nature human behaviour

SUPPLEMENTARY INFORMATION

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

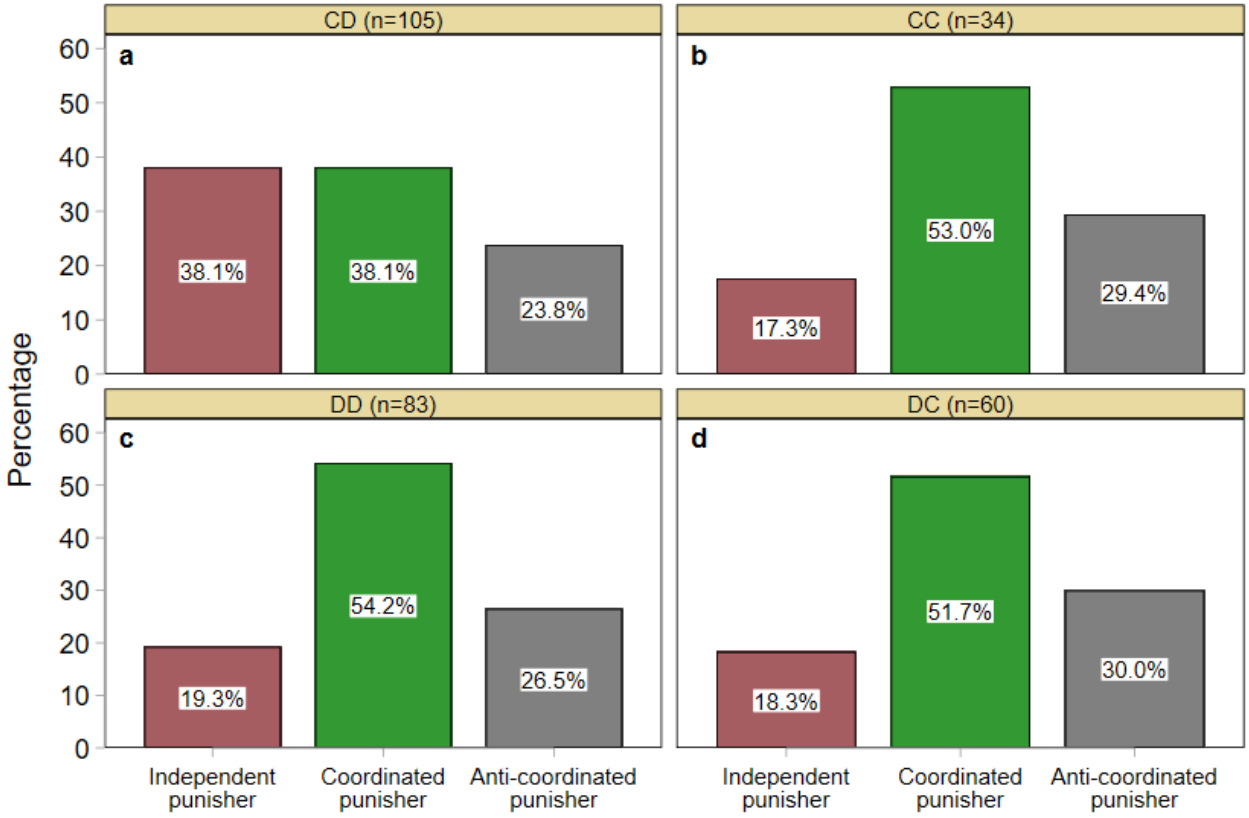
In the format provided by the authors and unedited.

People prefer coordinated punishment in cooperative interactions

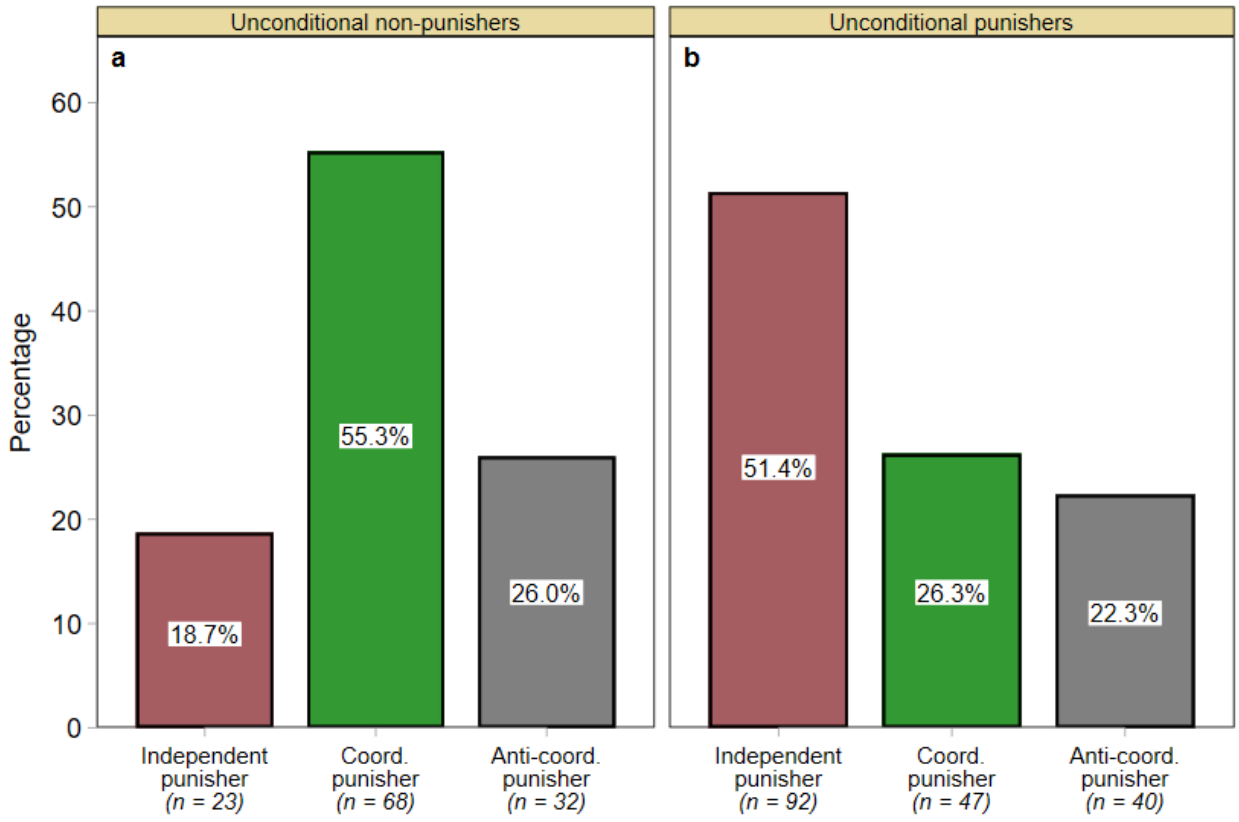
Lucas Molleman[1,2,3]*, Felix Kölle[2,4]*, Chris Starmer[2] and Simon Gächter[2,5,6]

[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

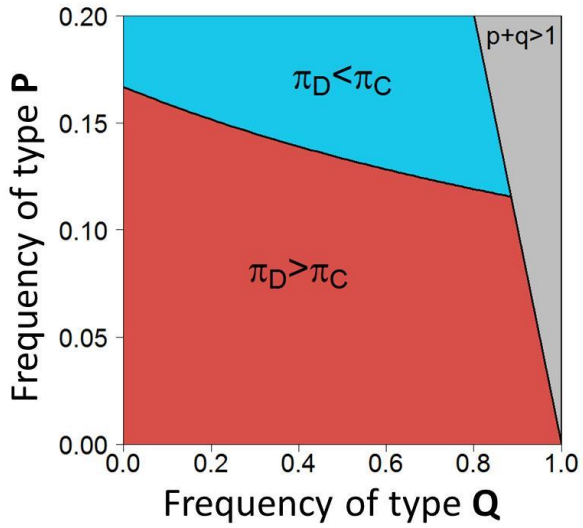
Supplementary Figures



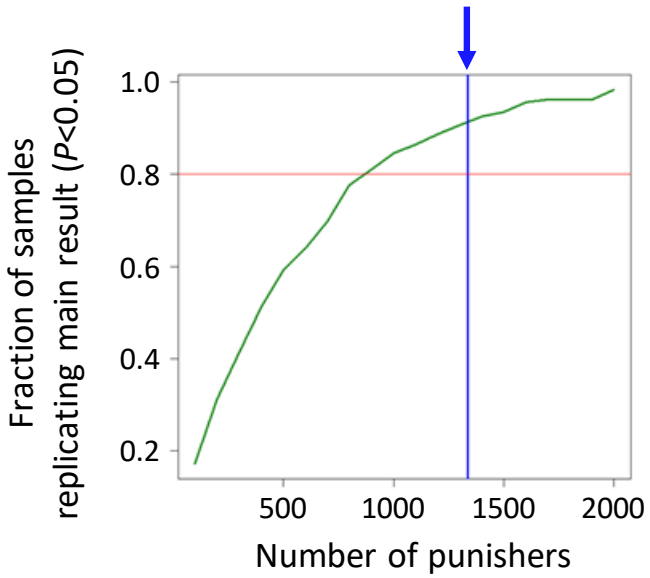
Supplementary Figure 1 | Conditional punishment types broken down by the outcome of the Public Goods Game (PGG). Bars reflect data from the strategy method, including only those individuals who punished at least once in the conditional stage (N = 282 out of the total of N = 1,336). **a,** Punisher cooperated, 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 depending 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"; 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 lines 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*2/3=) 1,200 participants. In our additional study, we 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 studies. '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. The 95% confidence intervals are in brackets and *P* values are in parentheses.

Dependent variable:	Punish (1 if yes, 0 otherwise)			
	Model (1)	Model (2)	Model (3)	Model (4)
Other punishes *(1 if other Punisher punished, 0 otherwise)*	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 *(1 if yes, 0 otherwise)*	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 *(1 if Target cooperated, 0 otherwise)*		-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 (using an alternative model specification). 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. The 95% confidence intervals are in brackets and *P* values are in parentheses.

Dependent variable:	Punish (1 if yes, 0 otherwise)		
	Model (1)	Model (2)	Model (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 Punisher cooperates and Target 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 Punisher defects and Target 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 Punisher and Target cooperate, 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 (reasons for decisions). 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 (personality and preferences). 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 (π_C) and for defection (π_D) can be written as

$\pi_C = b - c$, 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 $\pi_C > \pi_D$.

$\pi_C > \pi_D$ 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[1]. For simplicity, our analysis so far has only focused on punishment of defectors. Empirical evidence from a range of previous studies[2–4] as well as observations from our experiment (Supplementary Figure 1) indicate that at times, punishment is aimed at cooperators. Such anti-social punishment can have strongly detrimental consequences for cooperation[2,5–7]. Moreover, if individuals tend to coordinate their punishment towards cooperators, coordinated punishment may no longer be able to promote cooperation[8,9].

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-socially) 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 cooperation are then defined by:

$\pi_C > \pi_D$ 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 (see 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 (Note: 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'[10,11]]
2. When making my decisions, I was unsure what was the appropriate thing to do [concerns for social appropriateness or legitimitacy[12,13]]
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 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[14], gauged general dispositions towards positive and negative reciprocity[15] and elicited risk preferences[16].

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*'[17] 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 (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:

Option X

You: **+5**
Other 1: **+0**
Other 2: **+0**

Option Y

You: **+2**
Other 1: **+2**
Other 2: **+2**

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

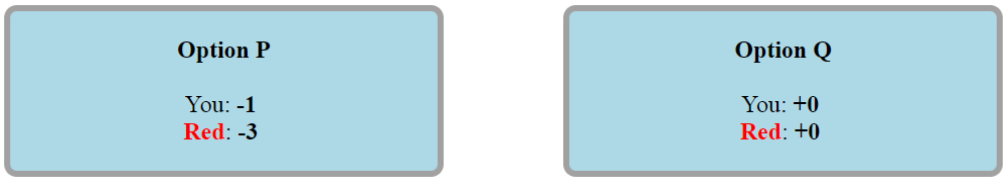
If you choose...	**and the two other participants choose...**	**then your earnings are:**
X	**X** and **X**	5
X	**X** and **Y**	7
X	**Y** and **Y**	9
Y	**X** and **X**	2
Y	**X** and **Y**	4
Y	**Y** and **Y**	6

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]]

Option X	or	**Option Y**
You: **+5** Other 1: **+0** Other 2: **+0**		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:

Option P	Option Q
You: **-1** **Red**: **-3**	You: **+0** **Red**: **+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:

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

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

b. 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.

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

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

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

[Continue]

[Back to instructions]

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

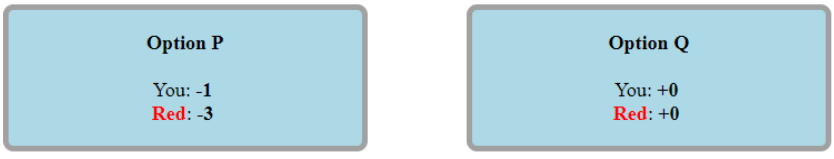
Step 1 (punishers only) [Replication experiment: "Your *first mover* decision"]

[[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.



[submit]

== [The *conditional* punishment decisions; both on the same screen; 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**?

Option P	Option Q
You: -1 Red: -3	You: +0 Red: +0

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

Option P	Option Q
You: -1 Red: -3	You: +0 Red: +0

[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?

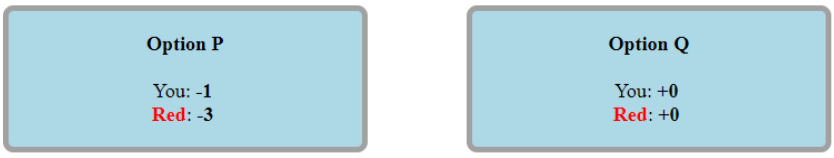
Not angry at all 0 0 0 0 0 0 0 Very angry

=== [Questionnaire eliciting possible motivations underlying the observed punishment preferences; Supplementary Results; 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:



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: **+3** They: **+0**	You: **+2** They: **+2**

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:

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

I choose

- ○ Option X
- ○ Option Y

Step 2

Recall the two choice options:

Option X	Option Y
You: **+3** They: **+0**	You: **+2** 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. Boyd, R., Gintis, H. & Bowles, S. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620 (2010).
2. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).
3. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *J. Public Econ.* **92**, 91–112 (2008).
4. Gächter, S., Herrmann, B. & Thöni, C. Culture and cooperation. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2651–2661 (2010).
5. Rand, D. G. & Nowak, M. A. The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2**, 434 (2011).
6. García, J. & Traulsen, A. Leaving the loners alone: Evolution of cooperation in the presence of antisocial punishment. *J. Theor. Biol.* **307**, 168–173 (2012).
7. 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).
8. 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. J.H. Gallo), pp. 1– 14. (Hauppauge, NY: Nova Publisher, 2014).
9. Huang, F., Chen, X. & Wang, L. Conditional punishment is a double-edged sword in promoting cooperation. *Sci. Rep.* **8**, 528 (2018).
10. 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).
11. Cialdini, R. B. & Cialdini, R. B. *Influence: The psychology of persuasion*. (Collins New York, 2007).
12. Boehm, C. *Hierarchy in the Forest: Egalitarianism and the Evolution of Human Altruism*. (Harvard University Press, 1999).
13. Wiessner, P. Norm enforcement among the Ju/'hoansi Bushmen. *Hum. Nat.* **16**, 115–145 (2005).
14. 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).
15. Perugini, M., Gallucci, M., Presaghi, F. & Ercolani, A. P. The personal norm of reciprocity. *Eur. J. Personal.* **17**, 251–283 (2003).
16. Dohmen, T. *et al.* Individual risk attitudes: Measurement, determinants, and behavioral consequences. *J. Eur. Econ. Assoc.* **9**, 522–550 (2011).
17. Arechar, A. A., Gächter, S. & Molleman, L. Conducting interactive experiments online. *Exp. Econ.* **21**, 99–131 (2018).