People prefer coordinated punishment in cooperative interactions

Molleman, L.; Kölle, F.; Starmer, C.; Gächter, S.

DOI
10.1038/s41562-019-0707-2

Publication date
2019

Document Version
Final published version

Published in
Nature Human Behaviour

nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-019-0707-2

People prefer coordinated punishment in cooperative interactions

Lucas Molleman[1,2,3*], Felix Kölle[2,4*], Chris Starmer[2] and Simon Gächter[2,5,6]

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

The ecological success of humans has often been attributed to our propensity for cooperation[2,4]. Many people are willing to go out of their way to help others, allowing human groups to deal with environmental challenges and to do things no individual can achieve on their own. But, as natural as it might seem, cooperation looks puzzling from the viewpoint of rational self-interest: why would one cooperate if others can reap the benefits of cooperation without paying the costs? This supposed 'free-rider problem' affects countless real-world situations, ranging from day-to-day work in teams and paying taxes to curbing overfishing and reducing carbon dioxide emissions. Moreover, a vast range of theoretical models and empirical studies from across the biological and social sciences have documented that, when studied in isolation, individually costly cooperation tends to break down through processes of natural selection or (social) learning, which often favour free riding[6–14].

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

Peer punishment has been extensively studied in a wide range of different experimental settings, both in decision-making laboratories and in the field, as well as across different interaction settings and across cultures[20,21,14,22–29]. By and large, these studies indicate that many people are inclined to punish free-riding interaction partners (often motivated by negative emotions such as anger), supporting the idea that such peer punishment can lead to a welfare-enhancing stabilization of cooperation at high levels.

While in experimental settings punishment has great potential to support cooperation, evolutionary models suggest that punishment can only emerge under very limited circumstances[30–32]. The reason is that punishment often entails costs for those who mete it out. This can create a 'second-order' free-rider problem: while only some individuals incur the costs of punishing, all members of a group may benefit from enhanced cooperation after non-cooperators are punished. Hence, from an individual perspective it can pay to refrain from punishment[24,33–35]. Over time, this may result in a decline of punishment in a population of self-interested agents, compromising its potential to support cooperation.

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

Despite these promising theoretical results, there is only scarce empirical support for the claim that individuals have a preference for

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. [2]Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham, Nottingham, UK. [3]Amsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam, the Netherlands. [4]Faculty of Management, Economics and Social Sciences, University of Cologne, Cologne, Germany. [5]Center for Economic Studies, Munich, Germany. [6]IZA Institute of Labour Economics, Bonn, Germany. *e-mail: l.s.molleman@uva.nl; felix.koelle@uni-koeln.de

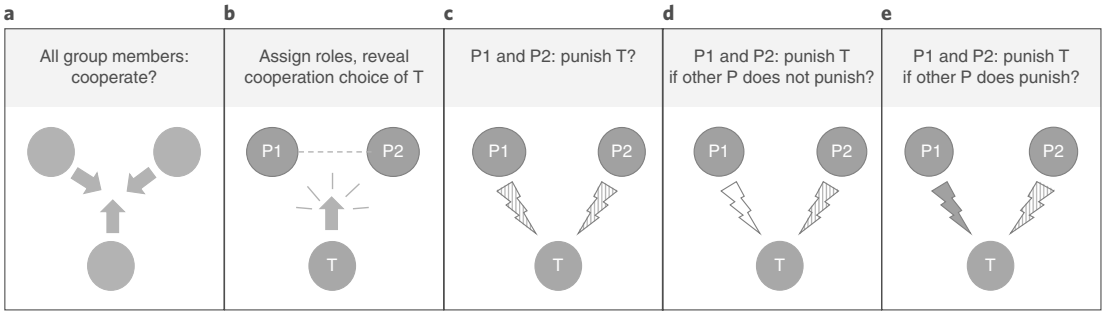

Fig. 1 | Experimental sequence. a, Participants are assigned to a group of three (grey circles) and make a binary decision in a PGG (grey arrows). **b**, Roles are randomly allocated among group members: two Punishers (P1 and P2; blue circles) and one Target (T; orange circle). T's PGG decision is revealed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

coordinated punishment. While field evidence[15,17,52–54] is consistent with the idea of coordinated punishment, it is not conclusive about whether people prefer coordinated punishment over independent punishment. Experimental evidence is needed to establish the existence of preferences for coordinated punishment. Hitherto, experimental studies of conditional punishment are rare, and mainly focus on how punishment is impacted by the distribution of cooperative behaviour (not the punishment behaviour) in the population, or analyse how aggregate levels of cooperation are affected when free riders can only be punished if multiple peers agree to do so[19,55–62]. Here, we investigate whether people prefer to coordinate their punishment in the context of an experiment in which individuals can explicitly condition their punishment on the punishment decisions of others. That is, we ask whether, analogously to conditional preferences observed in pro-social contexts such as cooperation (positive reciprocity), such conditionality is also characteristic of preferences to punish others (negative reciprocity), and if so, whether preferences for positive and negative reciprocity are correlated.

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

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

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

At the start of the second stage, two of the three group members are randomly allocated the roles of Punishers, and the remaining one is allocated the role of Target who can be punished but cannot punish the Punishers. The PGG decision of the Target is revealed, and then Punishers make binary choices whether or not to punish (that is, assign deduction points to) the Target. Assigning deduction points incurs a cost of 1 MU to the Punisher and 3 MU to the Target. The impact of punishment is additive: when a Target is punished by one group member they lose 3 MU and when they are punished by both of their group mates they lose 6 MU. Punishers cannot assign deduction points to one another, and neither can they observe each other's PGG decision when making their deduction point decision (mitigating possible effects of inequality between Punishers stemming from the contribution stage).

We examine whether people prefer to coordinate their punishment by having Punishers make two types of punishment decisions. First, they make an 'unconditional' punishment decision, deciding whether they want to punish the Target or not, irrespective of the decision of the other Punisher. Subsequently, they make two 'conditional' punishment decisions, in which they can condition their punishment on that of the other Punisher. To do this, we use the strategy method[63]: Punishers indicate whether or not they want to punish the Target in case the other Punisher chooses (1) to punish or (2) to not punish the Target. Our analysis mainly focuses on these two decisions made by the $n = 1{,}336$ Punishers in our experiment. Figure 1 summarizes the decision situations.

Once both Punishers have entered both types of decisions, one Punisher is randomly chosen and their unconditional punishment decision is implemented to initiate the punishment procedure. Subsequently, the corresponding conditional punishment decision of the other Punisher is implemented (Methods). This setup yields an incentive-compatible decision situation in a minimal social context that allows us to cleanly measure how people condition their punishment on the punishment decisions of others. Importantly, the strategy method yields a full punishment profile for each individual that is independent of their beliefs about other people's punishment decision. Because of the one-shot nature of the game, there

Table 1 | Behavioural determinants of conditional punishment

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

Coefficients from logistic generalized linear mixed models fitted to Punishers' decisions whether or not to punish the Target (1 if yes, 0 if no). '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 they 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 have 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[93]. The 95% confidence intervals are in brackets and *P* values are in parentheses.

are also no strategic incentives for punishment. Further, note that because punishment is costly and the game is played only once, if players are only interested in maximizing their own material payoff, no one is predicted to punish. As a result, full defection with no-punishment is the only Nash equilibrium of the game.

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

The cooperation rate in the PGG stage of the game was 48%. Among the 1,336 Punishers, the overall unconditional punishment rate was 11.4%. Unconditional punishment varied considerably with the cooperation decisions of Punishers and their Target. In particular, unconditional punishment rates were highest when the Punisher cooperated and the Target defected (24.0%), and lowest when both the Punisher and Target cooperated (5.5%). When the Target cooperated and the Punisher defected, punishment was also low (5.7%), while if both players defected punishment was intermediate occurring in 9.9% of the cases.

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

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

In Model 2 we confirm that the positive effect of the other punisher's decision is robust to alternative model specifications, and further investigate how preferences for coordinated punishment vary with the Target's decision to cooperate or defect. We find that a Target's decision to cooperate leads to lower overall levels of punishment, as indicated by the significant negative 'Target cooperated' dummy. Preferences for coordinated punishment, however, seem independent of the Target's PGG decision as the interaction term of 'Target cooperated' × 'Other punishes' is not statistically significant. This means that defection leads to higher levels of punishment, but does not make it more likely that people coordinate (or anticoordinate) their punishment. However, as unconditional punishment towards cooperators ('antisocial punishment')[25] occurred only in 5.7% of cases, coordinated antisocial punishment is very rare in our data; only 0.63% of the cooperators got punished in a coordinated way.

We now turn to the individual-level preferences for conditional punishment elicited with the strategy method (Fig. 1e,f). We distinguish four 'punishment types': (1) 'coordinated punisher', punishing only if the other does so as well; (2) 'anticoordinated punisher', punishing only if the other does not punish; (3) 'independent punisher', punishing regardless of the other's punishment decision; and (4) 'non-punisher', not punishing at all. Here we focus on the relative frequencies of these types.

In line with established findings for one-shot games without strategic incentives to punish[23,27,68,69], the majority of participants in the role of Punisher chose to never punish (78.9%). Among the $n = 282$ Punishers who punish at least once, we find a strong and striking pattern (Fig. 2a). In this group, coordinated punishers are most frequent (47.5%). The independent punishers and anticoordinated punishers are much less frequent (25.9% and 26.6%, respectively).

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

These behavioural patterns demonstrate that many people prefer coordinated punishment. As we developed our own experimental paradigm to examine rarely explored preferences for conditional

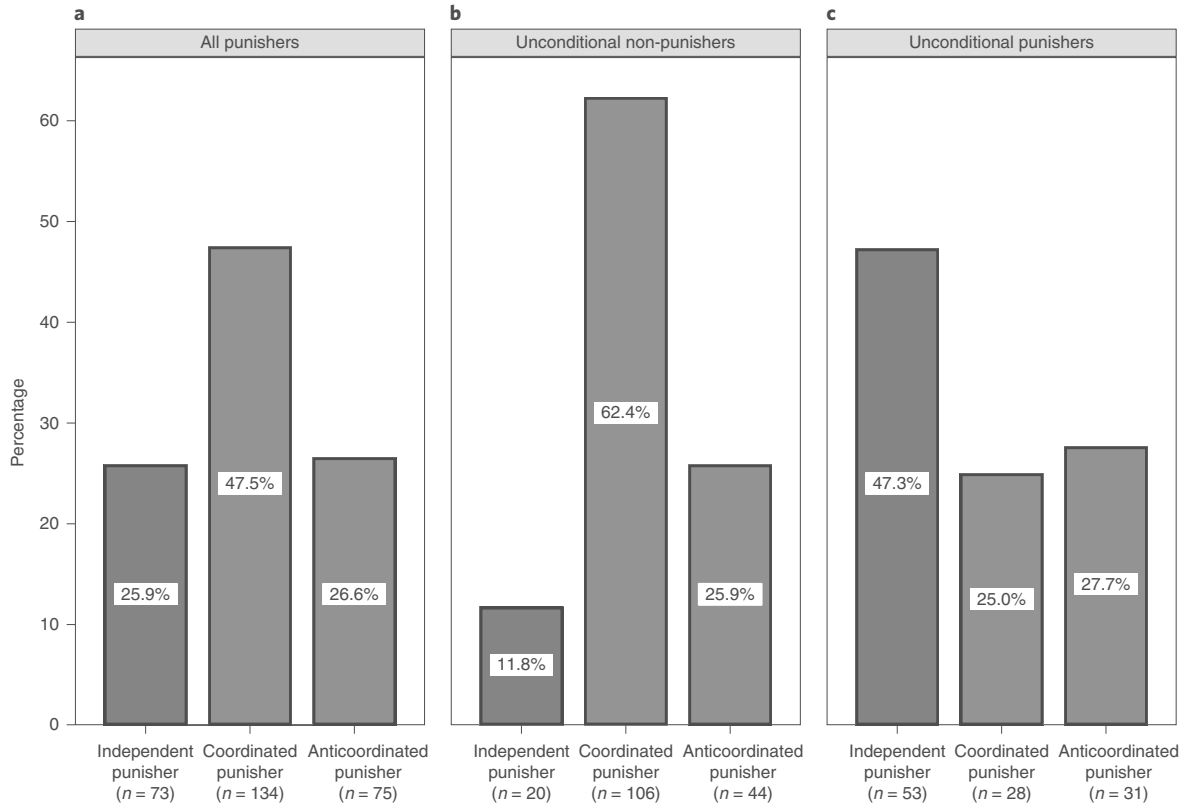

Fig. 2 | Distribution of punishment types by decision in unconditional punishment stage. Bars show data from the strategy method, restricted to those individuals who punished at least once ($n = 282$ out of the total of $n = 1{,}336$). **a**, All punishers. **b**, Only those who did not punish in the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

punishment, one might ask how robust our results are and, perhaps more fundamentally, why people would prefer coordinated punishment.

To test the robustness and replicability of our results, and to further probe motivations underlying conditional punishment preferences, we ran a new study with $n = 2{,}316$ additional participants. The design and the procedures of this new study were the same as in our original experiment, with the following exceptions. First, to rule out that the observed patterns in punishment preferences are due to confusion about the strategy method and the payoff consequences for punishment, we added a set of control questions right before participants entered the punishment stages. Second, as a further robustness check we counterbalanced the order in which participants made their punishment decisions, such that in half of the groups Punishers completed the 'conditional' stage (Fig. 1d,e) before the 'unconditional' stage (Fig. 1c). Third, to further explore the motivational factors underlying preferences for conditional punishment, we extended the postexperimental questionnaire with items probing not only experienced anger when making punishment decisions, but also other possible motivations such as a desire for revenge towards the Target, reciprocity towards the other Punisher, as well as inequality concerns.

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

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

To understand the potential drivers and underlying motivations of individuals' punishment preferences, we analyse Punishers' reported levels of anger (using data from both studies) as well as their responses to the postexperimental questionnaire in our

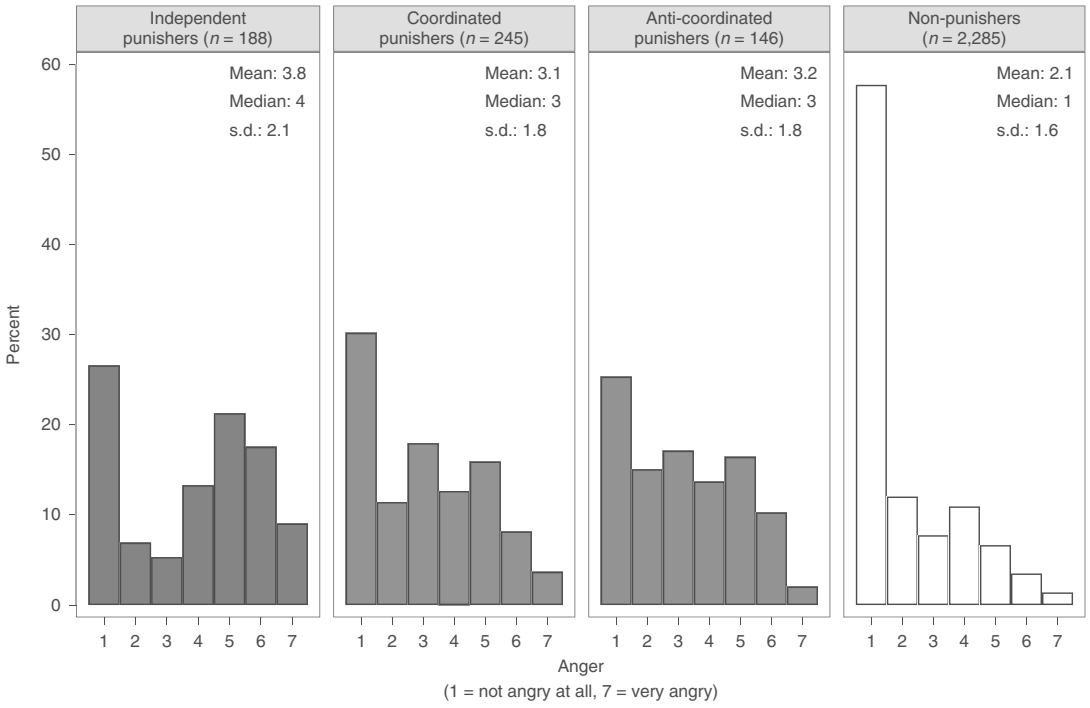

Fig. 3 | Anger levels per punishment type. Panels show distributions of Punishers' self-reported ratings of anger when they learned about the PGG decision of their Target. The distribution of anger levels differs significantly across punishment types (Kruskal–Wallis test: $\chi^2 = 198.29$, d.f. = 3,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

replication study. For eliciting anger levels, we asked punishers to rate their level of anger when making their punishment decisions on a 7-point scale (1, not angry at all; 7, very angry). On average, anger scores are highest when the Punisher cooperated and the Target defected (3.7) and lowest if both cooperated (1.5). We find anger levels to be significantly higher for unconditional punishers than for unconditional non-punishers (3.7 versus 2.2; Mann–Whitney U (MWU)-test, $z = 15.34$, d.f. = 1, $P < 0.001$, $r = 0.29$).

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

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

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

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

We further hypothesized that coordinated punishers might be only willing to punish if others do so too, because they do not want their payoffs to fall behind those of the other punisher. Consistent with this hypothesis, we found that, among those participants who punished at least once, coordinated punishers (CP) tended to agree the most with the statement 'I did not want to earn less than [the other punisher]' ($\mu_{CP} = 4.9$, $\mu_{\text{other punishers}} = 4.6$; MWU-test: $z = 2.09$, d.f. = 1, $P = 0.036$, $r = 0.12$). Another rationale for coordinated punishment might be that coordinated punishers see punishment by the other group member as a nice act (enforcing a norm of cooperation) and feel the need to reciprocate. This idea is supported by the observation that coordinated punishers agreed most with the statement 'I did not want to let [the other punisher] down in case they chose to punish' ($\mu_{CP} = 4.9$, $\mu_{\text{other punishers}} = 4.5$; MWU-test: $z = 2.26$, d.f. = 1, $P = 0.024$, $r = 0.13$). Finally, coordinated punishers may be unsure what to do when making their punishment decisions, or be unsure whether punishment is socially appropriate or legitimate[15,53], and

take others' punishment behaviour as a 'principle of social proof'[72]. We did not find support for these possible motives: coordinated punishers did not agree more with the statements 'When making my [conditional punishment] decisions, I was unsure what to do' ($\mu_{CP}=4.0$, $\mu_{\text{other punishers}}=4.0$; MWU-test: $z=0.11$, d.f.$=1$, $P=0.914$, $r=0.01$) or 'When making my [conditional punishment] decisions, I was unsure what was the appropriate thing to do' ($\mu_{CP}=4.2$, $\mu_{\text{other punishers}}=4.1$; MWU-test: $z=0.01$, d.f.$=1$, $P=0.998$, $r=0.01$).

Across our two sets of experiments, we find unambiguous evidence that many people like to condition their punishment decisions on those of other people. This conditionality is reminiscent of 'conditional cooperation', that is, many people's conditional willingness to contribute to a public good in the first place provided others do the same. This raises the interesting question whether conditional punishers are also conditional cooperators. To answer this question, in the following, we examine how preferences for conditional punishment (that is, negative reciprocity) relate to well-studied preferences for conditional cooperation (that is, positive reciprocity)[73]. That is, are these preferences linked and do they reflect a general sensitivity to social influence by peers, or are they unrelated, indicating that inclinations to reciprocate are context-specific? Existing evidence, which, with exceptions[62,74], only looked at cooperation and punishment decisions and did not elicit conditional preferences, suggests that positive and negative reciprocity are unrelated[75]. However, given that, conceptually, cooperation and punishment share the logic of a public good (while both cooperation and punishment are individually costly, all group members may benefit), one might expect that people who prefer to cooperate conditionally on others' cooperation also prefer to condition their punishment on others' punishment. To test this hypothesis, we conducted a follow-up experiment examining whether individuals' punishment preferences are related to their preferences for conditional cooperation.

Two weeks after participating in our conditional punishment experiments (both from the original and the new study), a subset of participants who were in the role of a Punisher was re-invited to participate in an additional study. In this follow-up experiment, participants were randomly matched with a partner to play a one-shot dyadic binary Prisoner's Dilemma Game, in which both players had to choose to either 'cooperate' or 'defect'. From the perspective of an individual participant, defecting yielded a personal benefit of 3 MU and 0 MU for their partner. Cooperating yielded 2 MU for both partners (for instructions, see Supplementary Methods).

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

To ensure sufficient statistical power, we selectively re-invited participants to obtain a more balanced sample with respect to the distribution of punishment types, compared to our full sample from the primary experiment. That is, we aimed to oversample those who punished at least once in the conditional stage in the primary experiment (cf. Fig. 2) and undersample those who never punished.

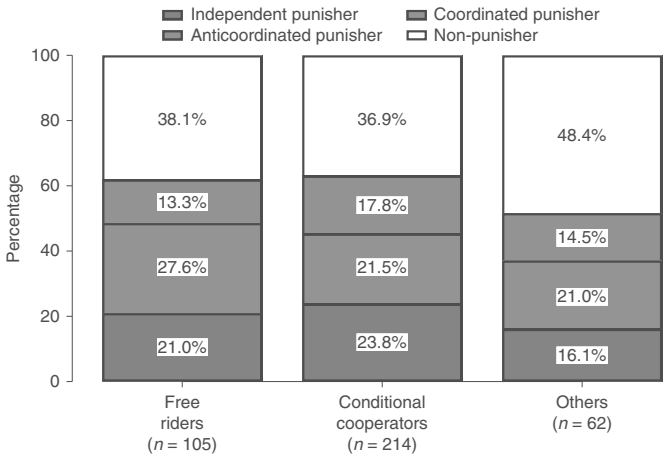

Fig. 4 | The (lack of) correlation between individuals' preferences for conditional cooperation and conditional punishment. Stacked bars show the distribution of punishment types as measured in the primary experiment, separated by conditional cooperation type as measured in the follow-up experiment.

This procedure indeed led to a more evenly distributed sample with respect to punishment types: among the $n=381$ participants in the follow-up experiment, 39% were non-punishers, 22% were unconditional punishers, 23% were coordinated punishers and 16% were anticoordinated punishers. In this sample we find a commonly observed pattern with regard to the distribution of cooperation types: more than half of the people (56%) are conditional cooperators, about 28% are free riders and the remaining 16% are classified as 'others'.

To investigate whether preferences for conditional cooperation and conditional punishment are linked at the individual level, we compare the distribution of punishment types across the different cooperation types. If these two types of preference reflect a more general behavioural tendency (for example, inclinations to reciprocate or to conform with the behaviour of others), we would expect that coordinated punishment is particularly frequent among conditional cooperators.

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

Our large-scale experiments provide firm empirical evidence that many people prefer to coordinate their punishment in cooperative interactions. Our results support theories that explain the emergence and maintenance of human cooperation based on individuals sanctioning their peers jointly rather than individually[5]. When deciding whether or not to punish a peer, many people are more willing to engage in costly punishment if others do so too. Intriguingly, preferences for coordinated punishment are particularly pronounced among those who do not punish unconditionally,

suggesting that punishment levels can rise substantially when people have the opportunity to coordinate their sanctions.

Our results indicate that conditional preferences are not limited to the domain of positive reciprocity (cooperation), but extend to the domain of negative reciprocity (punishment), too. On aggregate, in both domains conditional preferences lead individuals to align their decisions with others and conform to their actions. However, there is substantial heterogeneity in how individuals condition their punishment on the punishment behaviour of others (Fig. 2). Interestingly, our data suggest that people's conditional preferences in the domains of positive and negative reciprocity are unrelated (Fig. 4). This result supports the emerging view that behavioural strategies of cooperation and punishment are not closely associated with each other[14,74,75,77–80], and suggests that cooperation and punishment are separate phenomena, each driven by its own psychological processes. Indeed, the lack of correlation between conditional punishment and conditional cooperation indicates that, while first-order and second-order free-rider problems may look theoretically very similar, the underlying mechanisms supporting behaviour in them may be quite distinct.

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

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

We conducted our experiments online, with American participants from Amazon Mechanical Turk (MTurk). This platform is well suited for large-scale studies such as ours, and gives the opportunity to recruit a more demographically diverse sample than student samples typically recruited in laboratory studies[81]. Although collecting data online is associated with reduced levels of experimental control relative to the traditional decision-making laboratory, this does not have to compromise data quality[82–84], especially when the methodological challenges of conducting experiments online are adequately addressed[29]. Moreover, a recent study on cooperation and punishment found that MTurkers punish in a similar way as students in the laboratory[29], giving reason to be optimistic about the generalizability of our results. However, it is an empirical question whether the patterns of conditional punishment preferences from our study would be observed under more standard laboratory conditions or, for example, in samples from different cultural backgrounds.

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

Our study set out to investigate preferences for conditional punishment in a very simple and 'minimal' environment: that is, one which is just complex enough to allow clean tests for conditional punishment preferences. While we judge this is the right place to conduct initial tests for such preferences, having provided clear evidence for them, we believe that incrementally increasing the complexity of the experimental decision-making situation (that is, the number of factors at play) will help achieve a more complete empirical understanding of conditional punishment. Interesting extensions of our basic experiment would include nonlinear returns to scale of punishment[87]. For example, coordinated sanctions might be more efficient than individual, uncoordinated punishment, and less risky for those who mete them out as revenge is less likely[5]. Testing whether anticipating such 'synergy' modulates people's preferences for coordinated punishment would be of great value. Further experiments could test how coordinated punishment impacts the long-run dynamics of cooperation. When time horizons are longer than the one-shot interactions used in our study, decisions to cooperate and punish have a strategic dimension, potentially involving interactions between coordinated punishment and individuals' reputations. Such experiments could also test the theoretically predicted deleterious implications of antisocial punishment[88,89] in situations where defectors coordinate their punishment and 'gang up' against cooperators[90,91].

Methods

For the primary experiment, we recruited $n = 2{,}004$ participants from Amazon MTurk, two-thirds of whom ($n = 1{,}336$) had the role of Punisher (Fig. 1). Participants completed the experiment in about 10 min and earned a flat fee of US$0.50 plus their earnings from the game. At the end of a session, monetary units were converted into money at the rate 10 MU = US$1.00. Total average earnings were US$1.50, which corresponds to an average hourly wage of US$9.00. Before the start of the PGG, each participant had to correctly answer a set of control questions designed to test their understanding of the interaction setting. Participants were all US citizens; 55% were male and their mean age was 32.7 years. The online experiment was developed with the software LIONESS[92]. Ethical approval was provided by the Research Ethics Committee at the School of Economics, University of Nottingham. All experimental instructions are documented in the Supplementary Methods.

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

For the follow-up experiment, intended to measure preferences for conditional cooperation[76], we recruited $n = 177$ individuals who had participated in the primary experiment. The follow-up experiment was programmed in Qualtrics and took about 7–8 min. Participants were matched post-hoc to calculate their earnings, consisting of a flat fee of US$0.50 plus their earnings from the game, which were converted into money at the rate 5 MU = US$1.00. To calculate their game earnings, we matched 176 of the 177 participants in pairs. A random mechanism chose which type of decision was implemented for each partner. In particular, for one player the 'unconditional' cooperation decision was implemented (the first mover), while for the other player (the second mover) the corresponding conditional cooperation decision (depending on the first mover's decision) was implemented. The earnings for the remaining (177th)

participant were calculated by using their unconditional cooperation decision and implementing the corresponding conditional cooperation decision of a randomly chosen other participant. Total average earnings were US$1.25, which corresponds to an average hourly wage of US$10.00.

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

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

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data underlying the results reported in our manuscript can be found on Github at https://github.com/LucasMolleman/NHB_CoordinatedPunishment.

Code availability

The LIONESS code for the online experiment is available upon request from the corresponding authors. Analysis code for STATA can be found on Github at https://github.com/LucasMolleman/NHB_CoordinatedPunishment.

Received: 16 March 2018; Accepted: 23 July 2019;
Published online: 2 September 2019

References

1. Fehr, E., Fischbacher, U. & Gächter, S. Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13**, 1–25 (2002).
2. Bowles, S. & Gintis, H. *A Cooperative Species: Human Reciprocity and Its Evolution* (Princeton Univ. Press, 2011).
3. Rand, D. G. & Nowak, M. A. Human cooperation. *Trends Cogn. Sci.* **17**, 413–425 (2013).
4. Hammerstein, P. (ed.) *Genetic and Cultural Evolution of Cooperation.* (MIT Press, 2003).
5. Boyd, R., Gintis, H. & Bowles, S. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620 (2010).
6. Hamilton, W. D. The genetical evolution of social behaviour I and II. *J. Theor. Biol.* **7**, 1–52 (1964).
7. Gintis, H. *Game Theory Evolving: A Problem-centered Introduction to Modeling Strategic Interaction* (Princeton Univ. Press, 2000).
8. Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).
9. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
10. Lehmann, L. & Keller, L. The evolution of cooperation and altruism—a general framework and a classification of models. *J. Evol. Biol.* **19**, 1365–1376 (2006).
11. Egas, M. & Riedl, A. The economics of altruistic punishment and the maintenance of cooperation. *Proc. R. Soc. Lond. B* **275**, 871–878 (2008).
12. Fischbacher, U. & Gachter, S. Social preferences, beliefs, and the dynamics of free riding in public good experiments. *Am. Econ. Rev.* **100**, 541–556 (2010).
13. Burton-Chellew, M. N., El Mouden, C. & West, S. A. Social learning and the demise of costly cooperation in humans. *Proc. R. Soc. B* **284**, 20170067 (2017).
14. Gächter, S., Kölle, F. & Quercia, S. Reciprocity and the tragedies of maintaining and providing the commons. *Nat. Hum. Behav.* **1**, 650 (2017).
15. Boehm, C. *Hierarchy in the Forest: Egalitarianism and the Evolution of Human Altruism* (Harvard Univ. Press, 1999).
16. Sigmund, K. Punish or perish? Retaliation and collaboration among humans. *Trends Ecol. Evol.* **22**, 593–600 (2007).
17. Guala, F. Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35**, 1–15 (2012).
18. Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458 (2018).
19. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: self-governance is possible. *Am. Polit. Sci. Rev.* **86**, 404–417 (1992).
20. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
21. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
22. Henrich, J. et al. In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**, 73–78 (2001).
23. Henrich, J. et al. Costly punishment across human societies. *Science* **312**, 1767–1770 (2006).
24. Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. Winners don't punish. *Nature* **452**, 348 (2008).
25. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).
26. Gächter, S., Renner, E. & Sefton, M. The long-run benefits of punishment. *Science* **322**, 1510–1510 (2008).
27. Cubitt, R. P., Drouvelis, M. & Gächter, S. Framing and free riding: emotional responses and punishment in social dilemma games. *Exp. Econ.* **14**, 254–272 (2011).
28. Raihani, N. J., Thornton, A. & Bshary, R. Punishment and cooperation in nature. *Trends Ecol. Evol.* **27**, 288–295 (2012).
29. Arechar, A. A., Gächter, S. & Molleman, L. Conducting interactive experiments online. *Exp. Econ.* **21**, 99–131 (2018).
30. Panchanathan, K. & Boyd, R. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).
31. Gardner, A. & West, S. A. Cooperation and punishment, especially in humans. *Am. Nat.* **164**, 753–764 (2004).
32. Lehmann, L., Rousset, F., Roze, D. & Keller, L. Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment. *Am. Nat.* **170**, 21–36 (2007).
33. Heckathorn, D. D. Collective action and the second-order free-rider problem. *Ration. Soc.* **1**, 78–100 (1989).
34. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499 (2004).
35. Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863 (2010).
36. Barclay, P. Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* **27**, 325–344 (2006).
37. dos Santos, M., Rankin, D. J. & Wedekind, C. The evolution of punishment through reputation. *Proc. R. Soc. Lond. B* **278**, 371–377 (2011).
38. dos Santos, M., Rankin, D. J. & Wedekind, C. Human cooperation based on punishment reputation. *Evolution* **67**, 2446–2450 (2013).
39. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103 (2015).
40. Henrich, J. & Boyd, R. Why people punish defectors. *J. Theor. Biol.* **208**, 79–89 (2001).
41. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
42. Kiyonari, T. & Barclay, P. Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *J. Pers. Soc. Psychol.* **95**, 826 (2008).
43. Fu, T., Ji, Y., Kamei, K. & Putterman, L. Punishment can support cooperation even when punishable. *Econ. Lett.* **154**, 84–87 (2017).
44. Szolnoki, A., Szabó, G. & Perc, M. Phase diagrams for the spatial public goods game with pool punishment. *Phys. Rev. E* **83**, 036101 (2011).
45. Traulsen, A., Röhl, T. & Milinski, M. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B* **279**, 3716–3721 (2012).
46. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110 (1986).
47. Ostrom, E. *Governing the Commons* (Cambridge Univ. Press, 2015).
48. Hilbe, C., Traulsen, A., Röhl, T. & Milinski, M. Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *Proc. Natl Acad. Sci., USA* **111**, 752–756 (2014).
49. Szolnoki, A. & Perc, M. Effectiveness of conditional punishment for the evolution of public cooperation. *J. Theor. Biol.* **325**, 34–41 (2013).
50. FeldmanHall, O., Otto, A. R. & Phelps, E. A. Learning moral values: Another's desire to punish enhances one's own punitive behavior. *J. Exp. Psychol. Gen.* **147**, 1211 (2018).

51. Son, J.-Y., Bhandari, A. & FeldmanHall, O. Crowdsourcing punishment: individuals reference group preferences to inform their own punitive decisions. *Scientific Reports* **9**, 11625 (2019).
52. Mahdi, N. Q. Pukhtunwali: ostracism and honor among the Pathan hill tribes. *Ethol. Sociobiol.* **7**, 295–304 (1986).
53. Wiessner, P. Norm enforcement among the Ju/'hoansi Bushmen. *Hum. Nat.* **16**, 115–145 (2005).
54. Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl Acad. Sci. USA* **108**, 11375–11380 (2011).
55. Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111 (2006).
56. Ertan, A., Page, T. & Putterman, L. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *Eur. Econ. Rev.* **53**, 495–511 (2009).
57. Casari, M. & Luini, L. Cooperation under alternative punishment institutions: An experiment. *J. Econ. Behav. Organ.* **71**, 273–282 (2009).
58. Casari, M. & Luini, L. Peer punishment in teams: expressive or instrumental choice? *Exp. Econ.* **15**, 241–259 (2012).
59. Kamei, K. Conditional punishment. *Econ. Lett.* **124**, 199–202 (2014).
60. Cheung, S. L. New insights into conditional cooperation and punishment from a strategy method experiment. *Exp. Econ.* **17**, 129–153 (2014).
61. Peysakhovich, A. & Rand, D. G. Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Manag. Sci.* **62**, 631–647 (2015).
62. Albrecht, F., Kube, S. & Traxler, C. Cooperation and norm enforcement-the individual-level perspective. *J. Public Econ.* **165**, 1–16 (2017).
63. Selten, R. Die Strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperimentes. in *Beiträge zur experimentellen Wirtschaftsforschung: Seminar für Mathemat. Wirtschaftsforschung u. Ökonometrie* (ed. Sauermann, H.) 136–168 (J.C.B. Mohr, 1965).
64. Bosman, R. & Van Winden, F. Emotional hazard in a power-to-take experiment. *Econ. J.* **112**, 147–169 (2002).
65. Falk, A., Fehr, E. & Fischbacher, U. Driving forces behind informal sanctions. *Econometrica* **73**, 2017–2030 (2005).
66. Hopfensitz, A. & Reuben, E. The importance of emotions for the effectiveness of social punishment. *Econ. J.* **119**, 1534–1559 (2009).
67. Nelissen, R. M. A. & Zeelenberg, M. Moral emotions as determinants of third-party punishment: anger, guilt, and the functions of altruistic sanctions. *Judgm. Decis. Mak.* **4**, 543–553 (2009).
68. Gächter, S. & Herrmann, B. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philos. Trans. R. Soc. B* **364**, 791–806 (2009).
69. Gächter, S. & Herrmann, B. The limits of self-governance when cooperators get punished: experimental evidence from urban and rural Russia. *Eur. Econ. Rev.* **55**, 193–210 (2011).
70. Elster, J. Norms of revenge. *Ethics* **100**, 862–885 (1990).
71. Nikiforakis, N. Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Public Econ.* **92**, 91–112 (2008).
72. Cialdini, R. B. & Trost, M. R. *The Handbook of Social Psychology* 4th edn, Vols. 1 and 2 (eds. Gilbert, D. T. et al.) 151–192 (McGraw-Hill, 1998).
73. Thöni, C. & Volk, S. Conditional cooperation: review and refinement. *Econ. Lett.* **171**, 37–40 (2018).
74. Weber, T. O., Weisel, O. & Gächter, S. Dispositional free riders do not free ride on punishment. *Nat. Commun.* **9**, 2390 (2018).
75. Peysakhovich, A., Nowak, M. A. & Rand, D. G. Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nat. Commun.* **5**, 4939 (2014).
76. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404 (2001).
77. Dohmen, T., Falk, A., Huffman, D. & Sunde, U. Homo reciprocans: survey evidence on behavioural outcomes. *Econ. J.* **119**, 592–612 (2009).
78. Yamagishi, T. et al. Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proc. Natl Acad. Sci. USA* **109**, 20364–20368 (2012).
79. Egloff, B., Richter, D. & Schmukle, S. C. Need for conclusive evidence that positive and negative reciprocity are unrelated. *Proc. Natl Acad. Sci. USA* **110**, E786–E786 (2013).
80. Eriksson, K., Cownden, D., Ehn, M. & Strimling, P. 'Altruistic' and 'antisocial' punishers are one and the same. *Rev. Behav. Econ.* **1**, 209–221 (2014).
81. Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G. & Cudré-Mauroux, P. The dynamics of micro-task crowdsourcing: the case of amazon MTurk. In *Proc. 24th International Conference On World Wide Web* 238–247 (International World Wide Web Conferences Steering Committee, 2015).
82. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* **5**, 411–419 (2010).
83. Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* **14**, 399–425 (2011).
84. Berinsky, A. J., Huber, G. A. & Lenz, G. S. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit. Anal.* **20**, 351–368 (2012).
85. Balafoutas, L., Nikiforakis, N. & Rockenbach, B. Direct and indirect punishment among strangers in the field. *Proc. Natl Acad. Sci. USA* **111**, 15924–15927 (2014).
86. Balafoutas, L., Nikiforakis, N. & Rockenbach, B. Altruistic punishment does not increase with the severity of norm violations in the field. *Nat. Commun.* **7**, 13327 (2016).
87. Raihani, N. J. & Bshary, R. The evolution of punishment in n-player public goods games: a volunteer's dilemma. *Evolution* **65**, 2725–2728 (2011).
88. Rand, D. G. & Nowak, M. A. The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2**, 434 (2011).
89. Garcia, J. & Traulsen, A. Leaving the loners alone: evolution of cooperation in the presence of antisocial punishment. *J. Theor. Biol.* **307**, 168–173 (2012).
90. McCabe, C. M. & Rand, D. G. Coordinated punishment does not proliferate when defectors can also punish cooperators. in: *Antisocial Behavior: Etiology, Genetic and Environmental Influences and Clinical Management* (ed. Gallo, J. H.) 1– 14 (Nova Publisher, 2014).
91. Huang, F., Chen, X. & Wang, L. Conditional punishment is a double-edged sword in promoting cooperation. *Sci. Rep.* **8**, 528 (2018).
92. Giamattei, M., Molleman, L., Seyed Yahosseini, K. & Gächter, S. LIONESS Lab—A Free Web-based Platform for Conducting Interactive Experiments Online. *SSRN* https://doi.org/10.2139/ssrn.3329384 (2019).
93. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data* (MIT Press, 2010).

Acknowledgements

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

Author contributions

L.M., F.K., C.S. and S.G. designed the study, L.M. and F.K. collected and analysed the data. L.M., F.K., C.S. and S.G. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0707-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.M. or F.K.

Peer review information: Primary Handling Editor: Stavroula Kousta

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Lucas Molleman

Last updated by author(s): Jun 13, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data for the punishment experiment was collected using the open source software LIONESS (experimental code available upon request). Data for the conditional cooperation experiment was collected using the software Qualtrics.
Data analysis	Stata 15.1, StataCorp LLC

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The dataset and analysis code for this study are available in via GitHub: http://www.github.com/LucasMolleman/NHB_CoordinatedPunishment

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study includes quantitative behavioural and questionnaire data from online experiments.
Research sample	Participants were US citizens who were registered on the online platform Amazon Mechanical Turk (MTurk). Overall, n=4,320 subjects participated in our experiment. 48% of all participants were female and the average age was 33.9 (std. dev. = 10.4).
Sampling strategy	We sampled participants from MTurk by posting advertisements ('HITs') which prospective participants could choose to complete. Due to the lack of directly comparable experiments in the literature, our first study did not use explicit sample size calculations (we estimated that overall punishment in our one-shot experiment would be low, but wanted variation in punishment strategies; so we aimed for 2,000 participants in total). In our replication experiment, we used the data from our first study to do a power analysis, presented in Figure S3.
Data collection	Data was collected with online experiments. Participants completed the tasks remotely through their web browsers. Roles in the experiment (Punisher or Target) were randomly allocated within matching groups. During data collection or analysis, the researchers were not blind to the study hypothesis.
Timing	The data for our main study were collected between April and May 2015. The data for our replication study were collected in August 2018.
Data exclusions	No data were excluded.
Non-participation	From MTurk we cannot tell how many participants chose to decline participation (not-accept our MTurk HIT) after browsing its (general) description.
Randomization	Participants were randomly assigned a role (Punisher or Target) and randomly allocated into groups. All Punishers encountered both relevant conditions in the strategy method (one where the other participant chose to punish, and one there they chose to not-punish). In the replication study, the order of the unconditional and conditional decisions was counterbalanced between interaction groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited via the online crowdsourcing platform Amazon Mechanical Turk.
Ethics oversight	Ethical approval was provided by the Research Ethics Committee at the School of Economics, University of Nottingham.

Note that full information on the approval of the study protocol must also be provided in the manuscript.