, 2015a). As a result, being easily observable may lead to greater use of pro-social punishment and less use of anti-social punishment.

Nevertheless, the proximate importance of punishment to cooperation cannot be denied. In the current study, cooperation unravelled in the 'No Punishment' treatment, despite design features that typically support cooperation: fixed groups, repeated interactions, and fixed participant identities (Fehr & Gächter, 2000; Fudenberg & Tirole, 1991; Grimalda et al., 2016; Van Lange,

Joireman, Parks, & Van Dijk, 2013). Thus, the visibility of non-cooperation alone cannot explain the high levels of cooperation observed in the other treatments. When asked about their motives in the post-experiment survey, 40% of participants in the 'No Punishment' treatment indicated they reduced their cooperation in response to free-riding by others, compared to <6% in the other treatments (see Supplementary Information (B)). Defection as a form punishment is well recognised (e.g. Kroupa, 2014), but it is interesting that the presence of any punishment seemed to nudge the population away from conditional defection.

Finally, while experimental games are an abstraction of real life, previous studies have used designs that restrict opportunities for retaliation by either concealing player identities or reshuffling group memberships (e.g. Cinyabuguma et al., 2006; Denant-Boemont et al., 2007; Nikiforakis et al., 2012). We argue our set-up is more representative of a small-group environment (e.g. at school, at the workplace, in the neighbourhood) in terms of the availability of information on others and the possibility of long-term repercussion for ones actions (see, Guala, 2012; Kroupa, 2014).

Effects of power asymmetry on behaviour

The punishment behaviour of the powerful participants in the asymmetric treatments also yielded some expected and some more surprising results. In the 'Asymmetry in Cost' treatment the 'High Power' participants, who could punish for free, punished non-cooperation more often and more severely than the 'Low Power' participants; a result which is conceptually similar to other studies concerning asymmetry in punishment ability (e.g. Gross et al., 2016; Nikiforakis et al., 2009). This also conforms to field-experiments and anthropological studies, where punishment tends to be carried out by those with greater punishment ability (Diekmann & Przepiorka, 2016; Przepiorka & Diekmann, 2013; von Rueden, Gurven, Kaplan, & Stieglitz, 2014; Wiessner, 2005).

However, in the 'Asymmetry in Immunity' treatment, participants who could punish without the risk of retaliation did not punish non-cooperation more often than group members for whom retaliation was a risk (although when they did punish, they punished more severely)[1]. This may be explained by the fact the punitive sentiment that triggers punishment can also be sensitive to one's own participation (Price, Cosmides, & Tooby, 2002)[2]. Here powerful participants in the 'Asymmetry in Immunity' treatment were less cooperative than their equivalents in the 'Asymmetry in Cost' treatment, so may not have wished to take action against low-contributions in the cooperation stage. Still, the little policing that did occur was primarily conducted by 'High Power' participants in the 'Asymmetry in Immunity' treatment. The immunity from reprisals may explain why we did find some evidence of policing[3] where other studies have not (e.g., Denant-Boemont et al., 2007; Kamei & Putterman, 2015); when direct retaliation is possible it is unwise to involve yourself in the conflicts of others.

[1] It should be pointed out that our definition for punishing non-cooperation – punishing someone who contributed less than the actor – classifies all punishment by the least cooperative participant as anti-social. The low cooperation of 'High Power' participants in the AI treatment may thus complicate identifying the motive of their punishment behaviour. However, even if we would classify all Stage-1 punishment by 'High Power' participants as pro-social (targeted at non-cooperators), there would still be no significant difference in number of bouts of punishment of non-cooperation between the 'High' and 'Low Power' participants in AI treatment ($U=42.5$, $p=0.09$).

[2] In the post-experiment survey, some 'High Power' participants reported that they did not want to punish non-cooperation as they were not cooperating themselves; an aversion to hypocrisy (see Hopfensitz & Reuben, 2009, and Supplementary Information (B)). This also points to reputation concerns when making punishment decisions, as discussed previously in text. One HP participant in AI even claimed they would have deducted points from themselves, were that possible

[3] An alternative explanation may be that, by removing the threat of punishment, it was simply easier for HP participants in AI treatment to take in all the information presented to them. They could thus have noticed the anti-social use of punishment instead of concentrating on possible threats to themselves. While all steps were taken to ensure participants understood the information that was presented to them, participant understanding has been an issue for economic experiments (Burton-Chellew & West, 2013). Still, individuals in a powerful position *do* ignore social threat-cues in the environment (Dietze & Knowles, 2016; Watkins et al., 2010). Thus, even if HP participants in the AI treatment suffered less cognitive load, this can be seen conceptually similar to the experience of power outside of the laboratory.

The low cooperation of the immune 'High Power' participants in the 'Asymmetry in Immunity' treatment highlights the dilemma present in concentrating punishment power in a single entity, as immunity from retaliation allows for 'corrupt' behaviour (Eldakar et al., 2018; Piff et al., 2012; von Rueden & van Vugt, 2015). While the low cooperation of the immune participants was not unexpected, given punishment's traditional role in enforcing cooperation, it does contradict findings from several studies employing a concentration of punishment power that have found 'powerful' individuals to behave pro-socially (Diekmann & Przepiorka, 2016; Gross et al., 2016; O'Gorman et al., 2009). Instead, it lends partial support to recent models of 'corrupting' power (Eldakar et al., 2018; see also, Phillips, 2018), as the immune exploited their position to free-ride on the cooperation of others. However, the behaviour of the immune differed from theoretical predictions of (Eldakar et al., 2018), as they did not enforce others to cooperate fully. It seems that the immune participants hesitated enforcing double standards by punishing non-cooperation of others while not cooperating themselves, and this resulted in overall decrease in cooperation.

Limitations and future directions

We believe our study has produced interesting expected and unexpected results. However, it is not without weaknesses and these should be taken into account when the results are placed within the wider literature. The number of strictly independent observations (i.e. number of groups) was limited, and this limits the power of statistical tests. Limited of statistical power meant we could not to detect possible significant differences between some treatments once corrections for multiple comparisons were applied (For example, Figure 3B). This is especially the case for the treatment-level punishment data (see Figure 5, A-H). Further studies examining punishment effectiveness influences the use pro- or anti-social punishment would be valuable, especially in an in information-rich environment.

Secondly, our experiment environment differs from much of the literature. One of the key results in the current study was that cooperation flourished even when punishment was very ineffective (1:1 cost:impact ratio). We argue this was due to sensitivity to any punishment in an environment with full information and multiple punishment opportunities. As our main aim was to investigate power asymmetries and policing, we did not include systematic investigation of differences between designs of earlier studies (e.g. Egas & Riedl, 2008) and our game environment. However, others – notably Kamei and Putterman (2015) – have studied the effects of information availability and punishment stages more systematically. In line with our results and argumentation, they found that full information and multiple punishment stages resulted in higher contributions and lower use of anti-social punishments. Notably, comparing to Kamei and Putterman (2015), we also included a no-punishment treatment and a 1:1 cost:impact ratio treatment. Our results thus demonstrate that full information (and fixed groups) alone were not sufficient to maintain cooperation, but the threat of even highly ineffective punishment was.

The caveats of the current study present two clear avenues for future research. First, further studies on the conditions that promote pro-social versus and selfish behaviour of 'powerful' individuals would be highly warranted. Second, we have suggested that the full information environment contributed to the lack of anti-social behaviour, and thus to the lack of need for 'policing' in our study. However, information is often imperfect, noisy or costly to obtain (Akçay, Meirowitz, Ramsay, & Levin, 2012; Lee, Iwasa, Dieckmann, & Sigmund, 2019; Nowak & Sigmund, 1990). In an environment with limited or uncertain information, asymmetry in punishment power might exert a greater positive effect on cooperation and welfare, especially when acting on inaccurate information has costs (e.g. retaliation from those incorrectly punished).

Second, studies that examine the concentration of punishment power (including the current study) have not allowed high-power individuals to exploit their position by allowing a greater access to the resources generated by group contributions. For example, would participants be willing to transfer punishment power to others (e.g. Gross et al., 2016), if each unit of power contributed to their own disenfranchisement or potential exploitation? While asymmetries in power could lead to exploitation and lower group welfare, such asymmetries might be necessary for cooperation outside of an information-rich system or small-scale social environment (see Ghachem, 2016; Traulsen, Röhl, & Milinski, 2012).

Conclusion

In sum, we have shown that, provided that all participants in a group can be punished to some degree, cooperation and welfare can be maintained in a small-group environment. While many studies have demonstrated adverse effects of punishment to group welfare (Chaudhuri, 2011), we find that (the threat of) unrestricted punishment results in high cooperation, little anti-social punishment, and high group welfare. Group cooperation and welfare were maintained even with ineffective (1:1 cost:impact) punishment. There was also some evidence of higher-order punishment (policing) by participants immune to retaliation. However, immunity resulted in corrupt behaviour and reduced group welfare. The results suggest that in a small-group environment, where participants are aware of how others behave and can punish without restrictions, we can govern ourselves.

References

Acheson, J. M. (1988). *The lobster gangs of Maine*: University Press of New England.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences, 111*(45), 15924-15927.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature communications, 7*, 13327. doi:10.1038/ncomms13327

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325-344. doi:10.1016/j.evolhumbehav.2006.01.003

Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human prisoner's dilemmas. *Evolution and Human Behavior, 37*(4), 263-271.

Berger, J., & Hevenstone, D. (2016). Norm enforcement in the city revisited: An international field experiment of altruistic punishment, norm maintenance, and broken windows. *Rationality and Society, 28*(3), 299-319. doi:10.1177/1043463116634035

Bone, J. E., Wallace, B., Bshary, R., & Raihani, N. J. (2015). The effect of power asymmetries on cooperation and punishment in a prisoner's dilemma game. *PLoS ONE, 10*(1), e0117183.

Burns, J., & Visser, M. (2006). Bridging the great divide in south africa: Inequality and punishment in the provision of public goods. *rapport nr.: Working Papers in Economics*(219).

Burton-Chellew, M. N., & West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences, 110*(1), 216-221.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics, 14*(1), 47-83.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9*(3), 265-279.

Clutton-Brock, T., & Parker, G. (1995). Punishment in animal societies. *Nature, 373*, 209-216.

Dasgupta, P. (2011). Dark matters: Exploitation as cooperation. *Journal of Theoretical Biology*.

de Weerd, H., & Verbrugge, R. (2011). Evolution of altruistic punishment in heterogeneous populations. *Journal of Theoretical Biology*.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic theory, 33*(1), 145-167.

Diekmann, A., & Przepiorka, W. (2016). "Take One for the Team!" Individual Heterogeneity and the Emergence of Latent Norms in a Volunteer's Dilemma. *Social Forces, 94*(3), 1309-1333.

Dietze, P., & Knowles, E. D. (2016). Social Class and the Motivational Relevance of Other Human Beings: Evidence From Visual Attention. *Psychological Science*. doi:10.1177/0956797616667721

Dreber, A., & Rand, D. G. (2012). Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *The Behavioral and brain sciences, 35*(1), 24.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences, 275*(1637), 871-878. doi:10.1098/rspb.2007.1558

Eldakar, O. T., Kammeyer, J. O., Nagabandi, N., & Gallup, A. C. (2018). Hypocrisy and Corruption: How Disparities in Power Shape the Evolution of Social Control.

Evolutionary Psychology, 16(2), 1474704918756993. doi:10.1177/1474704918756993
Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785-791.
Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *The American Economic Review, 90*(4), 980-994.
Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are People Conditionally Cooperative Evidence from a public goods experiment. *Economics Letters, 71*(3), 397-404.
Frank, S. A. (1996). Policing and group cohesion when resources vary. *Animal Behaviour, 52*, 1163-1169.
Gordon, D. S., & Lea, S. E. G. (2016). Who Punishes? The Status of the Punishers Affects the Perceived Success of, and Indirect Benefits From,"Moralistic" Punishment. *Evolutionary Psychology, 14*(3), 1474704916658042.
Grimalda, G., Pondorfer, A., & Tracer, D. P. (2016). Social image concerns promote cooperation more than altruistic punishment. *Nature communications, 7*, 12288. doi:10.1038/ncomms12288
Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan–Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific reports, 6*.
Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35*(1), 1.
Hill, K., Barton, M., & Hurtado, A. M. (2009). The emergence of human uniqueness: Characters underlying behavioral modernity. *Evolutionary Anthropology: Issues, News, and Reviews, 18*(5), 187-200.
Kamei, K., & Putterman, L. (2015). In broad daylight: fuller information and higher-order punishment opportunities can promote cooperation. *Journal of Economic Behavior & Organization, 120*, 145-159.
Leibbrandt, A., & López-Pérez, R. (2011). The dark side of altruistic third-party punishment. *Journal of Conflict Resolution, 55*(5), 761-784.
Masclet, D. (2003). Ostracism in work teams: a public good experiment. *International Journal of Manpower, 24*(7), 867-887.
Masclet, D., Noussair, C., Tucker, S., & Villeval, M. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review, 93*(1), 366-380.
Nikiforakis, N., & Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization, 78*(3), 319-332.
Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public good experiments. *Experimental Economics, 11*(4), 358-369.
Nikiforakis, N., Normann, H., & Wallace, B. (2009). Asymmetric enforcement of cooperation in a social dilemma. *Southern Economic Journal, 76*(3), 638-659.
Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics, 96*(9), 797-807.
O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences, 276*(1655), 323.
Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review, 86*(2), 404-417
Palmstierna, M., Frangou, A., Wallette, A., & Dunbar, R. (2017). Family counts: deciding when to murder among the Icelandic Vikings. *Evolution and Human Behavior, 38*(2), 175-180.

Phillips, T. (2018). The concepts of asymmetric and symmetric power can help resolve the puzzle of altruistic and cooperative behaviour. *Biological Reviews, 93*(1), 457-468.
Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., & Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences, 109*(11), 4086–4091.
Przepiorka, W., & Diekmann, A. (2013). Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proceedings of the Royal Society B: Biological Sciences, 280*(1759), 2013-2247.
Reuben, E., & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior, 77*(1), 122-137.
Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences*.
Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution, 22*(11), 593-600. doi:10.1016/j.tree.2007.06.012
Singh, M., & Boomsma, J. J. (2015). Policing and punishment across the domains of social evolution. *Oikos, 124*(8), 971-982.
Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics, 6*(3), 167.
Tarling, R., & Morris, K. (2010). Reporting crime to the police. *British Journal of Criminology, 50*(3), 474.
Watkins, C. D., Fraccaro, P. J., Smith, F. G., Vukovic, J., Feinberg, D. R., DeBruine, L. M., & Jones, B. C. (2010). Taller men are less sensitive to cues of dominance in other men. *Behavioral Ecology, 21*(5), 943-947.
Yamagishi, T. (1988). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, 265-271.
Zar, J. (1999). Biostatistical analysis 4th ed. *New Jersey*.

Supplementary information A

Punishment behaviour over the game rounds. Please see Figure 1 for the instance of bouts of punishment over the game period, and Figure 2 for the amount allocated per bout of punishment.

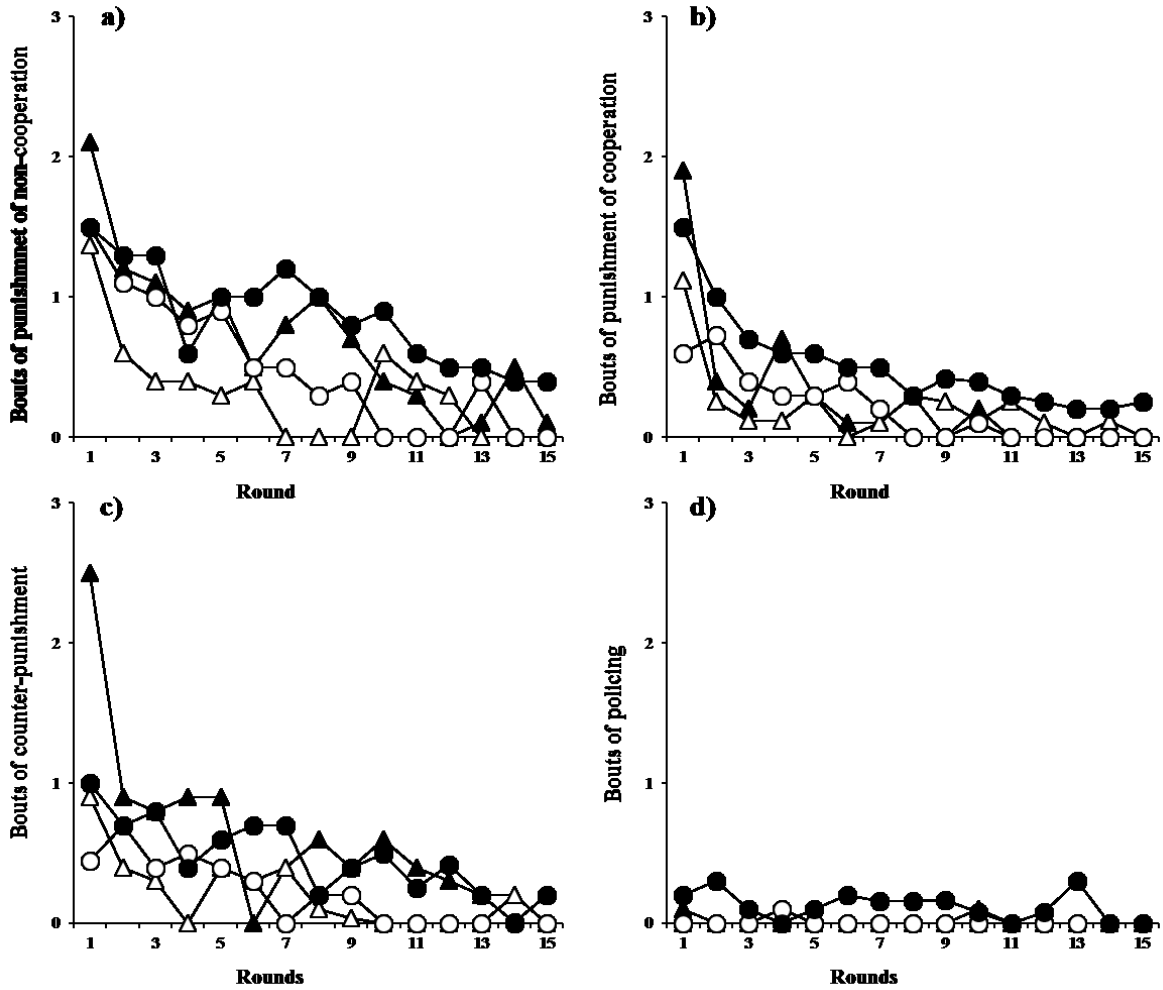

Figure 1: Bouts of punishment by treatment. □ = 'No Punishment'; ▲ = 'Symmetric and Free'; △ = 'Symmetric and Costly'; ○ = 'Asymmetry in Cost; ● = 'Asymmetry in Immunity'. a) Punishment of non-cooperation; b) Punishment of cooperation; c) Counter-punishment, d) Policing

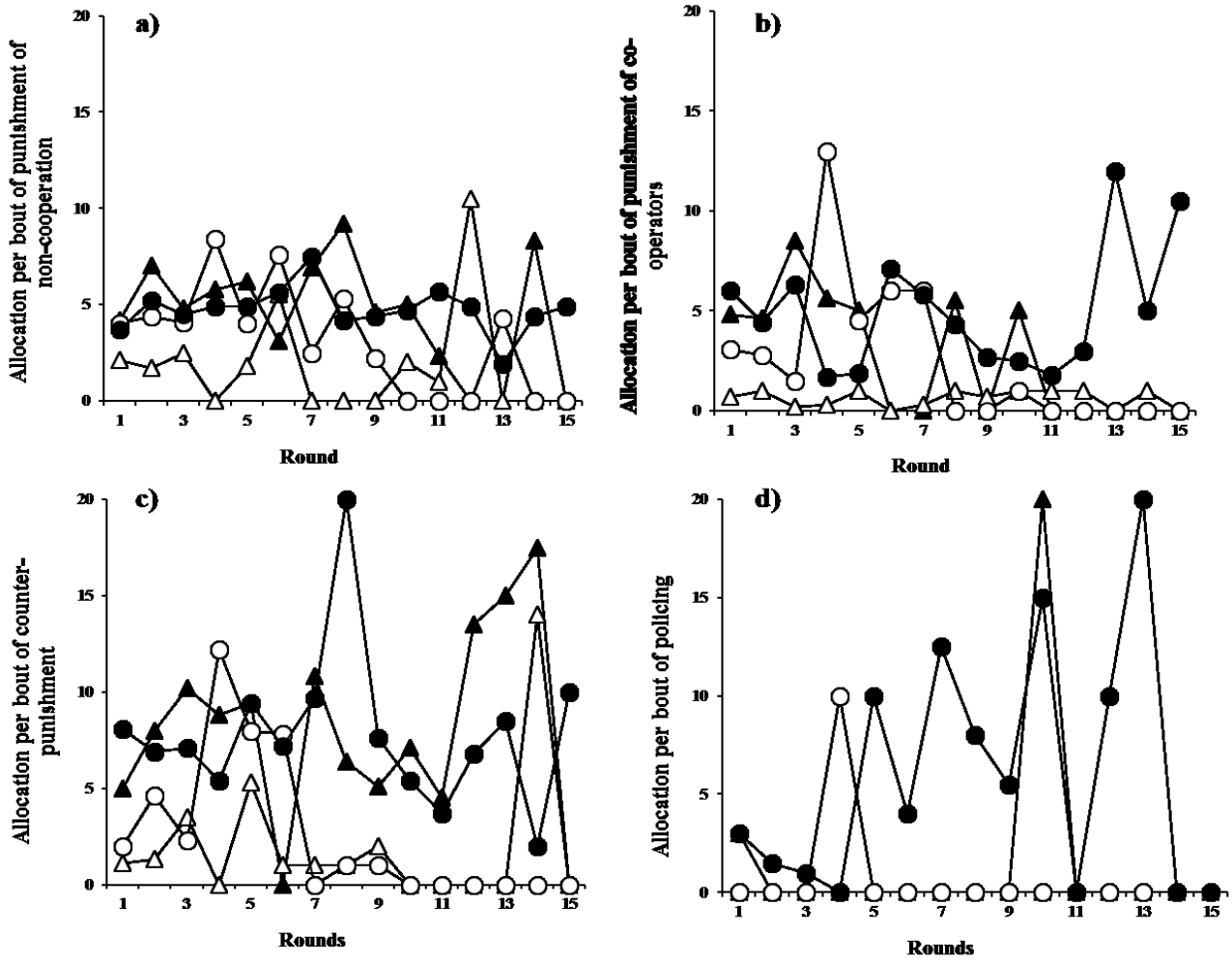

Figure 2: Amount allocated per bout of punishment by treatment. ☐ = 'No Punishment'; ▲ = 'Symmetric and Free'; △ = 'Symmetric and Costly'; ○ = 'Asymmetry in Cost; ● = 'Asymmetry in Immunity'. a) Punishment of non-cooperation; b) Punishment of cooperation; c) Counter-punishment, d) Policing

Supplementary information B

Responses to the question: "Briefly describe your contributions to the group project. Why did you contribute as you did"?

We would like to draw attention to two the more interesting findings from the open-question response. See Table SI-B1 for the coding of all responses. Full transcripts available upon request (in Finnish).

First, participants in all treatments except those in the 'No Punishment' treatment and participants who were immune in the 'Asymmetry in Immunity' treatment indicated that punishment was a factor in their cooperation decisions. This was also true of the 'High Power' participants in the 'Asymmetry in Cost' treatment even though they could retaliate at no cost (see Figure S1). That powerful individuals will still seek to avoid retaliation despite their increased capacity to take revenge has also been documented in dyadic interactions (Barclay & Raihani, 2016; Bone, Wallace, Bshary, & Raihani, 2015). It should also be noted that fear of punishment seemed only marginally related to the cost; individuals in treatments where punishment was free did not dramatically fear it more than participants in treatments where punishment was costly.

Second, participants in the 'No Punishment' treatment indicated they reduced their cooperation in order to avoid being exploited (Figure S2). The phenomenon of conditional cooperation has been well-documented, especially when there is no other way to 'punish' other group members for the latter's non-cooperation (Fischbacher, Gächter, & Fehr, 2001). The presented data emphasises the argument raised in the manuscript that the presence of punishment, even if it was not particular efficient, coupled with reputation concerns nudged the participants away from defection. The reframing caused by *any* punishment is also highlighted by the different attitudes of the 'High Power' participants; half the AC-HP participants believed contributing to be the best solution, but only one AI-HP participant felt the same (Table SI-B1)

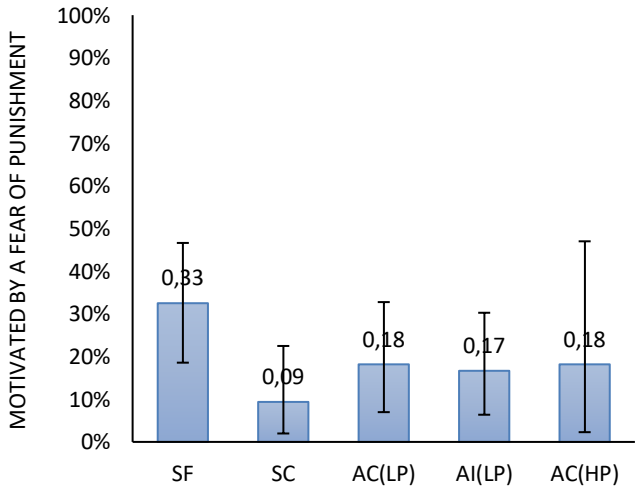

Figure S1: Percentage of participants who indicated their cooperation decisions were motivated by a fear of punishment. Bars =95% CI, calculated as per (Zar, 1999, pp. 527-529)

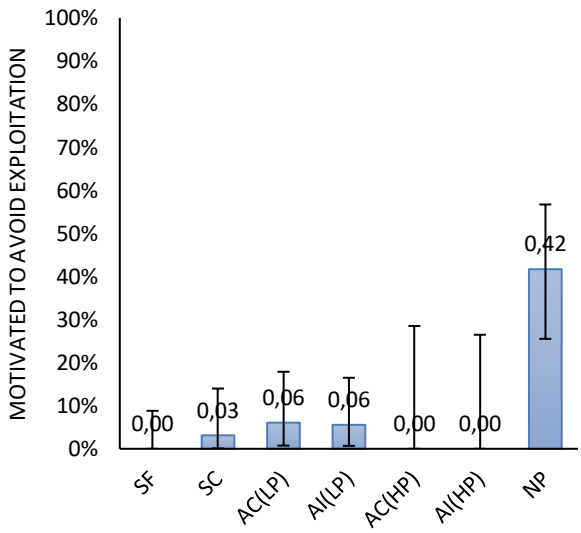

Figure S2: Percentage of participants who indicated their cooperation decisions were motivated by a desire to avoid defection. Bars =95% CI

Responses to the question: "Briefly explain why you gave (or did not give) deduction tokens to other players?"

As shown in Table SI-B2, a much greater variety of reasons were given as to why participants did or did not engage in punishment. Some of the responses relevant to the manuscript have been translated into English (Table SI-B3). Full transcripts (in Finnish) available upon request.

Across treatments participants indicated that they feared retaliation if they were to punishing others (Figure S3) or indicated a more general fear of triggering both retaliation and feuds (Figure S4). These fears were most apparent in the 'Symmetrical and Free' treatment, which makes sense given the zero cost of punishments in this treatment. However, while previous research has suggested that fear of retaliation and feuds leads to both reduced pro-social punishment and to reduced cooperation (e.g. Nikiforakis & Engelmann, 2011), the current study found consistently high cooperation despite the fear of both. That AC-HP individual did not fear retaliation from punishing is evident in their behaviour; punishing non-cooperation more often and more severely compared to their group-mates

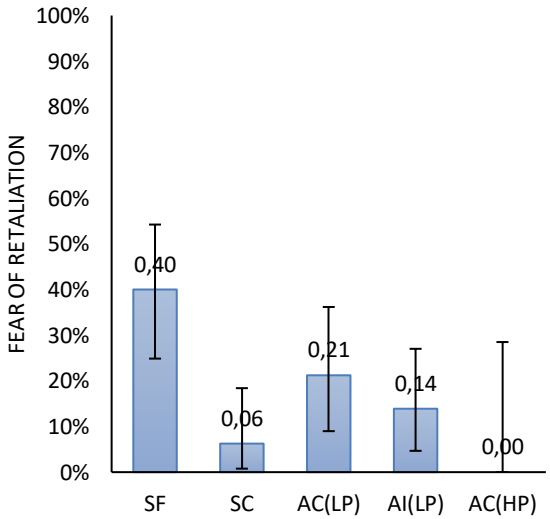

Figure S3: Percentage of participants who indicated their punishment decisions were affected by a fear of retaliation. Bars =95% CI

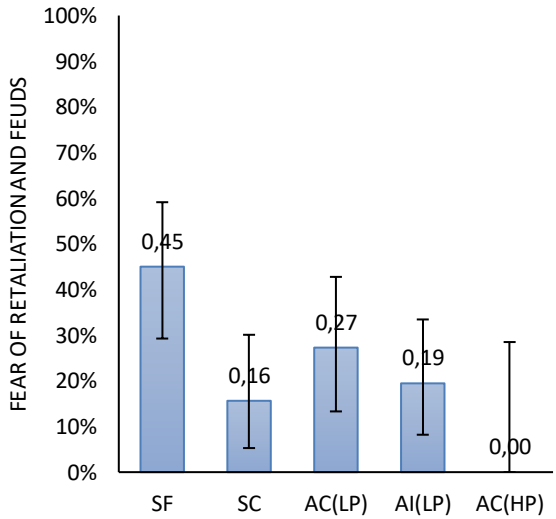

Figure 4: Percentage of participants who indicated their punishment decisions were affected by a fear of triggering retaliation and feuds. Bars =95% CI

Table SI-1: reponses to the question "Briefly describe your contributions to the group project. Why did you contribute as you did"

		Right thing to do (always tried to cooperate)		Interested in high earnings for everyone		Optimum solution ("simply the best way to get most points") Contributed		Optimum solution ("simply the best way to get most points") did not contribute		Fear of punishment		Impunity encouraged free riding		Conditional cooperation: Cooperated to match high cooperation of other		Conditional cooperation: Defected to avoid being exploited		Encourage others to cooperate		Not contributing for fun/variety		Not contributing to teach a lesson to others	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
SF	**40**	1	3%	16	40%	14	35%	6	15%	13	33%	0	0%	11	28%	0	0%	10	25%	3	8%	1	3%
SC	**32**	3	9%	10	31%	9	28%	4	13%	3	9%	0	0%	25	78%	1	3%	5	16%	3	9%	0	0%
AC (LP)	**33**	0	0%	12	36%	9	27%	3	9%	6	18%	0	0%	6	18%	2	6%	9	27%	1	3%	0	0%
AI (LP)	**36**	0	0%	11	31%	6	17%	2	6%	6	17%	0	0%	8	22%	2	6%	8	22%	1	3%	0	0%
AC (HP)	**11**	0	0%	1	9%	6	55%	3	27%	2	18%	0	0%	0	0%	0	0%	1	9%	0	0%	0	0%
AI (HP)	**12**	0	0%	3	25%	1	8%	4	33%	0	0%	4	33%	2	17%	0	0%	3	25%	1	8%	0	0%
NP	**36**	2	6%	10	28%	1	3%	11	31%	0	0%	0	0%	5	14%	15	42%	17	47%	2	6%	3	8%

Participants occasionally gave multiple reasons for their behaviour. Thus the count for each treatment may be greater than the total number of individuals in that treatment

Table SI-B2: reponses to the question "Briefly explain why you gave (or did not give) deduction tokens to other players?"

		No need to punish because everyone cooperated		"right thing to do" : punish anti-social individuals		"Right thing to do" :not punish anyone		Uneasy feeling about punishing others		Desire for eqaltarianism / feeling envy		Encourage cooperation		Desire for revenge		Fear of retaliation		Fear of initiating a feud		Fear of retaliation and/or feud		Optimum solution ("simply the best way to get most points"):punished	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
SF	**40**	2	5%	2	5%	1	3%	2	5%	12	30%	9	23%	11	28%	16	40%	4	10%	18	45%	0	0
SC	**32**	3	9%	1	3%	2	6%	1	3%	4	13%	7	22%	1	3%	2	6%	3	9%	5	16%	0	0
AC (LP)	**33**	5	15%	0	0%	0	0%	2	6%	7	21%	9	27%	2	6%	7	21%	3	9%	9	27%	0	0
AI (LP)	**36**	2	6%	0	0%	0	0%	1	3%	9	25%	6	17%	9	25%	5	14%	2	6%	7	19%	0	0
AC (HP)	**11**	1	9%	0	0%	0	0%	1	9%	4	36%	3	27%	1	9%	0	0%	0	0%	0	0%	0	0
AI (HP)	**12**	0	0%	0	0%	0	0%	2	17%	3	25%	5	42%	0	0%	0	0%	0	0%	0	0%	0	0

		Optimum solution ("simply the best way to get most points"):did not punish		Conditional punishment ("punished because other punished as well")		Conditional punishment ("did not punish because others did not punish either)		I liked to create chaos / test effects		Preemptive punishment to drain others power to punish back		Policing (punishing those who punished others)		Punishing those who did not punish free-riders		Reducing earnings of others		Expected B to punish		Specifically punished b	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
SF	**40**	3	8%	2	5%	3	8%	1	3%	1	3%	1	3%	0	0%	0	0%	0	0%	0	0%
SC	**32**	13	41%	0	0%	4	13%	0	0%	0	0%	1	3%	1	3%	0	0%	0	0%	0	0%
AC (LP)	**33**	10	30%	0	0%	1	3%	0	0%	0	0%	0	0%	0	0%	0	0%	3	9%	3	9%
AI (LP)	**36**	5	14%	0	0%	0	0%	1	3%	0	0%	0	0%	0	0%	1	3%	1	3%	0	0%
AC (HP)	**11**	0	0%	0	0%	3	27%	1	9%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
AI (HP)	**12**	1	8%	0	0%	0	0%	0	0%	0	0%	2	17%	0	0%	0	0%	0	0%	0	0%

Participants occasionally gave multiple reasons for their behaviour. Thus the count for each treatment may be greater than the total number of individuals in that treatment

Table SI-B3: Selected responses by powerful individuals to "Briefly explain why you gave (or did not give) deduction tokens to other players?". Highlights in blue were referred to in the manuscript.	
Treatment	**English translation**
AC-HP	At the beginning I gave deduction tokens because I could do so at no cost to myself. In the end, there were many rounds where everyone donated and earned equally. I was happy with being ahead of others a bit already.
AC-HP	I gave deduction tokens to others if they had donated less than others, or had bigger earnings than the others. Towards the end, I deducted a large sum from someone who was trying to earn by contributing much less than others (when all of us had donated the maximum already for several rounds).
AC-HP	I did not give deduction tokens because other did not give either, and because I would have felt I was playing the game badly.
AC-HP	I did not give deduction tokens because nobody else did either. There was no need to "punish" anyone, because all behaved group-friendly all the time, also helping me.
AC-HP	I gave deduction tokens if someone donated clearly less. Also if someone gave deduction tokens to me, I could give deduction tokens back.
AC-HP	I gave deduction tokens to discipline the group, to make sure everyone donated 20 on every round. Immediately if someone donated less than 20 I gave deduction points.
AC-HP	I did not give any deduction tokens, because it would not have deducted any points from me [apparently the player would have seen it unfair]. Others did not give deduction tokens either. If we only had played faster, we had gotten more to the common pot. I guess that is how the game works [the player apparently thought duration of the game was tied to time, not to number of rounds].
AI-HP	I did not give deduction tokens because it did not seem like a meaningful thing to do. As long as I was earning well, I was not interested in how much someone else earned. I was thinking that it is good if everyone wins as much as possible, and I did not want to punish anyone for a few points.
AI-HP	I did not give deduction tokens, unless others did not donate or punished other players with no reason.
AI-HP	I tried not to give deduction tokens, unless the player donated really very little. In this way, I tried to urge others to donate more.
AI-HP	I did not give deduction tokens because I was at a better position with respect to that, and I did not see how I could benefit from it.
AI-HP	I gave deduction tokens to balance the earnings of others. I would have also given deduction tokens to myself, had that been possible.
AI-HP	I thought it would be fair for others to have equal number of points.
AI-HP	I gave deduction tokens twice. First, when someone donated very little, and one put 15 right at the start. The other time was a blackout…
AI-HP	I did not give deduction points to anyone at any time, because it would have been unfair me being the player who could not be deducted from. In addition it would have noticeably undermined trust in the group and there would have been less token for everyone.
AI-LP	I was expecting that the untouchable player would punish "unproductive" behavior among group members. If I would start dealing out deduction points I would have to worry about revenge, or worsening the ambience in the group.
AI-HP	At the beginning, I gave deduction tokens to others in order to reduce their earnings. Towards the end I changed the strategy, and I did not give so many deduction tokens anymore.
AI-HP	Above I already mentioned that I punished for small donations if that was necessary.

AI-HP	If I was not given deduction points. I gave deduction points if others had more tokens or someone had given me deduction points.
Selected responses from participants in the symmetrical treatments	
SC	When I realized that everyone's earnings are greatest when everyone contributes maximally. I was thinking what to do if someone tries to free-ride. I decided to give one deduction token as a token slap on the wrist for the one who donates the least, and it seemed to have the desired effect. In the second round two players followed my example and donated the full amount. I put a third player to his place with one token punishment, and he followed suit on the next round as well. At one point one player, out of nowhere, donated 0 tokens. I think it may have been a mistake, but it was a serious offence nevertheless, so I gave 10 deduction points to express my disapproval. I was not going to tolerate free riding.
SC	Others did not give deduction tokens to me after the start, so some sort of solidarity (that seems to be missing from Finnish decision makers) was formed in our team. Anyway, as I am not a narcissist, I do not like to put others down.
SC	I gave deduction tokens to those who donated less than 20, so that they would not make it a habit of it and would behave correctly in future. The intention was to give enough deduction tokens together with others so that the earnings of the punished would be less than what you would get with cooperation (40 points).
SC	There was no need to give deduction tokens because everyone cooperated fully and did not give deduction tokens either. I am not sure what I would have done if someone had not contributed enough, or had given deduction points to me. Maybe I would have reacted on the next round, or in the deduction phase, by giving deduction points as a warning. But I did not need to think about it because everyone in the group behaved perfectly.
SC	I did not give deduction tokens because the others did not give to me either.
SC	I did not give deduction tokens because it diminished my earnings. I also did not give deduction tokens, if I was not given deduction tokens by others. Giving deductions seemed like revenge, which is childish.
SF	I did not give deduction tokens to anyone. It would not have benefitted me in any way; the game went much better when everyone cooperated fully and no-one took anything away from others. And this is how the team worked very soon, which was cool :-). Also, giving deduction tokens would probably have made me feel guilty, and it would have been punishing. When I did not "punish" for the deduction tokens I received on the first round, the cycle of punishing or rewarding did not come to be. (I presume such cycle could have arisen). So, I perceived other players as collaborators, not competitors.
SF	At the beginning , I gave the same amount of deduction tokens I had received. Later on, I gave even larger number of deduction tokens, so that the number of deduction tokens I received was relatively smaller. In the long run, this does not work, because others can give me 120 deduction points in total, when I can only give 40 to the three others.
SF	I gave deduction tokens to those who donated less than the others, but I tried to keep this punishment small, so that they would not start revenging me. Unfortunately one player did [revenge], and started to pick a fight by allocating unjustified deduction tokens to whomever. I tried to turn the public opinion agains him/her, but others did not give him/her deduction tokens like I did, apparently being afraid of revenge. But we would have been 3 against 1... I tried to cool the situation down, but this player continued to sabotage the earnings of others without profiting anything from it, and stopped it only on the last rounds. Quite a social simulator this research.
SF	The game converged to full cooperation very soon, where everyone donated fully to the project and no-one punished anyone with deduction tokens. There was no need to discipline anyone with deduction tokens.

SF	I gave deduction tokens if someone had donated less than me or the others, so that he/she would earn the same as others. I did not give deduction tokens if I had donated less than the others, or if I had the biggest earnings.
SF	I gave deduction token, and immediately I got a deduction point. After that I did not give any deduction tokens, nor did I receive any.
SF	I gave deduction tokens once, when a player donated 0 points - maybe by accident, but I saw it necessary to remark on the importance of being vigilant for the common good of the group.

Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human prisoner's dilemmas. *Evolution and Human Behavior, 37*(4), 263-271.

Bone, J. E., Wallace, B., Bshary, R., & Raihani, N. J. (2015). The effect of power asymmetries on cooperation and punishment in a prisoner's dilemma game. *PLoS ONE, 10*(1), e0117183.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are People Conditionally Cooperative Evidence from a public goods experiment. *Economics Letters, 71*(3), 397-404.

Nikiforakis, N., & Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization, 78*(3), 319-332.

Zar, J. (1999). Biostatistical analysis 4th ed. *New Jersey*.