

# 1 **People prefer coordinated punishment in cooperative interactions**

2

## 3 **Authors**

4 Lucas Molleman<sup>1,2,3\*</sup>, Felix Kölle<sup>2,4\*</sup>, Chris Starmer<sup>2</sup> and Simon Gächter<sup>2,5,6</sup>

5

## 6 **Affiliations**

7 <sup>1</sup> Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin,  
8 Germany

9 <sup>2</sup> Centre for Decision Research and Experimental Economics, School of Economics,  
10 University of Nottingham, United Kingdom

11 <sup>3</sup> Amsterdam Brain and Cognition Center, University of Amsterdam, The Netherlands

12 <sup>4</sup> Faculty of Management, Economics and Social Sciences, University of Cologne, Germany

13 <sup>5</sup> Center for Economic Studies, Munich, Germany

14 <sup>6</sup> IZA Institute of Labour Economics, Bonn, Germany

15

16 \* Corresponding authors: Lucas Molleman ([l.s.molleman@uva.nl](mailto:l.s.molleman@uva.nl); ORCID: 0000-0003-0184-  
17 4240); Felix Kölle ([felix.koelle@uni-koeln.de](mailto:felix.koelle@uni-koeln.de); ORCID: 0000-0003-4036-8566)

18

## 19 **Keywords**

20 cooperation, punishment, conditional preferences, coordination, decision-making experiment,  
21 social dilemmas

22

## 23 **Date**

24 15 July 2019

25 **Abstract**

26 **Human groups can often maintain high levels of cooperation despite the threat of**  
27 **exploitation by individuals who reap the benefits of cooperation without contributing to**  
28 **its costs<sup>1-4</sup>. Prominent theoretical models suggest that cooperation is particularly likely**  
29 **to thrive if people join forces to curb free riding and punish their non-contributing peers**  
30 **in a coordinated fashion<sup>5</sup>. However, it is unclear whether and, if so, how people actually**  
31 **condition their punishment of peers on punishment behaviour by others. Here we**  
32 **provide direct evidence that many people prefer coordinated punishment. With two**  
33 **large-scale decision-making experiments (total  $N = 4,320$ ), we create minimal and**  
34 **controlled conditions to examine preferences for conditional punishment and cleanly**  
35 **identify how individuals' punishment decisions are impacted by punishment behaviour**  
36 **by others. We find that the most frequent preference is to punish a peer only if another**  
37 **(third) individual does so as well. Coordinated punishment is particularly common**  
38 **among participants who shy away from initiating punishment. With an additional**  
39 **experiment we further show that preferences for conditional punishment are unrelated**  
40 **to well-studied preferences for conditional cooperation. Our results highlight the**  
41 **importance of conditional preferences in both positive and negative reciprocity, and**  
42 **provide strong empirical support for theories that explain cooperation based on**  
43 **coordinated punishment.**

## 44 **Main text**

45 The ecological success of humans has often been attributed to our propensity for  
46 cooperation<sup>2,4</sup>. Many people are willing to go out of their way to help others, allowing human  
47 groups to deal with environmental challenges and to do things no individual can achieve on  
48 their own. But, as natural as it might seem, cooperation looks puzzling from the viewpoint of  
49 rational self-interest: why would one cooperate if others can reap the benefits of cooperation  
50 without paying the costs? This supposed ‘free rider problem’ affects countless real-world  
51 situations, ranging from day-to-day work in teams and paying taxes to curbing overfishing  
52 and reducing carbon dioxide emissions. Moreover, a vast range of theoretical models and  
53 empirical studies from across the biological and social sciences have documented that, when  
54 studied in isolation, individually costly cooperation tends to break down through processes of  
55 natural selection or (social) learning, which often favour free riding<sup>1-9</sup>.

56 Peer punishment<sup>15-18</sup> is one of the key mechanisms proposed to explain why cooperation can  
57 thrive despite free rider incentives. When individuals sanction their uncooperative interaction  
58 partners, relative gains from free riding can be offset<sup>19-21</sup>. Influential theoretical arguments  
59 suggest that punishment can be particularly effective in promoting cooperation when  
60 individuals punish non-cooperators in a coordinated fashion<sup>5</sup>. In this paper we use large-scale  
61 decision-making experiments to provide systematic empirical evidence for coordinated  
62 punishment in a social dilemma situation.

63 Peer punishment has been extensively studied in a wide range of different experimental  
64 settings, both in decision-making laboratories and in the field, as well as across different  
65 interaction settings and across cultures<sup>14,20-29</sup>. By and large, these studies indicate that many  
66 people are inclined to punish free riding interaction partners (often motivated by negative  
67 emotions such as anger), supporting the idea that such peer punishment can lead to a welfare-  
68 enhancing stabilization of cooperation at high levels.

69 While in experimental settings punishment has great potential to support cooperation,  
70 evolutionary models suggest that punishment can only emerge under very limited  
71 circumstances<sup>30-32</sup>. The reason is that punishment often entails costs for those who mete it out.  
72 This can create a ‘second-order’ free rider problem: while only some individuals incur the  
73 costs of punishing, all members of a group may benefit from enhanced cooperation after non-  
74 cooperators are punished. Hence, from an individual perspective it can pay to refrain from

75 punishment<sup>24,33–35</sup>. Over time, this may result in a decline of punishment in a population of  
76 self-interested agents, compromising its potential to support cooperation.

77 Theoretical and experimental studies have explored various mechanisms that may address the  
78 second-order free rider problem, such as reputational benefits for punishers<sup>36–39</sup>, the  
79 punishment of those who fail to punish<sup>40–43</sup>, commitment of resources to prepare joint  
80 sanctioning of free riders before cooperative interactions take place<sup>35,44,45</sup>, or the  
81 establishment of specialized authorities that monitor behaviour and punish free riders<sup>46–48</sup>. In  
82 these studies, individuals' decisions to punish are typically considered in frameworks  
83 allowing only independent and uncoordinated actions. In real life, however, punishment does  
84 not typically take place in a social vacuum. Like cooperation, punishment can often be made  
85 dependent on the behaviour of others<sup>40,49–51</sup>, and empirical evidence from the field suggests  
86 that such coordinated punishment may be common in human groups<sup>15,17,52–54</sup>. Moreover, an  
87 influential evolutionary model indicates that punishment is likely to emerge when individuals  
88 can coordinate their punishment<sup>5</sup>. This model suggests that the second-order free rider  
89 problem can be largely avoided when individuals make their punishment conditional upon the  
90 punishment decisions of others. 'Ganging up' against free riders likely decreases the costs for  
91 individual punishers (*e.g.*, through reduced risks for retaliation), and may increase the impact  
92 and effectiveness of punishment as individuals join forces in meting it out<sup>2,5,17</sup>.

93 Despite these promising theoretical results, there is only scarce empirical support for the  
94 claim that individuals have a preference for coordinated punishment. While field  
95 evidence<sup>15,17,52–54</sup> is consistent with the idea of coordinated punishment, it is not conclusive  
96 about whether people prefer coordinated punishment over independent punishment.  
97 Experimental evidence is needed to establish the existence of preferences for coordinated  
98 punishment. Hitherto, experimental studies of conditional punishment are rare, and mainly  
99 focus on how punishment is impacted by the distribution of cooperative behaviour (not the  
100 punishment behaviour) in the population, or analyse how aggregate levels of cooperation are  
101 affected when free riders can only be punished if multiple peers agree to do so<sup>14,49–56</sup>. Here,  
102 we investigate whether people prefer to coordinate their punishment in the context of an  
103 experiment in which individuals can explicitly condition their punishment on the punishment  
104 decisions of others. That is, we ask whether, analogously to conditional preferences observed  
105 in pro-social contexts like cooperation (*cf.* positive reciprocity), such conditionality is also

106 characteristic of preferences to punish others (*cf.* negative reciprocity), and if so, whether  
107 positive and negative reciprocity are correlated.

108 Our results reveal that people do indeed tend to coordinate punishment with their peers.  
109 Among participants who are willing to punish in at least one instance, the most frequent  
110 preference is to use coordinated punishment: punishing only if others do so as well.  
111 Alternative punishment preferences (punishing irrespective of what others do, and punishment  
112 only if others do not) are observed much less frequently. Coordinated punishment particularly  
113 predominates among participants who refrain from initiating punishment. Furthermore, we  
114 confirm all these main results with a large-scale replication study. An analysis of participants'  
115 self-reported motivations suggests that anger is an important driver of punishment, and that  
116 coordinated punishment is associated with equality concerns towards the other punisher.

117 In a follow-up experiment, we demonstrate that preferences for conditional punishment are  
118 unrelated to preferences for conditional cooperation. That is, while at the aggregate level we  
119 observe that individuals are more likely to cooperate and punish if their peers do so too, at the  
120 individual level there is no correlation between preferences for cooperation and punishment.

121 We conduct a decision-making experiment with  $N = 2,004$  participants. After reading  
122 instructions and passing comprehension checks, participants are allocated to groups of three.  
123 In their group, they play a one-shot game consisting of two stages. The first stage is a binary  
124 linear public goods game (PGG) in which all individuals simultaneously choose to either  
125 'defect' or to 'cooperate'. From the perspective of an individual participant, defecting yields a  
126 personal benefit of 5 monetary units (MU), and 0 MU for the other two group members.  
127 Cooperating yields 2 MU for all three group members. This setup creates a social dilemma:  
128 while average payoffs in a group are maximised when all members cooperate, individuals can  
129 maximise their own monetary benefits by defecting (*i.e.* free riding), making defection a  
130 dominant strategy leading to a socially inefficient outcome.

131 At the start of the second stage, two of the three group members are randomly allocated the  
132 roles of Punishers, and the remaining one is allocated the role of Target who can be punished  
133 but cannot punish the Punishers. The PGG decision of the Target is revealed, and then  
134 Punishers make binary choices whether or not to punish (*i.e.* assign deduction points to) the  
135 Target. Assigning deduction points incurs a cost of 1 MU to the Punisher, and 3 MU to the  
136 Target. The impact of punishment is additive: when a Target is punished by one group

137 member they lose 3 MU and when they are punished by both of their group mates they lose 6  
138 MU. Punishers cannot assign deduction points to one another, neither can they observe each  
139 other's PGG decision when making their deduction point decision (mitigating possible effects  
140 of inequality between Punishers stemming from the contribution stage).

141 We examine whether people prefer to coordinate their punishment by having Punishers make  
142 two types of punishment decisions. First, they make an 'unconditional' punishment decision,  
143 deciding whether they want to punish the Target or not, irrespective of the decision of the  
144 other Punisher. Subsequently, they make two 'conditional' punishment decisions, in which  
145 Punishers can condition their punishment on that of the other Punisher. To do this, we use the  
146 strategy method<sup>63</sup>: Punishers indicate whether or not they want to punish the Target in case  
147 the other Punisher chooses (1) to punish or (2) to not punish the Target. Our analysis mainly  
148 focuses on these two decisions made by the  $N = 1,336$  Punishers in our experiment. Figure 1  
149 summarises the decision situations.

150 Once both Punishers have entered both types of decisions, one Punisher is randomly chosen  
151 and their unconditional punishment decision is implemented to initiate the punishment  
152 procedure. Subsequently, the corresponding conditional punishment decision of the other  
153 Punisher is implemented (Methods). This setup yields an incentive-compatible decision  
154 situation in a minimal social context that allows us to cleanly measure how people condition  
155 their punishment on the punishment decisions of others. Importantly, the strategy method  
156 yields a full punishment profile for each individual that is independent of their beliefs about  
157 other people's punishment decision. Because of the one-shot nature of the game, there are  
158 also no strategic incentives for punishment. Further, note that because punishment is costly  
159 and the game is played only once, if players are only interested in maximising their own  
160 material payoff, no one is predicted to punish. As a result, full defection with no-punishment  
161 is the only Nash equilibrium of the game.

162 In a post-experimental questionnaire we asked Punishers to self-report their experienced  
163 levels of anger when they learned about the Target's PGG decision (Methods). Negative  
164 emotions such as anger are commonly identified as key drivers of punishment<sup>21,27,64-67</sup>. With  
165 this questionnaire item we test whether anger is not only correlated with individual  
166 punishment decisions but also with preferences for conditional punishment.

167 The cooperation rate in the PGG stage of the game was 48%. Among the 1,336 Punishers, the  
168 overall unconditional punishment rate was 11.4%. Unconditional punishment varied  
169 considerably with the cooperation decisions of Punishers and their Target. In particular,  
170 unconditional punishment rates were highest when the Punisher cooperated and the Target  
171 defected (24.0%), and lowest when both the Punisher and Target cooperated (5.5%). When  
172 the Target cooperated and the Punisher defected, punishment was also low (5.7%), while if  
173 both players defected punishment was intermediate occurring in 10.0% of the cases.

174 On aggregate, people were much more likely to punish their peers when the other punisher  
175 did so as well. A decision of the other Punisher to punish the Target increased overall  
176 punishment rates by 40% (from 11.1% when others did not punish, to 15.5% when they did;  
177 McNemar test:  $\chi^2(1) = 16.66, P < 0.001, \phi_c = 0.11$ ). We interpret this as evidence that people  
178 tend to prefer to coordinate their punishment.

179 To test the robustness of this result, we fitted a logistic generalized linear model to conditional  
180 punishment decisions, confirming that this increase is statistically significant (Table 1, Model  
181 1,  $P < 0.001$ ). This model further shows that participants who punished unconditionally  
182 displayed much higher levels of punishment in the conditional stage ('Unconditional  
183 punishment';  $P < 0.001$ ). It also reveals that the relative influence of the other Punisher's  
184 punishment differs strongly between those who did and those who did not punish  
185 unconditionally. For those who did not punish unconditionally (the baseline case in Model 1),  
186 we observe a strong and positive effect of the other's punishment on punishment levels. By  
187 contrast, for those who did punish unconditionally, the analysis indicates that the decision of  
188 the other Punisher did not systematically affect their punishment (the joint effect of 'Other  
189 punishes' and its interaction term with 'Unconditional punishment' is not significantly  
190 different from zero; Wald test:  $\chi^2(1) = 0.15, P = 0.696$ ). Taken together, this analysis suggests  
191 that coordination effects are driven by those who did not punish unconditionally.

192 In Model 2 we confirm that the positive effect of the other punisher's decision is robust to  
193 alternative model specifications, and further investigate how preferences for coordinated  
194 punishment vary with the Target's decision to cooperate or defect. We find that a Target's  
195 decision to cooperate leads to lower overall levels of punishment, as indicated by the  
196 significant negative 'Target cooperated' dummy. Preferences for coordinated punishment,  
197 however, seem independent of the Target's PGG decision as the interaction term of 'Target  
198 cooperated'  $\times$  'Other punishes' is not statistically significant. This means that defection leads

199 to higher levels of punishment, but does not make it more likely that people coordinate (or  
200 anti-coordinate) their punishment. However, as unconditional punishment towards  
201 cooperators (“antisocial punishment”)<sup>25</sup> occurred only in 5.6% of cases (see above),  
202 coordinated antisocial punishment is very rare in our data; only 0.63% of the cooperators got  
203 punished in a coordinated way.

204 We now turn to the individual-level preferences for conditional punishment elicited with the  
205 strategy method (Figure 1e,f). We distinguish four ‘punishment types’: (i) ‘coordinated  
206 punisher’, punishing only if the other does so as well; (ii) ‘anti-coordinated punisher’,  
207 punishing only if the other does not punish; (iii) ‘independent punisher’, punishing regardless  
208 of the other’s punishment decision; and (iv) ‘non-punisher’, not punishing at all. Here we  
209 focus on the relative frequencies of these types.

210 In line with established findings for one-shot games without strategic incentives to  
211 punish<sup>23,27,68,69</sup>, the majority of participants in the role of Punisher chose to never punish  
212 (78.9%). Among the  $N = 282$  Punishers who punish at least once, we find a strong and  
213 striking pattern (Figure 2a). In this group, coordinated punishers are most frequent (47.5%).  
214 The independent punishers and anti-coordinated punishers are much less frequent (25.9% and  
215 26.6%, respectively).

216 Figure 2 also shows the distribution of punishment types, split by the unconditional  
217 punishment decision. In line with the regression results presented in Table 1, this  
218 decomposition reveals that coordinated punishment is particularly prevalent among those who  
219 do not punish in the unconditional stage (Figure 2b). Among those participants who do punish  
220 in the unconditional punishment stage, we most frequently observe independent punishment  
221 (Figure 2c). The distribution of punishment types across unconditional punishers and  
222 unconditional non-punishers is highly significantly different ( $\chi^2(2) = 52.88, P < 0.001, \varphi_c =$   
223  $0.43$ ). In particular, among unconditional non-punishers there is a significantly larger fraction  
224 of ‘coordinated punishers’ ( $\chi^2(1) = 37.77, P < 0.001, \varphi_c = 0.37$ ) and a significantly smaller  
225 fraction of ‘independent punishers’ ( $\chi^2(1) = 44.49, P < 0.001, \varphi_c = 0.40$ ). The fraction of ‘anti-  
226 coordinated punishers’, in contrast, is very similar across both subsets ( $\chi^2(1) = 0.11, P =$   
227  $0.738, \varphi_c = 0.02$ ; see Supplementary Figure 1 for a decomposition of conditional punishment  
228 types for each of the outcomes of the PGG).



229 These behavioural patterns demonstrate that many people prefer coordinated punishment. As  
230 we used a new experimental paradigm to examine rarely explored preferences for conditional  
231 punishment, one might ask how robust our results are, and, perhaps more fundamentally, why  
232 people would prefer coordinated punishment.

233 To test the robustness and replicability of our results, and to further probe motivations  
234 underlying conditional punishment preferences, we ran a new study with  $N = 2,316$  additional  
235 participants. The design and the procedures of this new study were the same as in our original  
236 experiment, with the following exceptions. First, to rule out that the observed patterns in  
237 punishment preferences are due to confusion about the strategy method and the payoff  
238 consequences for punishment, we added a set of control questions right before participants  
239 entered the punishment stages. Second, as a further robustness check we counterbalanced the  
240 order in which participants made their punishment decisions, such that in half of the groups,  
241 Punishers completed the ‘conditional’ stage (*cf.* Figure 1d,e) before the ‘unconditional’ stage  
242 (*cf.* Figure 1c). Third, to further explore the motivational factors underlying preferences for  
243 conditional punishment, we extended the post-experimental questionnaire with items probing  
244 not only experienced anger when making punishment decisions, but also other possible  
245 motivations such as a desire for revenge towards the Target, reciprocity towards the other  
246 Punisher, as well as inequality concerns (see below).

247 The results of the new study closely replicate our original findings. Again, participants were  
248 significantly more likely to punish when the other punisher did so as well (Supplementary  
249 Table 1), demonstrating that the observed punishment patterns in our original study were not  
250 driven by confusion. When we focus on the people who punish at least once in the conditional  
251 punishment stage and split up the data according to decisions in the unconditional punishment  
252 stage, we observe that the same punishment types predominate as before: coordinated  
253 punishment prevails among unconditional non-punishers (55.3%) and independent  
254 punishment prevails among unconditional punishers (51.4%; Supplementary Figure 2). As  
255 before, these distributions of conditional punishment preferences significantly differed from  
256 each other ( $\chi^2(2) = 37.01, P < 0.001, \varphi_c = 0.35$ ). Furthermore, each of them closely match  
257 their corresponding distribution from our original study (unconditional non-punishers:  $\chi^2(2) =$   
258  $2.94, P = 0.230, \varphi_c = 0.10$ ; unconditional punishers;  $\chi^2(2) = 1.07, P = 0.584, \varphi_c = 0.06$ ), and  
259 did not vary with order ( $\chi^2(2) = 1.61, P = 0.448, \varphi_c = 0.11$  for unconditional non-punishers;  
260  $\chi^2(2) = 2.63, P = 0.268, \varphi_c = 0.12$  for unconditional punishers).

261 Taking the data from both studies together, we find that among those  $N = 584$  participants  
262 who punish at least once in the conditional stage, 43% have a preference for coordinated  
263 punishment. A further 32% are independent punishers and 25% are anti-coordinated  
264 punishers. So, overall, our results suggest that coordinated punishment is a strong and robust  
265 phenomenon prevailing across different outcomes of cooperative interactions. Preferences for  
266 coordinated punishment are particularly common among people who do not punish  
267 unconditionally. In the Supplementary Information, we develop a simple model to explore  
268 how the relative frequencies of punishment preferences observed in our data may impact the  
269 relative payoffs of cooperation and defection. This model suggests that the range of  
270 conditions for which cooperation is favoured over defection can be substantially enhanced by  
271 the presence of individuals who do not punish unconditionally, but who are prepared to  
272 punish once another individual initiates it (Supplementary Figure 3; Supplementary Results).

273 To understand the potential drivers and underlying motivations of individuals' punishment  
274 preferences, we analyse Punishers' reported levels of anger (using data from both studies) as  
275 well as their responses to the post-experimental questionnaire in our replication study. For  
276 eliciting anger levels, we asked punishers to rate their level of anger when making their  
277 punishment decisions on a 7-point scale (1: not angry at all; 7: very angry). On average, anger  
278 scores are highest when the Punisher cooperated and the Target defected (3.7) and lowest if  
279 both cooperated (1.5). We find anger levels to be significantly higher for unconditional  
280 punishers than for unconditional non-punishers (3.7 versus 2.2; Mann Whitney U (MWU)-  
281 test,  $z = 15.34$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.29$ ).

282 Reported anger levels also vary markedly across the different conditional punishment types  
283 (Figure 3). Participants who never punished report average anger levels of only 2.1, which is  
284 significantly lower than those reported by any of the other types (MWU-tests for pairwise  
285 comparisons, d.f. = 1, all  $P < 0.001$ ). Furthermore, independent punishers tend to report  
286 higher average anger levels (3.8) than coordinated punishers (3.1, MWU-test,  $z = 3.67$ , d.f. =  
287 1,  $P < 0.001$ ,  $r = 0.18$ ) and anti-coordinated punishers (3.2, MWU-test,  $z = 2.86$ , d.f. = 1,  $P =$   
288 0.004,  $r = 0.16$ ), while there is no significant difference between the latter two types (MWU-  
289 test,  $z = 0.55$ , d.f. = 1,  $P = 0.586$ ,  $r = 0.03$ ). When accounting for multiple comparisons,  
290 differences that are significant in this analysis remain so (*i.e.*, by multiplying  $P$  values by 6;  
291 the number of comparisons). These findings suggest that conditional punishers may be less  
292 emotionally aroused than independent punishers, and are less driven by emotions of anger.

293 In the extended questionnaire from the new study, participants used a 7-point scale to indicate  
294 their agreement to a set of statements, designed to test a set of candidate motivations that we  
295 hypothesized to be associated with some specific punishment preferences but less so with  
296 others. Our approach was to perform targeted comparisons for each statement, singling out  
297 one specific punishment type which we linked, a priori, to the respective motivation. In  
298 particular, we tested whether agreement scores were higher in that punishment type than in  
299 others, giving us first correlational hints at possible motivations behind conditional  
300 punishment preferences.

301 In our analyses, we focus only on those participants who punished at least once in the  
302 conditional stage. That is, we disregard the non-punishers – who, unsurprisingly, report  
303 different motivations for their behaviour in the punishment stage of the experiment, relative to  
304 those who punished at least once, rendering all overall tests between distributions of types  
305 significant (Kruskal-Wallis tests, d.f. = 3, all  $P < 0.001$ ). For a full analysis of participants’  
306 responses broken down by punishment type, including non-punishers, see Supplementary  
307 Results.

308 Our analysis of anger outlined above suggests that independent punishers (IP) might be driven  
309 by a thirst for revenge<sup>70,71</sup>. This motivation is corroborated by the observation that  
310 independent punishers agree most with the statement ‘I wanted to reduce the other player’s  
311 earnings myself’ ( $\mu_{IP} = 4.8$ ,  $\mu_{\text{other punishers}} = 4.0$ ; MWU-test:  $z = 4.11$ , d.f. = 1,  $P < 0.001$ ,  $r =$   
312 0.24). By contrast, one motivation behind anti-coordinated punishment (ACP) could be  
313 wishing to see free riding being sanctioned, but only at moderate levels. The data does not  
314 support this idea: while anti-coordinated punishers agreed more with the statement ‘I did not  
315 want to reduce Red’s earnings by too much’ than independent punishers did, this difference  
316 was not significant ( $\mu_{ACP} = 3.8$ ,  $\mu_{\text{other punishers}} = 3.5$ ; MWU-test:  $z = 0.80$ , d.f. = 1,  $P = 0.423$ ,  $r$   
317 = 0.06).

318 We further hypothesized that coordinated punishers might be only willing to punish if others  
319 do so too, because they do not want their payoffs to fall behind those of the other punisher.  
320 Consistent with this hypothesis, we found that, among those participants who punished at  
321 least once, coordinated punishers (CP) tended to agree the most with the statement ‘I did not  
322 want to earn less than Blue [the other punisher]’ ( $\mu_{CP} = 4.9$ ,  $\mu_{\text{other punishers}} = 4.6$ ; MWU-test:  $z =$   
323 2.09, d.f. = 1,  $P = 0.036$ ,  $r = 0.12$ ). Another rationale for coordinated punishment might be  
324 that they see punishment by the other group member as a nice act (enforcing a norm of

325 cooperation) and feel the need to reciprocate. This idea is supported by the observation that  
326 coordinated punishers agreed most with the statement ‘I did not want to let Blue [the other  
327 punisher] down in case they chose to punish’ ( $\mu_{CP} = 4.9$ ,  $\mu_{\text{other punishers}} = 4.5$ ; MWU-test:  $z =$   
328  $2.26$ , d.f. = 1,  $P = 0.024$ ,  $r = 0.13$ ). Finally, coordinated punishers may be unsure what to do  
329 when making their punishment decisions, or be unsure whether punishment is socially  
330 appropriate or legitimate<sup>15,53</sup>, and take others’ punishment behaviour as a ‘principle of social  
331 proof’<sup>72</sup>. We did not find support for these possible motives: coordinated punishers did not  
332 agree more with the statements ‘When making my [conditional punishment] decisions, I was  
333 unsure what to do’ ( $\mu_{CP} = 4.0$ ,  $\mu_{\text{other punishers}} = 4.0$ ; MWU-test:  $z = 0.11$ , d.f. = 1,  $P = 0.914$ ,  $r =$   
334  $0.01$ ) or ‘When making my [conditional punishment] decisions, I was unsure what was the  
335 appropriate thing to do’ ( $\mu_{CP} = 4.2$ ,  $\mu_{\text{other punishers}} = 4.1$ ; MWU-test:  $z = 0.01$ , d.f. = 1,  $P =$   
336  $0.998$ ,  $r = 0.01$ ).

337 Across our two sets of experiments, we find unambiguous evidence that many people like to  
338 condition their punishment decisions on those of other people. This conditionality is  
339 reminiscent of ‘conditional cooperation’, that is, many people’s conditional willingness to  
340 contribute to a public good in the first place provided others do the same. This raises the  
341 interesting question whether conditional punishers are also conditional co-operators. We  
342 therefore now examine how preferences for conditional punishment (*i.e.* negative reciprocity)  
343 relate to well-studied preferences for conditional cooperation (*i.e.* positive reciprocity)<sup>73</sup>. That  
344 is, are these preferences linked and do they reflect a general sensitivity to social influence by  
345 peers, or are they unrelated, indicating that inclinations to reciprocate are context-specific?  
346 Existing evidence, which, with exceptions<sup>62,74</sup>, only looked at cooperation and punishment  
347 decisions and did not elicit conditional preferences, suggests that positive and negative  
348 reciprocity are unrelated<sup>75</sup>. However, given that conceptually, cooperation and punishment  
349 share the logic of a public good (while both cooperation and punishment are individually  
350 costly, all group members may benefit), one might expect that people who prefer to cooperate  
351 conditionally on others’ cooperation also prefer to condition their punishment on others’  
352 punishment. To test this hypothesis, we conducted a follow-up experiment examining whether  
353 individuals’ punishment preferences are related to their preferences for conditional  
354 cooperation.

355 Two weeks after participating in our conditional punishment experiments (both from the  
356 original and the new study reported above), a subset of participants who were in the role of a

357 punisher was re-invited to participate in an additional study (see below for subset details). In  
358 this follow-up experiment, participants were randomly matched with a partner to play a one-  
359 shot dyadic binary Prisoner's Dilemma Game, in which both players had to choose to either  
360 'cooperate' or 'defect'. From the perspective of an individual participant, defecting yielded a  
361 personal benefit of 3 monetary units (MU), and 0 MU for their partner. Cooperating yielded 2  
362 MU for both partners (for instructions, see Supplementary Methods).

363 As in the primary experiment, participants in the follow-up experiment had to make two types  
364 of decisions: an 'unconditional' and a 'conditional' decision<sup>76</sup>. After their unconditional  
365 decision to either 'cooperate' or 'defect', participants entered a second ('conditional') stage in  
366 which they could make their cooperation decision dependent on the cooperation decision of  
367 their partner. Again, we recorded these conditional cooperation decisions using the strategy  
368 method<sup>76</sup>: participants indicated their decision in case their partner would either cooperate or  
369 defect. Within a pair, earnings were determined by implementing the unconditional decision  
370 of one randomly chosen partner and the corresponding conditional decision of the other  
371 partner (Methods). This procedure allows us to classify participants into three distinct  
372 cooperation types<sup>14,74</sup>: 'conditional cooperators' (those who cooperate only if their partner  
373 cooperates, but defect otherwise), 'free riders' (those who always defect irrespective of their  
374 partner), and 'others' (those who fall under neither of the first two categories).

375 To ensure sufficient statistical power, we selectively re-invited participants to obtain a more  
376 balanced sample with respect to the distribution of punishment types, compared to our full  
377 sample from the primary experiment. That is, we aimed to oversample those who punished at  
378 least once in the conditional stage in the primary experiment (*cf.* Figure 2), and undersample  
379 those who never punished. This procedure indeed led to a more evenly distributed sample  
380 with respect to punishment types: among the  $N = 381$  participants in the follow-up  
381 experiment, 39% were non-punishers, 22% were unconditional punishers, 23% were  
382 coordinated punishers, and 16% were anti-coordinated punishers. In this sample we find a  
383 commonly observed pattern with regard to the distribution of cooperation types: more than  
384 half of the people (56%) are conditional cooperators, about 28% are free riders, and the  
385 remaining 16% are classified as 'others'.

386 To investigate whether preferences for conditional cooperation and conditional punishment  
387 are linked at the individual level, we compare the distribution of punishment types across the  
388 different cooperation types. If these two types of preference reflect a more general

389 behavioural tendency (*e.g.*, inclinations to reciprocate, or to conform with the behaviour of  
390 others), we would expect that coordinated punishment is particularly frequent among  
391 conditional cooperators.

392 We find no evidence for a systematic relation between preferences for conditional cooperation  
393 and preferences for conditional punishment (Figure 4). The overall distribution of conditional  
394 punishment types does not differ across the cooperation types ( $\chi^2(6) = 5.26, P = 0.510, \varphi_c =$   
395  $0.08$ ). Furthermore, individuals with preferences for coordinated punishment were not  
396 disproportionately more likely to have preferences for conditional cooperation ( $\chi^2(1) = 0.71, P$   
397  $= 0.401, \varphi_c = 0.04$ ). Interestingly, free riders who, by definition, are unwilling to contribute to  
398 a first-order public good, are not less likely to contribute to the second-order public good of  
399 punishment compared with conditional cooperators. That is, their preference for conditional  
400 punishment is not different from those of conditional cooperators ( $\chi^2(3) = 2.25, P = 0.522, \varphi_c$   
401  $= 0.08$ ). These results suggest that conditional preferences in positive and negative reciprocity  
402 do not follow the same logic<sup>74</sup>. Positive effects of peer behaviour do not, for example, reflect  
403 a simple conformist heuristic of blindly following others.

404 Our large-scale experiments provide firm empirical evidence that many people prefer to  
405 coordinate their punishment in cooperative interactions. Our results support theories that  
406 explain the emergence and maintenance of human cooperation based on individuals  
407 sanctioning their peers jointly rather than individually<sup>5</sup>. When deciding whether or not to  
408 punish a peer, many people are more willing to engage in costly punishment if others do so  
409 too. Intriguingly, preferences for coordinated punishment are particularly pronounced among  
410 those who do not punish unconditionally, suggesting that punishment levels can rise  
411 substantially when people have the opportunity to coordinate their sanctions.

412 Our results indicate that conditional preferences are not limited to the domain of positive  
413 reciprocity (cooperation), but extend to the domain of negative reciprocity (punishment), too.  
414 On aggregate, in both domains conditional preferences lead individuals to align their  
415 decisions with others and conform to their actions. However, there is substantial heterogeneity  
416 in how individuals condition their punishment on the punishment behaviour of others (Figure  
417 2). Interestingly, our data suggest that people's conditional preferences in the domains of  
418 positive and negative reciprocity are unrelated (Figure 4). This result supports the emerging  
419 view that behavioural strategies of cooperation and punishment are not closely associated  
420 with each other<sup>14,74,75,77–80</sup>, and suggests that cooperation and punishment are separate

421 phenomena, each driven by its own psychological processes. Indeed, the lack of correlation  
422 between conditional punishment and conditional cooperation indicates that, while first-order  
423 and second-order free rider problems may look theoretically very similar, the underlying  
424 mechanisms supporting behaviour in them may be quite distinct.

425 Our analysis of anger levels provides a first step in understanding the possible drivers of  
426 conditional punishment. ‘Independent punishers’ – who punished regardless of the  
427 punishment of others – reported the highest levels of anger. This indicates that negative  
428 emotions are an important factor explaining punishment behaviour in our experiment.  
429 Interestingly, we observe lower levels of anger for individuals who condition their  
430 punishment on that of others (‘coordinated punishers’ who only punished if the other  
431 punished as well, and ‘anti-coordinated punishers’ who punished only if the other refrained  
432 from punishment). This suggests that, compared to independent punishers, the preferences of  
433 conditional punishers might perhaps reflect a more deliberative attitude, with behaviour  
434 relatively less likely to be driven by negative emotions.

435 Our additional questionnaires provide further correlational hints regarding possible  
436 motivations underlying conditional punishment preferences. Independent punishment was  
437 associated with a desire to mete out punishment oneself, supporting the idea that this  
438 preference might be driven by a ‘thirst for revenge’. By contrast, preferences for coordinated  
439 punishment were linked with increased concerns for equality and not letting the other  
440 punisher down, suggesting (positive) reciprocity towards the other punisher. While these  
441 questionnaire results provide some initial indication of why people might prefer to punish  
442 conditionally, the observed differences between the punishment types were relatively small.  
443 Moreover, our examination of the possible motivations is by no means exhaustive, and we  
444 consider our current analysis to be a first step. Systematically uncovering the motivations  
445 underlying conditional punishment preferences would be an interesting direction for future  
446 study, which could contribute to a more comprehensive understanding of the psychological  
447 determinants of peer punishment.

448 We conducted our experiments online, with American participants from Amazon Mechanical  
449 Turk. This platform is well suited for large-scale studies like ours, and gives the opportunity  
450 to recruit a more demographically diverse sample than student samples typically recruited in  
451 laboratory studies<sup>81</sup>. Although collecting data online is associated with reduced levels of  
452 experimental control relative to the traditional decision making laboratory, this does not have

453 to compromise data quality<sup>82-84</sup>, especially when the methodological challenges of conducting  
454 experiments online are adequately addressed<sup>29</sup>. Moreover, a recent study on cooperation and  
455 punishment found that MTurkers punish in a similar way as students in the lab<sup>29</sup>, giving  
456 reason to be optimistic about the generalizability of our results. However, it is an empirical  
457 question whether the patterns of conditional punishment preferences from our study would be  
458 observed under more standard laboratory conditions, or, for example, in samples from  
459 different cultural backgrounds.

460 We designed our experiments to identify preferences for conditional punishment in a highly  
461 controlled experimental scenario that isolates the impact of a peer's punishment behaviour on  
462 people's tendencies to punish. At the same time, our design strived to minimize potential  
463 confounding effects due to factors like non-anonymity, the possibility of future interactions  
464 with those you punish, or anticipated counter-punishment. Of course, using a stylized social  
465 context comes at a cost of realism, and might limit the generalisability of experimental  
466 findings. In our case, the observed strong association between conditional punishment types  
467 and (self-reported) anger suggests that decisions in our experiment are at least partly  
468 motivated by factors that are commonly considered to drive punishment in the wild.  
469 Nevertheless, studies of conditional punishment in more contextualized settings<sup>85,86</sup> would  
470 make valuable complements to the experiments presented here.

471 Our study set out to investigate preferences for conditional punishment in a very simple and  
472 'minimal' environment: that is, one which is just complex enough to allow clean tests for  
473 conditional punishment preferences. While we judge this is the right place to conduct initial  
474 tests for such preferences, having provided clear evidence for them, we believe that  
475 incrementally increasing the complexity of the experimental decision-making situation (*i.e.*,  
476 the number of factors at play) will help achieve a more complete empirical understanding of  
477 conditional punishment. Interesting extensions of our basic experiment would include non-  
478 linear returns to scale of punishment<sup>87</sup>. For example, coordinated sanctions might be more  
479 efficient than individual, uncoordinated punishment, and less risky for those who mete them  
480 out as revenge is less likely<sup>5</sup>. Testing whether anticipating such 'synergy' modulates people's  
481 preferences for coordinated punishment would be of great value. Further experiments could  
482 test how coordinated punishment impacts the long-run dynamics of cooperation. When time  
483 horizons are longer than the one-shot interactions used in our study, decisions to cooperate  
484 and punish have a strategic dimension, potentially involving interactions between coordinated



485 punishment and individuals' reputation. Such experiments could also test the theoretically  
486 predicted deleterious implications of anti-social punishment<sup>88,89</sup> in situations where defectors  
487 coordinate their punishment and 'gang up' against cooperators<sup>90,91</sup>.

## 488 **Methods**

489 For the primary experiment, we recruited  $N = 2,004$  participants from Amazon Mechanical  
490 Turk, two thirds of whom ( $N = 1,336$ ) had the role of Punisher (Figure 1). Participants  
491 completed the experiment in about 10 minutes and earned a flat fee of \$0.50 plus their  
492 earnings from the game. At the end of a session, monetary units were converted into money at  
493 the rate 10 MU = \$1.00. Total average earnings were \$1.50, which corresponds to an average  
494 hourly wage of \$9.00. Before the start of the PGG, each participant had to correctly answer a  
495 set of control questions designed to test their understanding of the interaction setting.

496 Participants were all US citizens; 55% were male, and their mean age was 32.7 years. The  
497 online experiment was developed with the software LIONESS<sup>92</sup>; code available upon request  
498 from the corresponding authors. Ethical approval was provided by the Research Ethics  
499 Committee at the School of Economics, University of Nottingham. All experimental  
500 instructions are documented in the Supplementary Methods.

501 Experimental sessions ended with a short questionnaire. In the questionnaire, we asked  
502 Punishers to indicate how angry they felt when they learned about the Target's PGG decision  
503 on a Likert scale from 1 (not angry at all) to 7 (very angry; see Supplementary Methods for  
504 exact question wording). We also recorded age and gender.

505 For the follow-up experiment, intended to measure preferences for conditional cooperation<sup>76</sup>,  
506 we recruited  $N = 177$  individuals who had participated in the primary experiment. The follow-  
507 up experiment was programmed in Qualtrics and took about 7-8 minutes. Participants were  
508 matched post-hoc to calculate their earnings, consisting of a flat fee of \$0.50 plus their  
509 earnings from the game, which were converted into money at the rate 5 MU = \$1.00. To  
510 calculate their game earnings, we matched 176 of the 177 participants in pairs. A random  
511 mechanism chose which type of decision was implemented for each partner. In particular, for  
512 one player the 'unconditional' cooperation decision was implemented (the first mover), while  
513 for the other player (the second mover) the corresponding conditional cooperation decision  
514 (depending on the first mover's decision) was implemented. The earnings for the remaining  
515 (177<sup>th</sup>) participant were calculated by using their unconditional cooperation decision and

516 implement the corresponding conditional cooperation decision of a randomly chosen other  
517 participant. Total average earnings were \$1.25, which corresponds to an average hourly wage  
518 of \$10.00.

519 Our replication study had the same general setup as our original study. For the conditional  
520 punishment experiments, we recruited  $N = 2,316$  additional participants on MTurk (all US  
521 citizens; 53% male, mean age 34.9 years; sample size based on a power analysis, presented in  
522 Supplementary Figure 4). Relative to the original study, we made three changes: on top of the  
523 control questions prior to the cooperation stage of the game (as used in the original study), we  
524 added control questions prior to the punishment stage. Furthermore, we counterbalanced the  
525 order of the ‘unconditional’ (Figure 1c) and the ‘conditional stage’ (Figure 1d,e), so that half  
526 of the participants made their conditional punishment decisions first. Finally, we added a set  
527 of questionnaire items directly probing possible motivations for conditional punishment  
528 preferences, as well as items to explore links between conditional punishment preferences  
529 with personality characteristics. These items and the analysis of participants’ responses are  
530 detailed in the Supplementary Results. We recruited  $N = 204$  individuals who participated in  
531 the replication study for the follow-up experiment measuring preferences for conditional  
532 cooperation, which was identical to the follow-up experiment from the original study.

533 Reported tests were two-tailed, unless stated otherwise. Sample sizes for the original study  
534 were not based on an explicit power analysis due to a lack of directly comparable experiments  
535 to base a power analysis on. We used the data from the original study to perform a power  
536 analysis for the replication study (Supplementary Figure 4). After being matched into groups,  
537 participants were randomly assigned a role (Punisher or Target). All Punishers encountered  
538 both relevant conditions in the strategy method (one where the other participant chose to  
539 punish, and one where they chose to not-punish). In the replication study, the order of the  
540 unconditional and conditional decisions was counterbalanced between interaction groups.  
541 Data collection and analysis were not performed blind to the conditions of the experiments.  
542 No data from interaction groups who completed the experiment was excluded from the  
543 reported analyses.

544

545 **Data availability.** All data underlying the results reported in our manuscript can be found on  
546 Github at [https://github.com/LucasMolleman/NHB\\_CoordinatedPunishment](https://github.com/LucasMolleman/NHB_CoordinatedPunishment).

547 **Code Availability.** Analysis code (for STATA) can be found on Github at  
548 [https://github.com/LucasMolleman/NHB\\_CoordinatedPunishment](https://github.com/LucasMolleman/NHB_CoordinatedPunishment).

549

550

## 551 **References**

- 552 1. Fehr, E., Fischbacher, U. & Gächter, S. Strong reciprocity, human cooperation, and the  
553 enforcement of social norms. *Hum. Nat.* **13**, 1–25 (2002).
- 554 2. Bowles, S. & Gintis, H. *A Cooperative Species: Human Reciprocity and Its Evolution*.  
555 (Princeton University Press, 2011).
- 556 3. Rand, D. G. & Nowak, M. A. Human cooperation. *Trends Cogn. Sci.* (2013).
- 557 4. Henrich, J. *The Secret of Our Success: How Culture Is Driving Human Evolution,*  
558 *Domesticating Our Species, and Making us Smarter*. (Princeton University Press, 2015).
- 559 5. Boyd, R., Gintis, H. & Bowles, S. Coordinated punishment of defectors sustains  
560 cooperation and can proliferate when rare. *Science* **328**, 617–620 (2010).
- 561 6. Hamilton, W. D. The genetical evolution of social behaviour I and II. *J. Theor. Biol.* **7**, 1–  
562 52 (1964).
- 563 7. Gintis, H. *Game theory evolving: A problem-centered introduction to modeling strategic*  
564 *interaction*. (Princeton University Press, 2000).
- 565 8. Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302**,  
566 1907–1912 (2003).
- 567 9. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563  
568 (2006).

- 569 10. Lehmann, L. & Keller, L. The evolution of cooperation and altruism—a general framework  
570 and a classification of models. *J. Evol. Biol.* **19**, 1365–1376 (2006).
- 571 11. Egas, M. & Riedl, A. The economics of altruistic punishment and the maintenance of  
572 cooperation. *Proc. R. Soc. Lond. B Biol. Sci.* **275**, 871–878 (2008).
- 573 12. Fischbacher, U. & Gächter, S. Social preferences, beliefs, and the dynamics of free riding  
574 in public good experiments. *Am. Econ. Rev.* **100:1**, 541–556 (2010).
- 575 13. Burton-Chellew, M. N., El Mouden, C. & West, S. A. Social learning and the demise of  
576 costly cooperation in humans. *Proc R Soc B* **284**, 20170067 (2017).
- 577 14. Gächter, S., Kölle, F. & Quercia, S. Reciprocity and the tragedies of maintaining and  
578 providing the commons. *Nat. Hum. Behav.* **1**, 650 (2017).
- 579 15. Boehm, C. *Hierarchy in the Forest: Egalitarianism and the Evolution of Human Altruism*.  
580 (Harvard University Press, 1999).
- 581 16. Sigmund, K. Punish or perish? Retaliation and collaboration among humans. *Trends Ecol.*  
582 *Evol.* **22**, 593–600 (2007).
- 583 17. Guala, F. Reciprocity: Weak or strong? What punishment experiments do (and do not)  
584 demonstrate. *Behav. Brain Sci.* **35**, 1–15 (2012).
- 585 18. Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nat. Hum.*  
586 *Behav.* **2**, 458 (2018).
- 587 19. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: Self-  
588 governance is possible. *Am. Polit. Sci. Rev.* 404–417 (1992).

- 589 20. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am.*  
590 *Econ. Rev.* **90**, 980–994 (2000).
- 591 21. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
- 592 22. Henrich, J. *et al.* In search of homo economicus: behavioral experiments in 15 small-scale  
593 societies. *Am. Econ. Rev.* **91**, 73–78 (2001).
- 594 23. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767–1770  
595 (2006).
- 596 24. Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. Winners don't punish. *Nature*  
597 **452**, 348 (2008).
- 598 25. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science*  
599 **319**, 1362–1367 (2008).
- 600 26. Gächter, S., Renner, E. & Sefton, M. The long-run benefits of punishment. *Science* **322**,  
601 1510–1510 (2008).
- 602 27. Cubitt, R. P., Drouvelis, M. & Gächter, S. Framing and free riding: emotional responses  
603 and punishment in social dilemma games. *Exp. Econ.* **14**, 254–272 (2011).
- 604 28. Raihani, N. J., Thornton, A. & Bshary, R. Punishment and cooperation in nature. *Trends*  
605 *Ecol. Evol.* **27**, 288–295 (2012).
- 606 29. Arechar, A. A., Gächter, S. & Molleman, L. Conducting interactive experiments online.  
607 *Exp. Econ.* **21**, 99–131 (2018).
- 608 30. Panchanathan, K. & Boyd, R. A tale of two defectors: the importance of standing for  
609 evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).

- 610 31. Gardner, A. & West, S. A. Cooperation and punishment, especially in humans. *Am. Nat.*  
611 **164**, 753–764 (2004).
- 612 32. Lehmann, L., Rousset, F., Roze, D. & Keller, L. Strong reciprocity or strong ferocity? A  
613 population genetic view of the evolution of altruistic punishment. *Am. Nat.* **170**, 21–36  
614 (2007).
- 615 33. Heckathorn, D. D. Collective action and the second-order free-rider problem. *Ration. Soc.*  
616 **1**, 78–100 (1989).
- 617 34. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the  
618 second-order free rider problem. *Nature* **432**, 499 (2004).
- 619 35. Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. Social learning promotes  
620 institutions for governing the commons. *Nature* **466**, 861–863 (2010).
- 621 36. Barclay, P. Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* **27**, 325–  
622 344 (2006).
- 623 37. dos Santos, M., Rankin, D. J. & Wedekind, C. The evolution of punishment through  
624 reputation. *Proc. R. Soc. Lond. B Biol. Sci.* **278**, 371–377 (2011).
- 625 38. dos Santos, M., Rankin, D. J. & Wedekind, C. Human cooperation based on punishment  
626 reputation. *Evolution* **67**, 2446–2450 (2013).
- 627 39. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103  
628 (2015).
- 629 40. Henrich, J. & Boyd, R. Why people punish defectors. *J. Theor. Biol.* **208**, 79–89 (2001).

- 630 41. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything  
631 else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
- 632 42. Kiyonari, T. & Barclay, P. Cooperation in social dilemmas: free riding may be thwarted  
633 by second-order reward rather than by punishment. *J. Pers. Soc. Psychol.* **95**, 826 (2008).
- 634 43. Fu, T., Ji, Y., Kamei, K. & Putterman, L. Punishment can support cooperation even when  
635 punishable. *Econ. Lett.* **154**, 84–87 (2017).
- 636 44. Szolnoki, A., Szabó, G. & Perc, M. Phase diagrams for the spatial public goods game  
637 with pool punishment. *Phys. Rev. E* **83**, 036101 (2011).
- 638 45. Traulsen, A., Röhl, T. & Milinski, M. An economic experiment reveals that humans  
639 prefer pool punishment to maintain the commons. *Proc R Soc B* rspb20120937 (2012).
- 640 46. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc.*  
641 *Psychol.* **51**, 110 (1986).
- 642 47. Ostrom, E. *Governing the Commons*. (Cambridge University Press, 2015).
- 643 48. Hilbe, C., Traulsen, A., Röhl, T. & Milinski, M. Democratic decisions establish stable  
644 authorities that overcome the paradox of second-order punishment. *Proc. Natl. Acad. Sci.*  
645 **111**, 752–756 (2014).
- 646 49. Szolnoki, A. & Perc, M. Effectiveness of conditional punishment for the evolution of  
647 public cooperation. *J. Theor. Biol.* **325**, 34–41 (2013).
- 648 50. FeldmanHall, O., Otto, A. R. & Phelps, E. A. Learning moral values: Another’s desire to  
649 punish enhances one’s own punitive behavior. *J. Exp. Psychol. Gen.* **147**, 1211 (2018).

- 650 51. Son, J.-Y., Bhandari, A. & FeldmanHall, O. Crowdsourcing punishment: Individuals  
651 reference group preferences to inform their own punitive decisions. (2019).  
652 doi:doi:10.31234/osf.io/ph3wn
- 653 52. Mahdi, N. Q. Pukhtunwali: Ostracism and honor among the Pathan hill tribes. *Ethol.*  
654 *Sociobiol.* **7**, 295–304 (1986).
- 655 53. Wiessner, P. Norm enforcement among the Ju/'hoansi Bushmen. *Hum. Nat.* **16**, 115–145  
656 (2005).
- 657 54. Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare.  
658 *Proc. Natl. Acad. Sci.* **108**, 11375–11380 (2011).
- 659 55. Güreker, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning  
660 institutions. *Science* **312**, 108–111 (2006).
- 661 56. Ertan, A., Page, T. & Putterman, L. Who to punish? Individual decisions and majority  
662 rule in mitigating the free rider problem. *Eur. Econ. Rev.* **53**, 495–511 (2009).
- 663 57. Casari, M. & Luini, L. Cooperation under alternative punishment institutions: An  
664 experiment. *J. Econ. Behav. Organ.* **71**, 273–282 (2009).
- 665 58. Casari, M. & Luini, L. Peer punishment in teams: expressive or instrumental choice? *Exp.*  
666 *Econ.* **15**, 241–259 (2012).
- 667 59. Kamei, K. Conditional punishment. *Econ. Lett.* **124**, 199–202 (2014).
- 668 60. Cheung, S. L. New insights into conditional cooperation and punishment from a strategy  
669 method experiment. *Exp. Econ.* **17**, 129–153 (2014).



- 670 61. Peysakhovich, A. & Rand, D. G. Habits of virtue: Creating norms of cooperation and  
671 defection in the laboratory. *Manag. Sci.* **62**, 631–647 (2015).
- 672 62. Albrecht, F., Kube, S. & Traxler, C. Cooperation and Norm Enforcement-The Individual-  
673 Level Perspective. *Journal of Public Economics* 1 (2017).
- 674 63. Selten, R. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens  
675 im Rahmen eines Oligopolexperimentes. in (Seminar für Mathemat. Wirtschaftsforschung  
676 u. Ökonometrie, 1965).
- 677 64. Bosman, R. & Van Winden, F. Emotional hazard in a power-to-take experiment. *Econ. J.*  
678 **112**, 147–169 (2002).
- 679 65. Falk, A., Fehr, E. & Fischbacher, U. Driving forces behind informal sanctions.  
680 *Econometrica* **73**, 2017–2030 (2005).
- 681 66. Hopfensitz, A. & Reuben, E. The importance of emotions for the effectiveness of social  
682 punishment. *Econ. J.* **119**, 1534–1559 (2009).
- 683 67. Nelissen, R. M. A. & Zeelenberg, M. Moral emotions as determinants of third-party  
684 punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgm. Decis. Mak.*  
685 543 (2009).
- 686 68. Gächter, S. & Herrmann, B. Reciprocity, culture and human cooperation: previous  
687 insights and a new cross-cultural experiment. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 791–  
688 806 (2009).
- 689 69. Gächter, S. & Herrmann, B. The limits of self-governance when cooperators get  
690 punished: Experimental evidence from urban and rural Russia. *Eur. Econ. Rev.* **55**, 193–  
691 210 (2011).

- 692 70. Elster, J. Norms of revenge. *Ethics* **100**, 862–885 (1990).
- 693 71. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really  
694 govern ourselves? *J. Public Econ.* **92**, 91–112 (2008).
- 695 72. Cialdini, R. B. & Trost, M. R. Social Influence: Social Norms, Conformity and  
696 Compliance. in *The handbook of social psychology, Vols. 1 and 2 (4th ed.)* (eds. Gilbert,  
697 D. T., Fiske, S. T. & Lindzey, G.) 151–192 (McGraw-Hill, 1998).
- 698 73. Thöni, C. & Volk, S. Conditional cooperation: Review and refinement. *Econ. Lett.* **171**,  
699 37–40 (2018).
- 700 74. Weber, T. O., Weisel, O. & Gächter, S. Dispositional free riders do not free ride on  
701 punishment. *Nat. Commun.* **9**, 2390 (2018).
- 702 75. Peysakhovich, A., Nowak, M. A. & Rand, D. G. Humans display a ‘cooperative  
703 phenotype’ that is domain general and temporally stable. *Nat. Commun.* **5**, 4939 (2014).
- 704 76. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence  
705 from a public goods experiment. *Econ. Lett.* **71**, 397–404 (2001).
- 706 77. Dohmen, T., Falk, A., Huffman, D. & Sunde, U. Homo reciprocans: Survey evidence on  
707 behavioural outcomes. *Econ. J.* **119**, 592–612 (2009).
- 708 78. Yamagishi, T. *et al.* Rejection of unfair offers in the ultimatum game is no evidence of  
709 strong reciprocity. *Proc. Natl. Acad. Sci.* **109**, 20364–20368 (2012).
- 710 79. Egloff, B., Richter, D. & Schmukle, S. C. Need for conclusive evidence that positive and  
711 negative reciprocity are unrelated. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E786–E786 (2013).

- 712 80. Eriksson, K., Cownden, D., Ehn, M. & Strimling, P. ‘Altruistic’ and ‘Antisocial’  
713 Punishers are One and the Same. *Rev. Behav. Econ.* **1**, 209–221 (2014).
- 714 81. Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G. & Cudré-Mauroux, P. The  
715 dynamics of micro-task crowdsourcing: The case of amazon mturk. in *Proceedings of the*  
716 *24th international conference on world wide web* 238–247 (International World Wide  
717 Web Conferences Steering Committee, 2015).
- 718 82. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on amazon mechanical  
719 turk. *Judgm. Decis. Mak.* **5**, 411–419 (2010).
- 720 83. Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: Conducting  
721 experiments in a real labor market. *Exp. Econ.* **14**, 399–425 (2011).
- 722 84. Berinsky, A. J., Huber, G. A. & Lenz, G. S. Evaluating online labor markets for  
723 experimental research: Amazon. com’s Mechanical Turk. *Polit. Anal.* **20**, 351–368  
724 (2012).
- 725 85. Balafoutas, L., Nikiforakis, N. & Rockenbach, B. Direct and indirect punishment among  
726 strangers in the field. *Proc. Natl. Acad. Sci.* **111**, 15924–15927 (2014).
- 727 86. Balafoutas, L., Nikiforakis, N. & Rockenbach, B. Altruistic punishment does not increase  
728 with the severity of norm violations in the field. *Nat. Commun.* **7**, 13327 (2016).
- 729 87. Raihani, N. J. & Bshary, R. The evolution of punishment in n-player public goods games:  
730 A volunteer’s dilemma. *Evolution* **65**, 2725–2728 (2011).
- 731 88. Rand, D. G. & Nowak, M. A. The evolution of antisocial punishment in optional public  
732 goods games. *Nat. Commun.* **2**, 434 (2011).

- 733 89. Garcia, J. & Traulsen, A. Leaving the loners alone: Evolution of cooperation in the  
734 presence of antisocial punishment. *J. Theor. Biol.* **307**, 168–173 (2012).
- 735 90. McCabe, C. M. & Rand, D. G. Coordinated punishment does not proliferate when  
736 defectors can also punish cooperators. *J. Commun. Res.* **6**, (2014).
- 737 91. Huang, F., Chen, X. & Wang, L. Conditional punishment is a double-edged sword in  
738 promoting cooperation. *Sci. Rep.* **8**, 528 (2018).
- 739 92. Giamattei, M., Molleman, L., Seyed Yahosseini, K. & Gächter, S. LIONESS Lab – a free  
740 web-based platform for conducting interactive experiments online. Available at SSRN:  
741 <https://ssrn.com/abstract=3329384> or <http://dx.doi.org/10.2139/ssrn.3329384>. (Social  
742 Science Research Network, 2019).
- 743 93. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data*. (MIT press,  
744 2010).

745

746 **Acknowledgements.** We thank B. Beranek, P. van den Berg, J. Schulz, T. Weber and O.  
747 Weisel for insightful comments and useful discussions. This work was supported by the  
748 European Research Council [grant number ERC-AdG 295707 COOPERATION], the  
749 Economic and Social Research Council [grant numbers ES/K002201/1 and ES/P008976/1],  
750 the Nottingham School of Economics and the Center of Adaptive Rationality, Max Planck  
751 Institute for Human Development Berlin. L.M. was further supported by the Open Research  
752 Area grant ASTA [grant number 176] and the Amsterdam Brain and Cognition Project Grant  
753 2018. The funders had no role in study design, data collection and analysis, decision to  
754 publish, or preparation of the manuscript.

755

756 **Author contributions.** L.M., F.K., C.S. and S.G. designed the study, L.M. and F.K. collected  
757 and analysed the data, L.M., F.K., C.S. and S.G. wrote the paper.

758

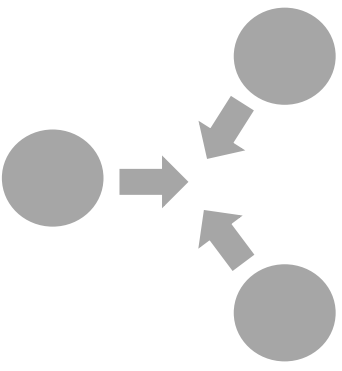
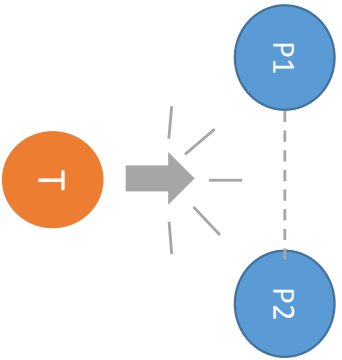
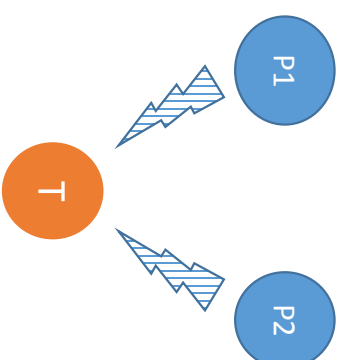
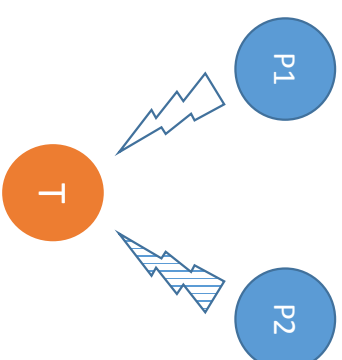
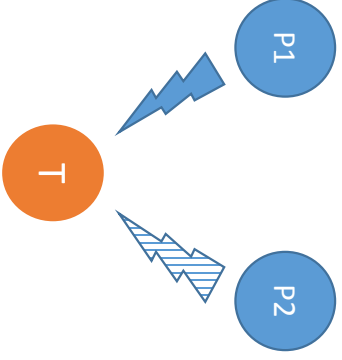
759 **Competing interests.** The authors declare no competing interests.

760

761 **Materials & Correspondence.** Please contact L.M. for correspondence and material  
762 requests.

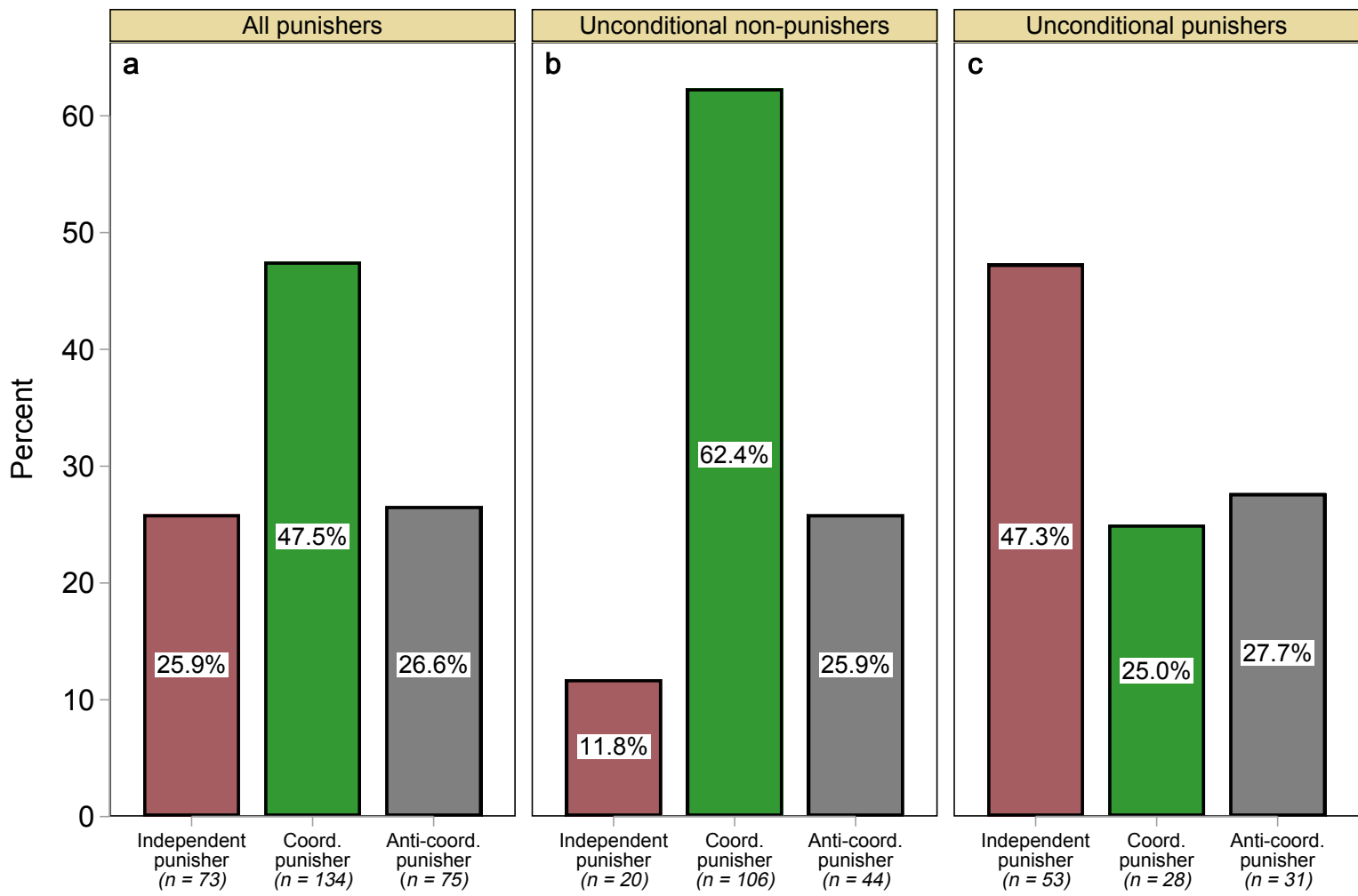
763 **Figure legends**

764 **Figure 1 | Experimental sequence. a**, Participants are assigned to a group of three (grey  
765 circles) and make a binary decision in a public goods game (PGG; grey arrows). **b**, Roles are  
766 randomly allocated among group members: two Punishers (P1 and P2; blue circles) and one  
767 Target (T; orange circle); T's PGG decision is revealed to P1 and P2 but no information about  
768 the other Punisher's PGG decision is provided; Punishers are informed about the steps  
769 comprising the punishment procedure. **c**, P1 and P2 each make an unconditional binary  
770 decision whether or not to punish T (blue hatched bolts). **d,e**, P1 and P2 make conditional  
771 binary punishment decisions; shown is the situation from the perspective of P2. First, they  
772 decide whether they would punish in case the other P player chose to not punish (**d**; empty  
773 bolt) in step c; or to punish (**e**; solid bolt) in step c. Once all decisions have been made, P1 or  
774 P2 is randomly selected and their unconditional punishment decision (step c) is implemented,  
775 along with the corresponding conditional decision of the other P player (step d or e). For  
776 experimental instructions, see Supplementary Methods.

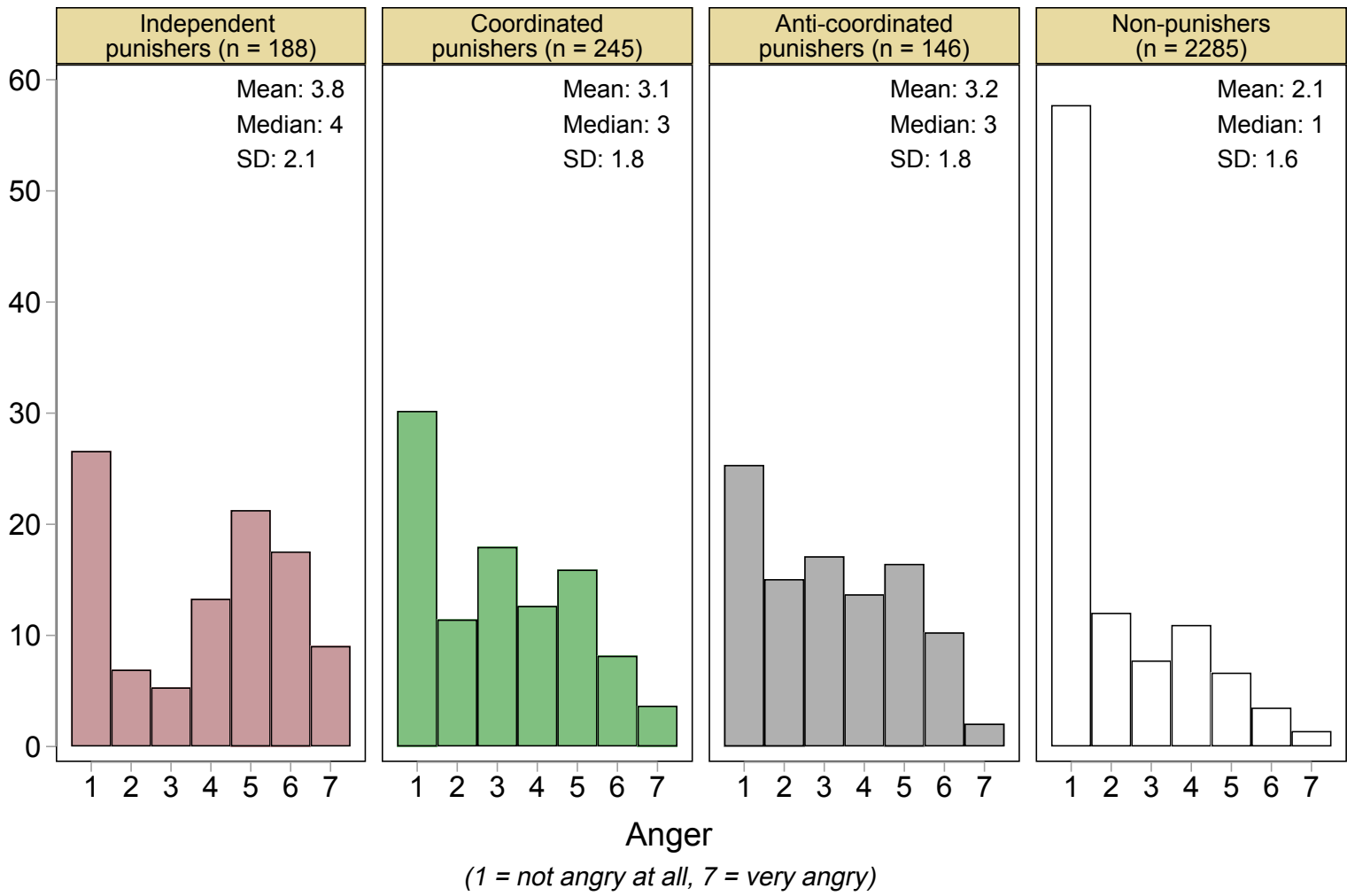
<p><b>a.</b> all group members: cooperate?</p>	
<p><b>b.</b> assign roles, reveal cooperation choice of T</p>	
<p><b>c.</b> P1 and P2: punish T?</p>	
<p><b>d.</b> P1 and P2: punish T if other P does <i>not</i> punish?</p>	
<p><b>e.</b> P1 and P2: punish T if other P <i>does</i> punish?</p>	

777 **Figure 2 | Distribution of punishment types by decision in unconditional punishment**  
778 **stage.** Bars show data from the strategy method, restricted to those individuals who punished  
779 at least once ( $N = 282$  out of the total of  $N = 1,336$ ). **a**, All punishers. **b**, Only those who did  
780 *not* punish in the unconditional punishment stage. **c**, Only those who did punish in the  
781 unconditional punishment stage.

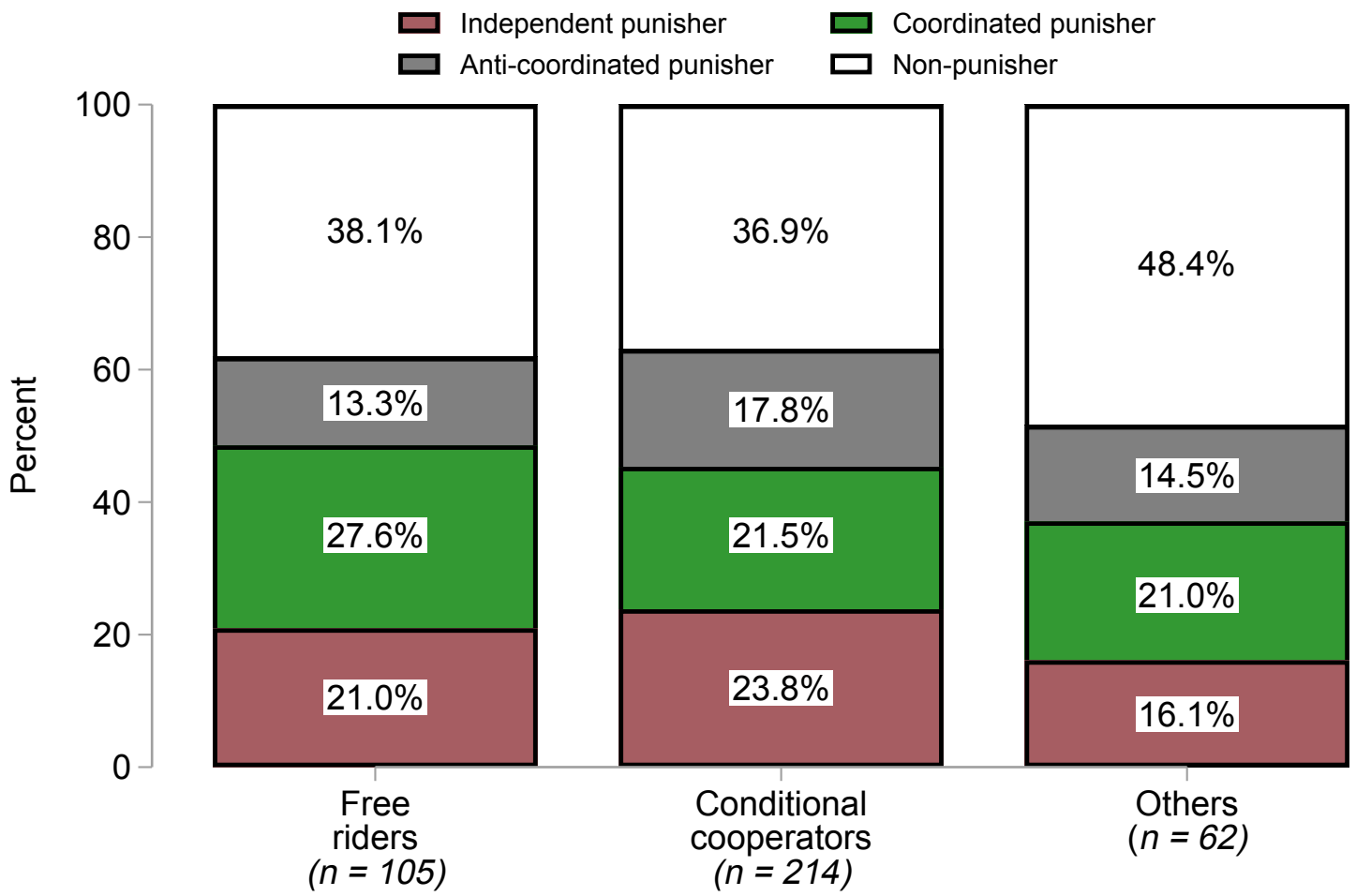




782 **Figure 3 | Anger levels per punishment type.** Panels show distributions of Punishers' self-  
783 reported ratings of anger when they learned about the PGG decision of their Target. The  
784 distribution of anger levels differ significantly across punishment types (Kruskal-Wallis test:  
785  $\chi^2 = 198.29$ , d.f. = 3,  $P < 0.001$ ).



786 **Figure 4 | The (lack of) correlation between individuals' preferences for conditional**  
787 **cooperation and conditional punishment.** Stacked bars show the distribution of punishment  
788 types as measured in the primary experiment, separated by conditional cooperation type as  
789 measured in the follow-up experiment.

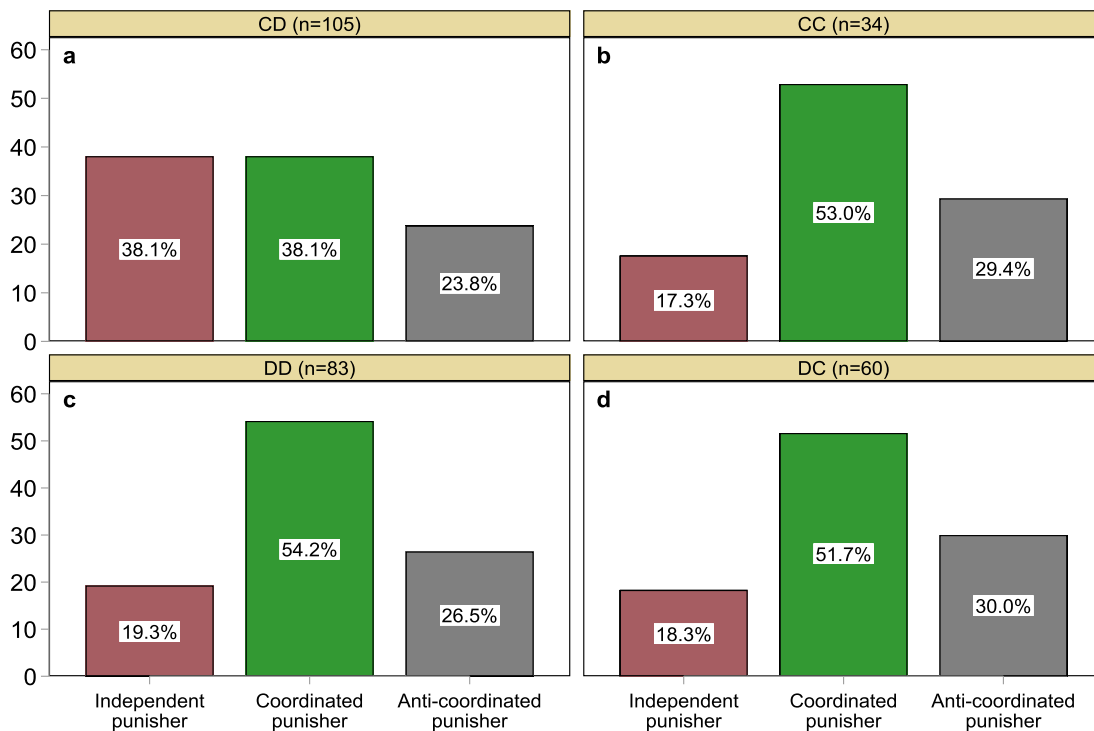


790 **Tables**

791 **Table 1 | Behavioural determinants of conditional punishment.** Coefficients from logistic  
792 generalized linear mixed models fitted to Punishers' decisions whether or not to punish the  
793 Target (1 if yes, 0 if no). 'Other punishes' is a dummy variable with value 1 in case the other  
794 Punisher punishes and 0 otherwise. 'Unconditional punishment decision' is a dummy variable  
795 indicating whether a participant punished unconditionally (=1) or not (=0). 'Unconditional  
796 punishment × Other punishes' is an interaction term between the two variables. 'Target  
797 cooperated' is a dummy variable with value 1 if the Target cooperated and 0 if she defected.  
798 'Target cooperated' × Other punishes' is an interaction term between this variable and others'  
799 punishment decision to test whether coordinated punishment varies with the Target's  
800 cooperation decision. Additional regressions including controls for gender and age revealed  
801 that neither of these demographic items has a significant effect. Including gender and age did  
802 not significantly change any of the effects reported above. We cluster standard errors at the  
803 individual level, correcting for repeated observations<sup>93</sup>. 95% confidence interval in brackets  
804 and p-values in parentheses.

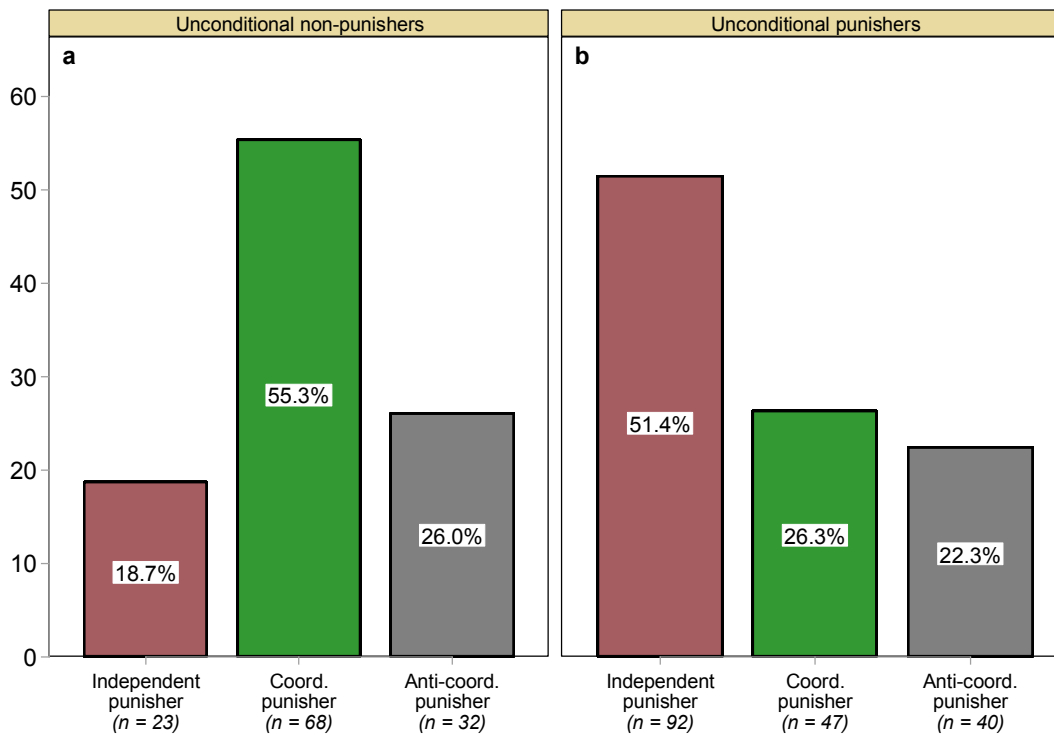
Dependent variable:	Punish (1 if yes, 0 otherwise)	
	(1)	(2)
Other punishes <i>(1 if other Punisher punished, 0 otherwise)</i>	0.734 ( $<0.001$ ) [0.447 – 1.022]	0.380 ( $<0.001$ ) [0.155 – 0.605]
Unconditional punishment <i>(1 if yes, 0 otherwise)</i>	3.074 ( $<0.001$ ) [2.666 – 3.481]	
Unconditional punishment $\times$ Other punishes	-0.814 (0.001) [-1.306 – -0.322]	
Target cooperated <i>(1 if Target cooperated, 0 otherwise)</i>		-0.831 ( $<0.001$ ) [-1.120 – -0.463]
Target cooperated $\times$ Other punishes		0.039 (0.854) [-0.373 – 0.450]
Constant	-2.862 ( $<0.001$ ) [-3.114 – -2.610]	-1.750 ( $<0.001$ ) [-1.960 – -1.541]
Number of observations	2,672	2,672
Number of participants	1,336	1,336

## Supplementary Figures

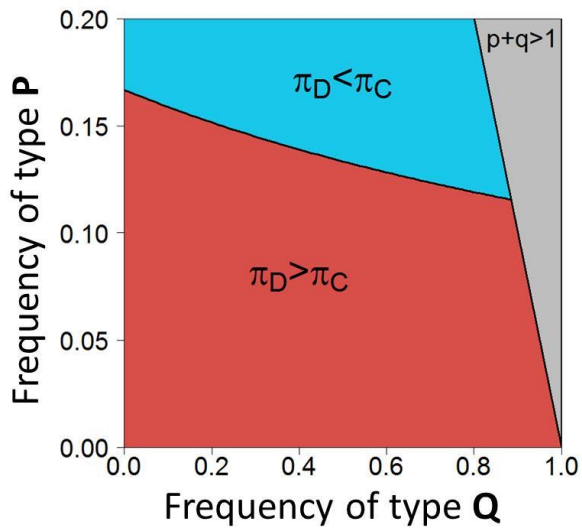


**Supplementary Figure 1 | Conditional punishment types broken down by outcome of the Public Goods Game (PGG).** Bars reflect data from the strategy method, and includes only those individuals who punished at least once in the conditional stage ( $N = 282$  out of the total of  $N = 1,336$ ). **a**, Punisher cooperated, Target defected (CD). **b**, Punisher and Target cooperated (CC). **c**, Punisher and Target defected (DD). **d**, Punisher defected, Target cooperated (DC). We observe that overall, coordinated punishers tend to be most frequent (52% - 54%). The exception to this general pattern is the case when the Punisher cooperated and the Target defected (panel a). For that outcome, the frequency of *independent punishers* is higher than in the other cases ( $\chi^2(1) = 12.99$ ,  $P < 0.001$ ,  $\varphi_c = 0.21$ ), and the frequency of *coordinated punishers* is lower ( $\chi^2(1) = 5.96$ ,  $P = 0.015$ ,  $\varphi_c = 0.15$ ; see Supplementary Table 2 below for a more detailed statistical analysis). For the other outcomes of the PGG, the fraction of *independent punishers* varies between 17% and 19%, and the fraction of *anti-coordinated punishers* varies between 27% and 30%. Overall, the distribution of types is remarkably similar across these situations ( $\chi^2(4) = 0.25$ ,  $P = 0.993$ ,  $\varphi_c = 0.03$ ).

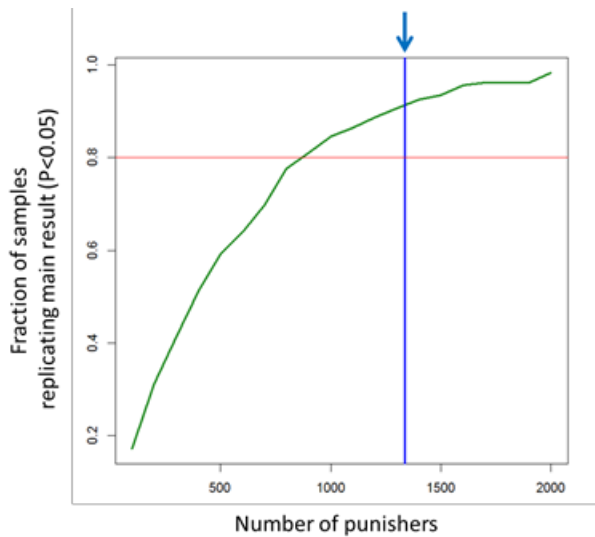




**Supplementary Figure 2 | Distribution of punishment types in the replication study.** Bars report data from the strategy method, restricted to those individuals who punished at least once ( $N = 302$  out of the total of  $N = 1,544$ ). **a**, Only participants who did not punish unconditionally. **b**, Only participants who did punish unconditionally.



**Supplementary Figure 3 | Relative payoffs of cooperation and defection depend on the relative frequencies of punishment types.** This levelplot summarizes the analysis of the model presented in the Supplementary Results. The horizontal and vertical axes show the frequency of punishment types  $Q$  and  $P$ , respectively. For each combination of frequencies of these punishment types, the colours indicate whether expected payoffs of cooperation are lower (red) or higher (blue) than expected payoffs of defection. Theoretically impossible cases (where the total frequency of  $P$  and  $Q$  would exceed 1) are shown in grey. The black line separating the red and blue areas indicates the values for which  $\pi_C = \pi_D$ . That is,  $p = \frac{c}{k(2+q)}$ . For this illustration, we assume that  $c=1$  and  $k=3$ . For full model details, see Supplementary Results, section “Effects of coordinated punishment on relative payoffs of cooperation and defection”.



**Supplementary Figure 4 | Power analysis for the additional study.** Based on the data of our original experiments we calculated the probability to reproduce our main result (“people were more likely to punish their peers when the other punisher did so as well”; cf. Table 1, Model 1) for given sample sizes. For a range of possible sample sizes  $N$ , we sampled  $N$  participants from our data (with replacement) who were in the role of Punisher, and ran Model (1) on that sample. For each  $N$  we repeated that 1,000 times and tracked the number of replications in which the main effect was positive and significant at the  $P < 0.05$  level. The green line indicates the expected probability of detecting a significant result (at the 5% significance level) as a function of the number of punishers in our sample. The vertical blue line indicates the sample size in our original submission; the horizontal red line indicates 80% probability of replicating our main result. The red and green line intersect at  $N \approx 800$ , indicating that this number of Punishers is expected to have 80% power. Note that in our experimental setup, only  $2/3$  of the participants are in the Punisher role, so we would require  $(800 \cdot 2/3 =)$  1,200 participants. In our additional study, counterbalanced the order of the punishment decisions, aiming for 1,200 participants in both ‘orders’, yielding 2,400 participants in total.

## Supplementary Tables

**Supplementary Table 1 | Determinants of conditional punishment in our replication study, and for all data pooled.** Coefficients from logistic generalized linear mixed models fitted to Punishers' decisions whether or not to punish the Target (1 if yes, 0 if no). The models presented in this table mirror those from Table 1 of the main text. Models (1) and (2) use data from our replication study. Models (3) and (4) use all data, pooling across our main and our replication study. 'Other punishes' is a dummy variable with value 1 in case the other Punisher punishes and 0 otherwise. 'Unconditional punishment decision' is a dummy variable indicating whether a participant punished unconditionally (=1) or not (=0). 'Unconditional punishment × Other punishes' is an interaction term between the two variables. 'Target cooperated' is a dummy variable with value 1 if the Target cooperated and 0 if she defected. 'Target cooperated' × Other punishes' is an interaction term between this variable and others' punishment decision to test whether coordinated punishment varies with the Target's cooperation decision. Additional regressions including controls for gender and age revealed that neither of these demographic items has a significant effect. Including gender and age did not significantly change any of the effects reported above. We cluster standard errors at the individual level, correcting for repeated observations. 95% confidence interval in brackets and p-values in parentheses.

Dependent variable:	Punish (1 if yes, 0 otherwise)			
	(1)	(2)	(3)	(4)
Other punishes <i>(1 if other Punisher punished, 0 otherwise)</i>	0.533 ( $<0.001$ ) [0.241 – 0.825]	0.227 (0.009) [0.055 – 0.397]	0.643 ( $<0.001$ ) [0.437 – 0.848]	0.294 ( $<0.001$ ) [0.156 – 0.431]
Unconditional punishment <i>(1 if yes, 0 otherwise)</i>	3.407 ( $<0.001$ ) [3.032 – 3.782]		3.246 ( $<0.001$ ) [2.973 – 3.519]	
Unconditional punishment × Other punishes	-0.408 (0.067) [-0.844 – 0.028]		-0.600 ( $<0.001$ ) [-0.925 – 0.276]	
Target cooperated <i>(1 if Target cooperated, 0 otherwise)</i>		-1.342 ( $<0.001$ ) [-1.705 – -0.979]		-1.103 ( $<0.001$ ) [-1.360 – -0.846]
Target cooperated × Other punishes		0.087 (0.658) [-0.299 – 0.474]		0.075 (0.598) [-0.205 – 0.355]
Constant	-3.129 ( $<0.001$ ) [-3.399 – -2.859]	-1.500 ( $<0.001$ ) [-1.679 – -1.320]	-2.994 ( $<0.001$ ) [-3.179 – -2.810]	-1.611 ( $<0.001$ ) [-1.747 – -1.475]
Number of observations	3,088	3,088	5,760	5,760
Number of participants	1,544	1,544	2,880	2,880

**Supplementary Table 2 | Determinants of conditional punishment.** Coefficients from logistic generalized linear mixed models fitted to Punishers' decisions whether or not to punish the Target (1 if yes, 0 if no). Model (1) uses data from our main experiment. Model (2) uses data from our replication study. Model (3) provides the results using all data. 'Other punishes' is a dummy variable that takes the value 1 in case the other Punisher punishes and 0 otherwise. Dummy variables CD (DC) indicate a situation in which a Punisher cooperated (defected) and the Target defected (cooperated), while CC indicates a situation where both players cooperated. The case where both the Punisher and the Target defected is the baseline. We include interaction terms between these variables and others' punishment decision to investigate whether coordinated punishment is more or less prevalent for the different outcomes of the PGG stage of the game. We cluster standard errors at the individual level, correcting for repeated observations. 95% confidence interval in brackets and p-values in parentheses.

Dependent variable:	Punish (1 if yes, 0 otherwise)		
	(1)	(2)	(3)
Other punishes (1 if yes, 0 if no)	0.547 (0.005) [0.165 – 0.929]	0.272 (0.032) [0.024 – 0.521]	0.382 (<0.001) [0.169 – 0.594]
CD (1 if self cooperates and other defects, 0 otherwise)	0.723 (0.001) [0.291 – 1.155]	0.473 (0.010) [0.112 – 0.834]	0.566 (<0.001) [0.290 – 0.841]
DC (1 if self defects and other cooperates, 0 otherwise)	-0.207 (0.426) [-0.715 – 0.302]	-0.986 (<0.001) [-1.474 – -0.498]	-0.632 (<0.001) [-0.979 – -0.285]
CC (1 if self and other cooperates, 0 otherwise)	-0.758 (0.014) [-1.363 – -0.152]	-1.271 (<0.001) [-1.823 – -0.719]	-1.054 (<0.001) [-1.459 – -0.648]
Other punishes × CD	-0.282 (0.243) [-0.755 – 0.191]	-0.087 (0.622) [-0.431 – 0.258]	-0.159 (0.266) [-0.439 – 0.121]
Other punishes × DC	-0.133 (0.654) [-0.714 – 0.448]	0.136 (0.615) [-0.393 – 0.665]	0.029 (0.883) [-0.355 – 0.412]
Other punishes × CC	-0.114 (0.744) [-0.795 – 0.568]	-0.101 (0.729) [-0.671 – 0.470]	-0.076 (0.732) [-0.513 – 0.361]
Constant	-2.143 (<0.001) [-2.479 – -1.807]	-1.728 (<0.001) [-1.987 – -1.468]	-1.899 (<0.001) [-2.104 – -1.693]
Number of observations	2,672	3,088	5,760
Number of participants	1,336	1,544	2,880

**Supplementary Table 3 | Responses to the extended questionnaire in the replication study.**

Numbers show mean [median] responses, separated by punishment types. Numbers in brackets are standard deviations.

Statement	Agreement (1 = ‘disagree strongly’, 7 = ‘agree strongly’)			
	Non-punisher (NP)	Independent punisher (IP)	Coordinated punisher (CP)	Anti-coordinated punisher (ACP)
1. When making my decisions, I was unsure what to do	3.2 [3] (1.8)	3.7 [4] (1.7)	4.0 [4] (1.6)	4.4 [5] (1.6)
2. When making my decisions, I was unsure what was the appropriate thing to do	3.2 [3] (1.8)	3.8 [4] (1.8)	4.2 [4] (1.5)	4.5 [5] (1.6)
3. I did not want to let Blue down in case they chose to punish	3.9 [4] (1.8)	4.5 [5] (1.6)	4.9 [5] (1.5)	4.5 [4] (1.4)
4. I wanted to reduce Red’s earnings myself	2.4 [2] (1.5)	4.8 [5] (1.6)	4.1 [4] (1.6)	4.0 [4] (1.5)
5. I did not want to earn less than Blue	4.3 [4] (1.6)	4.6 [5] (1.7)	4.9 [5] (1.2)	4.5 [4] (1.4)
6. I did not want to reduce my own earnings	6.3 [7] (1.1)	5.4 [5] (1.5)	5.9 [6] (1.0)	5.6 [6] (1.4)
7. I did not want to reduce Red’s earnings by too much	5.0 [5] (1.7)	3.2 [3] (1.8)	4.0 [4] (1.7)	3.8 [4] (1.6)

**Supplementary Table 4 | Responses to the extended questionnaire in the replication study.**

Numbers shows mean [median] responses, separated by punishment types. Numbers in brackets are standard deviations.

	Non-punisher (NP)	Independent punisher (IP)	Coordinated punisher (CP)	Anti-coordinated punisher (ACP)
Positive reciprocity	5.5 [5.7] (1.4)	5.6 [6.0] (1.5)	5.4 [5.7] (1.6)	5.5 [5.7] (1.0)
Negative reciprocity	2.7 [2.7] (1.5)	3.1 [3.0] (1.7)	3.2 [3.0] (1.7)	2.9 [2.7] (1.4)
Risk	5.9 [6] (2.4)	6.6 [7.0] (2.5)	6.7 [7.0] (2.4)	6.6 [7.0] (2.5)
Extraversion	3.4 [3.5] (1.6)	3.9 [4.0] (1.7)	3.8 [4.0] (1.4)	3.3 [3.3] (1.3)
Agreeableness	5.1 [5.0] (1.3)	5.1 [5.0] (1.4)	5.1 [5.0] (1.2)	4.9 [5.0] (1.3)
Conscientiousness	5.4 [5.5] (1.2)	5.5 [5.5] (1.2)	5.2 [5.5] (1.3)	5.2 [5.5] (1.3)
Emotional stability	4.7 [5.0] (1.6)	4.9 [5.0] (1.6)	4.5 [4.5] (1.4)	4.6 [4.5] (1.4)
Openness	5.1 [5.0] (1.2)	5.2 [5.2] (1.2)	4.9 [5.0] (1.3)	5.0 [5.0] (1.3)

# Supplementary Results

## Effects of coordinated punishment on relative payoffs of cooperation and defection

To explore how coordinated punishment could affect cooperation, we derive a simple model with the most common punishment preferences identified in our experiment. Our aim here is, with the help of some simplifying assumptions, to illustrate how the frequency of these punishment preferences in a population affects the relative payoffs of cooperation and defection. Note that we do not aim to explore how these punishment preferences may have emerged through evolutionary processes (e.g. through natural selection or social learning); we simply aim to explore the *consequences* for expected payoffs of cooperation and defection, under given relative frequencies of punishment preferences in the population.

Our experimental results indicate that the most common punishment preferences are: (1) to punish in the unconditional decision as well as in both conditional decisions (main text Figure 2c; red bar; let us denote this as punishment type  $P$ ); and (2) to not-punish in the unconditional decision, and only punish conditional upon the other Punisher initialising punishment (in their unconditional punishment decision; main text Figure 2b; green bar; let us denote this as punishment type  $Q$ ). We assume that individuals not belonging to either of these types ( $P$  or  $Q$ ) do not punish at all.

For exploration purposes, we consider a population of infinite size, in which individuals are randomly matched to interact in groups of three. The structure of interactions is similar to our experiment. First, individuals interact in a binary Public Goods Game (PGG) in which they make a cooperation decision ('cooperate' or 'defect'). Subsequently, we randomly assign roles to the group members: two individuals will act a 'Punisher', and remaining one as the 'Target'. Punishers each make a binary decision whether or not to punish the Target. For our purposes, we first focus on punishment directed at Targets who defected in the PGG. At the end of this analysis we consider (anti-social) punishment towards cooperators.

As in the experiment, punishment takes place in two stages. (1) an 'unconditional' stage in which one of the Punishers observes the cooperation decision of the Target and independently chooses whether or not to punish them; (2) a 'conditional' stage in which the remaining punisher observes that unconditional punishment decision and decides whether or not to punish the Target.

Punishment takes place according to the punisher's 'type'. For the sake of simplicity, we only focus on the impact of punishment on the relative expected payoffs of cooperation and defection, and ignore any costs that conducting punishment may impose on Punishers.



To examine how coordinated punishment may affect the relative payoffs of cooperation and defection, we calculate how relative expected payoffs of these two decisions vary with the frequency of coordinated punishers in the population. In the PGG stage, all group members receive a benefit  $b$  of the cooperation of all group members. Defectors avoid the cost  $c$  of cooperating, so defection pays off better than cooperation. This ‘cost of cooperation’ can be offset if defectors receive punishment from their peers. Targets incur a cost  $k$  for each peer that punishes them.

To compare the expected relative payoffs of cooperators and defectors, we need some notation. Let  $p$  denote the fraction of individuals in the population who have punishment preference  $P$  (see above), and punish both unconditionally and also in the conditional stage. Further, let  $q$  denote the fraction of individuals in the population who have punishment preference  $Q$ , who do not punish unconditionally, but in the conditional stage only if the other Punisher has punished unconditionally. Individuals can have only one type of punishment preference, so  $0 \leq p + q \leq 1$ . We assume that the other  $(1 - p - q)$  do not punish. The expected payoffs for cooperation ( $\pi_C$ ) and for defection ( $\pi_D$ ) can be written as

$$\pi_C = b - c, \text{ and}$$

$$\pi_D = b - k \cdot [p \cdot (2 + q)]$$

The term between the square brackets is the expected number of individuals that punish a defector. It is calculated as follows. We first take the probability of unconditional punishment, which is simply equal to  $p$ , the frequency of punishment type  $P$  in the population. Then we calculate the probability of conditional punishment. This punishment can be meted out by an individual of punishment type  $P$  or  $Q$ : this probability is given by the sum of  $p$  (again, the frequency of  $P$  who punish independently) *plus* the probability that unconditional punishment has taken place (again  $p$ ) times  $q$  (the frequency of coordinated punishers  $Q$ ). We obtain the term between the square brackets by factoring out  $p$  in  $p + p + p \cdot q$ .

To illustrate how coordinated punishers (type  $Q$ ) affect the relative payoffs of cooperation and defection, we can derive a minimal frequency of type  $P$  for which cooperation has higher expected payoffs than defection; that is  $\pi_C > \pi_D$ .

$$\pi_C > \pi_D \text{ if } p > \frac{c}{k(2+q)}$$

Supplementary Figure 3 shows that types that do not punish unconditionally, but will engage in coordinated punishment can substantially increase the range of conditions for which cooperation leads to higher expected payoffs than defection. In other words, for cooperation to thrive, a

population requires a considerably lower frequency of individuals who punish free riding when there are individuals around who would not punish unconditionally, but who are ready to step in as soon as they observe punishment taking place.

These results are in line with a more detailed analysis showing that coordinated punishment can promote the evolutionary emergence of costly cooperation<sup>8</sup>. For simplicity, our analysis so far has only focused on punishment of defectors. Empirical evidence from a range of previous studies<sup>9–11</sup> as well as observations from our experiment (cf Supplementary Figure 1) indicate that at times, punishment is aimed at cooperators. Such anti-social punishment can have strongly detrimental consequences for cooperation<sup>9,12–14</sup>. Moreover, if individuals tend to coordinate their punishment towards cooperators, coordinated punishment may no longer be able to promote cooperation<sup>15,16</sup>.

In our model, anti-social punishment can be accounted for by writing the expected payoffs for cooperation and defection as:

$$\pi_C = b - c - k \cdot [p' \cdot (2 + q)]$$

$$\pi_D = b - k \cdot [p \cdot (2 + q)]$$

where  $p'$  indicates the frequency of individuals who (anti-social) punish cooperators. Note that we assume that individuals with type  $Q$ , who coordinate their punishment, do not distinguish whether the target of punishment had defected or cooperated.

The conditions for cooperation to have higher expected payoffs than defection are then defined by:

$$\pi_C > \pi_D \text{ if } (p - p') > \frac{c}{k(2+q)}$$

This shows that anti-social punishment decreases the scope for cooperation to thrive, and that in the context of this simple model, the relative expected payoffs of cooperation and defection reflect the frequency differences between pro-social ( $p$ ) and anti-social ( $p'$ ) punishers.

## Questionnaire targeted at motivations underlying punishment preferences

Here we provide details on the extended questionnaire from the replication study probing possible motivations underlying conditional punishment preferences. To measure these motivations, participants in the role of Punisher were asked to think back to their decisions in the conditional punishment stage (*cf.* Figure 1d,e of the main text). Then they had to use a 7-point scale to indicate their agreement with each of seven statements, where 1 means ‘disagree strongly’ and 7 means ‘agree strongly’. Each of these statements was designed to measure a candidate motivation for punishment, and/or conditioning punishment on the punishment of others. (NB: remember that on the Punishers’ experimental screens, the other Punisher was referred to as ‘Blue’ and the Target was referred to as ‘Red’.)

The seven statements about the conditional punishment decisions were [with, in square brackets, the underlying motivation they aim to tap]:

1. When making my decisions, I was unsure what to do [requiring ‘social proof’<sup>1,2</sup>]
2. When making my decisions, I was unsure what was the appropriate thing to do [concerns for social appropriateness or legitimacy<sup>3,4</sup>]
3. I did not want to let Blue down in case they chose to punish [positive reciprocity towards the other Punisher]
4. I wanted to reduce Red’s earnings myself [thirst for revenge]
5. I did not want to earn less than Blue [disadvantageous inequality aversion with regard to the other Punisher]
6. I did not want to reduce my own earnings [monetary concerns]
7. I did not want to reduce Red’s earnings by too much [wanting to apply a fitting punishment (i.e. applying a sanction of proper magnitude)]

In the Supplementary Table 3, we show the mean and the median responses to these statements, broken down by conditional punishment type.

In the main text we focus on the different possible underlying motivations of those punishers who punish at least once in their conditional punishment decision, i.e., independent punisher (IP), coordinated punisher (CP), and anti-coordinated punisher (ACP). Here we complement these results by comparing these types with those who never punish, i.e., non-punishers (NP; Supplementary Table 3). Relatively speaking, non-punishers were less unsure about what to do (Mann Whitney U (MWU) test,  $z = 6.59$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.17$ ) and less unsure about what the appropriate thing to do was (MWU test,  $z = 8.21$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.21$ ). Furthermore, they were less concerned about letting the other punisher [Blue] down (MWU test,  $z = 6.93$ , d.f. = 1,  $P < 0.001$ ,  $r =$

0.18) and they reported to be less driven by a thirst of revenge (MWU test,  $z = 16.67$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.42$ ). They further scored significantly higher on statements 6 (MWU test,  $z = 8.32$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.21$ ) and 7 (MWU test,  $z = 11.68$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.30$ ), indicating that concerns about their own and the target's earnings played an important role for not meting out any punishment.

In addition to these possible motivations underlying participants' behaviour in the particular punishment situation they encountered in our experiment, we also measured personality-level characteristics using established psychological scales. In particular, we administered a brief measurement of the big five personality scale<sup>5</sup>, gauged general dispositions towards positive and negative reciprocity<sup>6</sup> and elicited risk preferences<sup>7</sup>.

Supplementary Table 4 shows mean [median] scores of these personality scales broken down by punishment type. While the different punishment types do not seem to differ with regard to dispositions towards positive reciprocity (Kruskal-Wallis (KW) test,  $\chi^2 = 6.03$ , d.f. = 3,  $P = 0.110$ ), we find them to differ with regard to their attitudes towards negative reciprocity (KW-test,  $\chi^2 = 12.48$ , d.f. = 3,  $P = 0.006$ ).

A closer inspection reveals that this effect is driven by non-punishers who display a significantly lower disposition towards negative reciprocity than any other type (MWU-test,  $z = 3.44$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.09$ ), while there is no pronounced difference among the remaining types (KW-test,  $\chi^2 = 0.83$ , d.f. = 2,  $P = 0.658$ ). A similar pattern can be observed with regard to risk attitudes. Non-punishers are significantly less willing to take risks than the other three types (MWU-test,  $z = 4.71$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.13$ ), but there is no difference between the latter (KW-test,  $\chi^2 = 0.03$ , d.f. = 2,  $P = 0.983$ ,  $r = 0.17$ ).

With regard to the big five personality dimensions, the only notable difference across punishment types is with regard to extraversion (KW-test,  $\chi^2 = 17.18$ , d.f. = 3,  $P < 0.001$ ). In particular, independent punisher and coordinated punisher score higher than non-punisher and anti-coordinated punisher (MWU-test,  $z = 3.79$ , d.f. = 1,  $P < 0.001$ ,  $r = 0.08$ ), with no difference between the former (MWU-test,  $z = 0.43$ , d.f. = 1,  $P = 0.666$ ,  $r = 0.01$ ) or the latter two (MWU-test,  $z = 0.46$ , d.f. = 1,  $P = 0.648$ ,  $r = 0.03$ ). No significant differences are observed with regard to the other personality characteristics (KW-tests, all  $P > 0.173$ ).

## Supplementary Methods: Experimental materials

Below we show on-screen instructions as displayed to participants. We start with the conditional punishment experiment. Then we show the follow-up experiment on conditional cooperation. Participants could not navigate the experimental pages at will. Each time they pressed a button, the browser history was automatically overwritten. See Aréchar et al (2018) ‘*Conducting interactive experiments online*’<sup>17</sup> for details.

NB: ‘== [notes] ==’ indicates a new screen, with occasional notes in brackets. Experimental code for both experiments are available upon request from the corresponding author.

### Conditional punishment experiment

Differences between our main and the replication study are highlighted throughout, in **purple**. These differences were (i) addition of control questions prior to the punishment stages of the game; (ii) rewording of instructions to accommodate the counterbalanced design (so, half of the punishers made their ‘conditional decision’ before their ‘unconditional decision’); and (iii), addition of the questionnaires probing candidate motivation underlying conditional punishment preferences.

== == ==

### Welcome!

In this HIT you will be interacting with two other real MTurkers who also accepted this HIT, and who are participating **at the same time** as you. It is therefore important that you complete this HIT **without interruptions**. Including the time to read these instructions, the HIT will take about 8 minutes to complete.

During this HIT you can earn Points. The number of Points you earn depends on your decisions and the decisions of the other participants. You receive 4 Points to start with. At the end of the HIT your Points will be converted into real money (**10 Points = \$1,00**). In addition, you will receive \$0.50 on top of however much you earn during the HIT. You will receive a code to enter into MTurk to collect your payment once you have finished.

Please click the link below to start the HIT.

[continue]

== [instructions for Stage 1 (cf. main text Figure 1a)] ==

### Your task

At the beginning of the HIT **you and two other real MTurkers will form a group**. We will refer to the other members of your group simply as **Other 1** and **Other 2**.

In your group, you will make decisions in two Stages which can affect your earnings.

## Stage I

In Stage I you and the two other participants each will choose between two options: Options **X** and **Y**. Your choice can affect your own earnings and the earnings of the other two participants. The earnings (in Points) of Options X and Y for each of the participants are:

<p><b>Option X</b></p> <p>You: +5 Other 1: +0 Other 2: +0</p>	<p><b>Option Y</b></p> <p>You: +2 Other 1: +2 Other 2: +2</p>
---	---

The following table illustrates how the possible outcomes of Stage I depend on yourself and the two other participants:

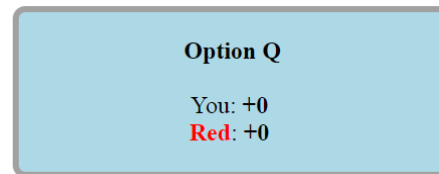
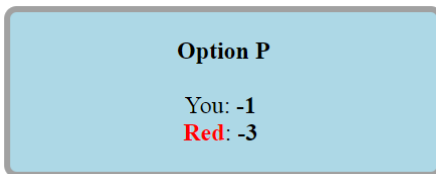
If you choose...	and the two other participants choose...	then your earnings are:
<b>X</b>	<b>X and X</b>	5
<b>X</b>	<b>X and Y</b>	7
<b>X</b>	<b>Y and Y</b>	9
<b>Y</b>	<b>X and X</b>	2
<b>Y</b>	<b>X and Y</b>	4
<b>Y</b>	<b>Y and Y</b>	6

Important: All group members make their choice between Option X and Option Y **at the same time**. Once both you and the other two participants have made a decision, you proceed to Stage II.

## Stage II

Before the beginning of Stage II, every group member is randomly assigned a color label. Two group members will be labeled **Blue** and one will be labeled **Red**. If you are assigned **Red**, you do not have to make a decision in Stage II.

If you are assigned **Blue**, you will be informed about **Red**'s decision in Stage I. Then you will choose between Options P and Q. Your choice can only affect **Red** and yourself. The earnings (in Points) of Options P and Q for you and **Red** are:



**Remember:** in this HIT you will be interacting with **real** other MTurkers who are completing this HIT **at the same time**. Please observe the **time limit** shown on your screen to avoid long waiting times. If you fail to respond in time, you will be **excluded from the task** and we will not be able to pay you.

Please click below if you understood your task. The link will open in a new window, so that you can always refer back to these instructions.

[I have read the instructions and I understood my task. Continue]

== [Compulsory comprehension questions. Participants could only proceed once they have all questions correctly] ==

## Control questions

Please answer the following questions to check your understanding of the decision situation.

**Question 1:** Suppose that all three group members (including you) choose **Option Y**.

- a. How many Points will **you** earn in Stage I?
- b. How many Points will **each of the other two group members** earn in Stage I?

**Question 2:** Suppose that all three group members (including you) choose **Option X**.

- a. How many Points will **you** earn in Stage I?
- b. How many Points will **each of the other two group members** earn in Stage I?

**Question 3:** Suppose that the other two group members chose **Option Y**.

- a. How many Points will **you** earn in Stage I if you would choose **Option Y**?
- b. How many Points will **you** earn in Stage I if you would choose **Option X**?

[submit]

== [A 'lobby' page, where participants waited to be matched with others. In the below screen, the 'X' was updated as other participants entered the lobby. Once 3 participants were in the lobby, they were automatically matched and directed to the next page. The countdown timer is initially set to 2 minutes. If this timer reaches 0, participants are given the option to leave the HIT and collect their participation fee, or to return to the lobby and wait for an additional 2 minutes] ==

Please wait until the other members of your group are ready.  
We are currently waiting for  $X$  participants.

If you are still waiting when the time below is up,  
you can leave this HIT and collect your participation fee.

[[countdown timer]]

== [Public Goods Game decision. Countdown timer set to 30 seconds.] ==

## Stage I

You have been grouped with two other participants, Other 1 and Other 2.  
Please select your choice and submit.

[[countdown timer]]

<b>Option X</b> You: +5 Other 1: +0 Other 2: +0	or	<b>Option Y</b> You: +2 Other 1: +2 Other 2: +2
--	----	--

[submit]

== [Instructions Stage 2. Instructions for Punishers (Target in italics); From this page, Targets are directed to a waiting screen and could only proceed once the two Punishers had made their decisions.] ==

## Beginning of Stage II

All members of your group have made their decision for Stage I. Stage II will start now.

The computer program has randomly assigned color labels to each of the members of your group.  
Two group members received a **blue** label, and one received a **red** label.

**You have been assigned a blue label. *[You have been assigned a red label]***

This means that you **do *[not]*** have to make a decision in Stage II. After Stage II all group members will be informed about all decisions and their final earnings in both Stages.

Please click below to continue.



== [The following pages were specific to Punishers (**Blue** players); Targets (**Red** player) were directed to the questionnaire. As soon as the Blue players in their group had made their punishment decisions, they proceeded to the results screen (see below) ] ==

## Stage II (punishers only)

You and one other group member were assigned the **blue** label. We will refer to this other group member simply as **Blue**. Similarly, we will refer to the group member with the red label as **Red**.

In this Stage, both you and **Blue** will make **two types of decisions**. You will make these decisions in two Steps: **Step 1** and **Step 2**.

=== [in the replication experiment, the last two sentences were changed to accommodate the counterbalanced design. In particular, we avoided introducing ‘Step 1’ and ‘Step 2’ and then say for half the participants that they had to make their Step 2 decision first. So, this sentence read: “(...) **both you and Blue will make two types of decisions: as first mover and as second mover.** ”] ===

To begin with, you will be informed about the decision of **Red** in Stage I.

In **Step 1** [replication experiment: *as first mover*] you will choose between **Option P** and **Option Q**. The earnings (in Points) of Options P and Q for you and **Red** are:

<b>Option P</b> You: -1 <b>Red</b> : -3	<b>Option Q</b> You: +0 <b>Red</b> : +0
---	---

**Blue** will make this decision at the same time.

In **Step 2** you will again choose between **Option P** and **Option Q**. However, now you can make your decision dependent on what **Blue** decided in Step 1. This means that we will ask you:

- What would you choose if in Step 1 **Blue** chose **Option P**? [replication experiment: “**What would you choose if Blue chose Option P as first mover?**”]
- What would you choose if in Step 1 **Blue** chose **Option Q**? [replication experiment: “**What would you choose if Blue chose Option Q as first mover?**”]

[replication experiment only, for original (reversed) decision order: “**You will start with making your first (second) mover decision, followed by your second (first) mover decision.**”]

Once both you and **Blue** have completed Step 1 and Step 2, the computer program randomly selects either you or **Blue** as the **first mover**. The remaining participant will be the **second mover**.

[replication experiment: “Once you and **Blue** have completed your *first mover* and *second mover* decisions, the computer program randomly determines which decisions will be applied. This means that the computer selects either you or **Blue** as the **first mover** The remaining participant will be the **second mover**.”]

The following table illustrates how the outcome of Stage II depends on the choices of the first and the second mover:

<b>If the first mover chooses...</b>	<b>and the corresponding choice of the second mover is...</b>	<b>then the earnings in Stage II are:</b>	<b>first mover:</b>	<b>second mover:</b>	<b>Red</b>
<b>P</b>	<b>P</b>		-1	-1	-6
<b>P</b>	<b>Q</b>		-1	0	-3
<b>Q</b>	<b>P</b>		0	-1	-3
<b>Q</b>	<b>Q</b>		0	0	0

After Stage II all group members will be informed about all decisions and their final earnings for both Stages.

Please click below if you understood your task.

=== [Compulsory comprehension questions. Participants could only proceed once they have all questions correctly; *only shown to punishers in replication experiment*] ===

## Control questions

Please answer the following questions to check your understanding of the decision situation

**Question 1:** Suppose that the computer program selects you as the *first mover*.

- How many Points will you earn in Stage II if you selected **P** for that case?
- How many Points will you earn in Stage II if you selected **Q** for that case?

**Question 2:** Suppose that the computer program selects you as the *second mover*. Suppose that **Blue** chose **P** as first mover.

- How many Points will **you** earn in Stage II if you selected **P** for that case?
- How many Points will **Blue** earn in Stage II if you selected **P** for that case?
- How many Points will **Red** earn in Stage II if you selected **P** for that case?

[Continue]

[Back to instructions]

== [The *unconditional* punishment decision; cf. main text Figure 1c; countdown timer set to 60 seconds] ==

**Step 1 (punishers only)** [Replication experiment: “*Your first mover decision*”]

[[countdown timer]]

In Stage I, **Red** chose **Option X**.

The earnings (in Points) from this choice are:

**You: +0, Blue: +0, Red: +5.**

Please select your choice and submit.

<p><b>Option P</b></p> <p>You: -1 Red: -3</p>
---

<p><b>Option Q</b></p> <p>You: +0 Red: +0</p>
---

[submit]

== [The *conditional* punishment decisions; both on the same screen; cf. main text Figure 1d and e; countdown timer set to 60 seconds] ==

## Step 2 (punishers only) [Replication experiment: “Your *second mover* decision”]

[[countdown timer]]

In Stage I, **Red** chose **Option X**.

The earnings (in Points) from this choice are:

**You: +0, Blue: +0, Red: +5.**

What would you choose if in Step 1 **Blue** chose **Option P**?

<b>Option P</b> You: -1 <b>Red: -3</b>
--

<b>Option Q</b> You: +0 <b>Red: +0</b>
--

What would you choose if in Step 1 **Blue** chose **Option Q**?

<b>Option P</b> You: -1 <b>Red: -3</b>
--

<b>Option Q</b> You: +0 <b>Red: +0</b>
--

[submit]

== [Decision phase of the experiment is over. Questionnaire items follow; anger was elicited in both the original study and the replication study] ===

## Questionnaire (punishers only)

In Stage I, **Red** chose Option Y.

Your earnings from **Red**'s choice: +0.

How **angry** did you feel when you found out **Red** decision in Stage I?

Not angry at all 0 0 0 0 0 0 Very angry

=== [Questionnaire eliciting possible motivations underlying the observed punishment preferences; cf. Supplementary Results B; [only shown in replication study](#)] ===

## Questionnaire

Now please think back of **Stage II** of the game you just played.

In that Stage you chose between Option **P** and Option **Q**.  
The earnings (in Points) of Options **P** and **Q** for you and **Red** were:

Option P
You: -1
Red: -3

Option Q
You: +0
Red: +0

You made this decision in two situations:

- (1) in case **Blue** chose Option **P**, you chose [XXX]
- (2) in case **Blue** chose Option **Q**, you chose [XXX]

Below we list seven statements about your decisions in this Stage.  
Please indicate for each *to which extent you agree* with the statement.

==[For each of the following questions, participants had to choose one of the following answers: 'Disagree strongly', 'Disagree moderately', 'Disagree a little', 'Neither agree nor disagree', 'Agree a little', 'Agree moderately', 'Agree strongly']

1. When making my decision, I was unsure what to do.
2. When making my decision, I was unsure what one *should* do.
3. I did not want to earn less than **Blue**.
4. I did not want let **Blue** down in case they chose P.
5. I did not want to reduce **Red**'s earnings by too much.
6. I wanted to reduce **Red**'s earnings *myself*.
7. I did not want to reduce my own earnings.

=== [Elicitation of Big Five Personality traits (based on Gosling et al., 2003); [only shown in replication study](#)] ===

This part of the questionnaire is about yourself.

Below we list ten brief statements.  
Please indicate for each to which extent you agree with the statement.

==[For each of the following questions, participants had to choose one of the following answers: 'Disagree strongly', 'Disagree moderately', 'Disagree a little', 'Neither agree nor disagree', 'Agree a little', 'Agree moderately', 'Agree strongly']

I see myself as...

- ... Extraverted, enthusiastic
- ... Critical, quarrelsome
- ... Dependable, self-disciplines
- ... Anxious, easily upset
- ... Open to new experiences, complex
- ... Reserved, quiet
- ... Sympathetic, warm
- ... Disorganized, careless
- ... Calm, emotionally stable
- ... Conventional, uncreative

====[ Elicitation of attitudes towards positive and negative reciprocity (based on Perugini et al., 2003); [only shown in replication study](#)]====

Below we list some brief statements about yourself.  
Please indicate for each to which extent you agree with the statement.

==[For each of the following questions, participants had to choose one of the following answers: 'Disagree strongly', 'Disagree moderately', 'Disagree a little', 'Neither agree nor disagree', 'Agree a little', 'Agree moderately', 'Agree strongly']

- If someone does me a favour, I am prepared to return it.
- If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost.
- If somebody puts me in a difficult position, I will do the same to him/her.
- I go out of my way to help somebody who has been kind to me before.
- If somebody offends me, I will offend him/her back.
- I am ready to incur personal costs to help somebody who helped me before

==== [Elicitation of risk attitudes (based on Dohmen et al., 2011); [only shown in replication study](#)]  
]====

How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risk?

==[For this questions, participants had to choose on a 10 item Likert scale where 1 means: 'avoid risks' and 10 means 'fully prepared to take risks']

==== [Final questionnaire screen shown to all participants]====

## Questionnaire

What is your gender?

What is your age?

== [Results screen listing the outcome of the first and the second stage of the game] ==

## Results

### Stage I

You chose Option X.

One **Blue** chose Option Y.

The other **Blue** chose Option Y.

Your earnings from Stage I: **9 Points**.

### Stage II

The decisions of the two **Blue** participants were:

Option P and Option P.

Your earnings from Stage II: **-6 Points**.

You started with **4 Points**.

So, in total you have earned **7 Points**.

== [Final earnings screen; the 'unique code' was specific to each participant and allowed us to pay out bonus earnings based on decisions in the game] ==

## Your earnings

In this experiment you earned XXX Points.

These Points are worth \$XXX. This is your bonus for this HIT.

The guaranteed participation fee for completing this HIT is \$0.50.

So, in total you will receive \$XXX.

Note that your participation fee and your bonus will be paid separately.

Thank you very much for your participation!

To receive your payment please copy the following unique code and enter it into MTurk:

**[10-digit code specific to each participant to match earnings between our records and MTurk]**

After entering your code you can close this window.

## Conditional cooperation experiment

Below we show the instructions for the follow-up experiment in which we elicited participants' preferences for conditional cooperation.

### Instructions

Thank you for accepting this HIT. In this HIT you will be interacting with another real MTurker who also accepted this HIT. Including the time to read these instructions, the HIT will take about 5 minutes to complete.

During this HIT you can earn Points. The number of Points you earn depends on your decisions and the decisions of the other MTurker. At the end of the HIT your Points will be converted into real money (**5 Points = \$1.00**). In addition, you will receive \$0.50 on top of however much you earn during the HIT. You will receive a code to enter into MTurk to collect your payment once you have finished.

Please click the link below to start the HIT.

### Your Task

In this HIT you and the other real MTurker will form a group. You and the other participant will make **two types of decisions**. You will make these decisions in two Steps: **Step 1** and **Step 2**.

In **Step 1** you will choose between **Option X** and **Option Y**. The earnings (in Points) of Options X and Y for you and the other participant are:

Option X	Option Y
You: +3	You: +2
They: +0	They: +2

They will make the same decision.

In **Step 2** you will again choose between **Option X** and **Option Y**. However, now you can make your decision dependent on what they decided in Step 1. This means that we will ask you:

- What would you choose if in Step 1 they chose **Option X**?
- What would you choose if in Step 1 they chose **Option Y**?

### Determining Outcomes

Once both you and the other participant have completed Step 1 and Step 2, the computer program randomly



selects either you or the other participant as the **first mover**. The remaining participant will be the **second mover**.

The following table illustrates how the outcome depends on the first mover's and the second mover's choice:

If the first mover in Step 1 chooses...	and the corresponding choice of the second mover in Step 2 is...	then the first mover's earnings are	and the second mover's earnings are
X	X	3	3
X	Y	5	2
Y	X	2	5
Y	Y	4	4

Please click below if you understood your task.

### Control questions

Please answer the following questions to check your understanding of the decision situation.

Recall the two options:

Option X	Option Y
You: <b>+3</b>	You: <b>+2</b>
They: <b>+0</b>	They: <b>+2</b>

**Question 1:** Suppose that you and the other participant both choose **Option X**.

- a. How many Points will **you** earn?
- b. How many Points will **the other participant** earn?

**Question 2:** Suppose that you and the other participant both choose **Option Y**. a. How many Points will **you** earn?

- a. How many Points will **you** earn?
- b. How many Points will **the other participant** earn?

**Question 3:** Suppose that the other participant chooses **Option Y**.

- a. How many Points will **you** earn if you would choose **Option Y**
- b. How many Points will **you** earn if you would choose **Option X**?

**Step 1**

Please make a choice between the following two options:

<b>Option X</b>  You: <b>+3</b> They: <b>+0</b>	<b>Option Y</b>  You: <b>+2</b> They: <b>+2</b>
--	--

I choose

- Option X
- Option Y

**Step 2**

Recall the two choice options:

Option X	Option Y
You: +3	You: +2
They: +0	They: +2

What would you choose if in Step 1 **they** chose **Option X**?

- Option X
- Option Y

What would you choose if in Step 1 **they** chose **Option Y**?

- Option X
- Option Y

### **Your bonus earnings**

Your bonus earnings for this HIT will be determined as follows. On DD-MM-YYYY, the decisions of all MTurkers who have participated in this HIT will be collected, and you will be randomly matched with another participant.

As explained before, a computer program will randomly select either you or the other participant as the first mover. The remaining participant will be the second mover. Your earnings will be calculated by implementing the first mover's decision in Step I, and the corresponding decision of the second mover in Step II.

Please note that your guaranteed participation fee of \$0.50 and your bonus will be paid separately.

Please click below to continue and receive your completion code to input on MTurk.

## Supplementary References

1. Cialdini, R. B. & Trost, M. R. Social influence: Social norms, conformity and compliance. in *The handbook of social psychology, Vols. 1 and 2 (4th ed.)* (eds. Gilbert, D. T., Fiske, S. T. & Lindzey, G.) 151–192 (McGraw-Hill, 1998).
2. Cialdini, R. B. & Cialdini, R. B. *Influence: The psychology of persuasion*. (Collins New York, 2007).
3. Boehm, C. *Hierarchy in the forest: Egalitarianism and the evolution of human altruism*. (Harvard University Press.[aDSW], 1999).
4. Wiessner, P. Norm enforcement among the Ju/'hoansi Bushmen. *Hum. Nat.* **16**, 115–145 (2005).
5. Gosling, S. D., Rentfrow, P. J. & Swann, W. B. A very brief measure of the Big-Five personality domains. *J. Res. Personal.* **37**, 504–528 (2003).
6. Perugini, M., Gallucci, M., Presaghi, F. & Ercolani, A. P. The personal norm of reciprocity. *Eur. J. Personal.* **17**, 251–283 (2003).
7. Dohmen, T. *et al.* Individual risk attitudes: Measurement, determinants, and behavioral consequences. *J. Eur. Econ. Assoc.* **9**, 522–550 (2011).
8. Boyd, R., Gintis, H. & Bowles, S. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620 (2010).
9. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).
10. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *J. Public Econ.* **92**, 91–112 (2008).
11. Gächter, S., Herrmann, B. & Thöni, C. Culture and cooperation. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2651–2661 (2010).
12. Rand, D. G. & Nowak, M. A. The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2**, 434 (2011).
13. García, J. & Traulsen, A. Leaving the loners alone: Evolution of cooperation in the presence of antisocial punishment. *J. Theor. Biol.* **307**, 168–173 (2012).
14. Hauser, O. P., Nowak, M. A. & Rand, D. G. Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible. *J. Theor. Biol.* **360**, 163–171 (2014).
15. McCabe, C. M. & Rand, D. G. Coordinated punishment does not proliferate when defectors can also punish cooperators. *J. Commun. Res.* **6**, (2014).
16. Huang, F., Chen, X. & Wang, L. Conditional punishment is a double-edged sword in promoting cooperation. *Sci. Rep.* **8**, 528 (2018).
17. Arechar, A. A., Gächter, S. & Molleman, L. Conducting interactive experiments online. *Exp. Econ.* **21**, 99–131 (2018).