Communications Psychology https://doi.org/10.1038/s44271-025-00372-w

Article in Press

Individual differences in motives for costly punishment

Received: 31 July 2025

Accepted: 19 November 2025

Cite this article as: Claessens, S., Atkinson, Q.D., Raihani, N.J. Individual differences in motives for costly punishment. *Commun Psychol* (2026). https://doi.org/10.1038/s44271-025-00372-w

Scott Claessens, Quentin D. Atkinson & Nichola J. Raihani

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

ARTICLE IN PRESS

Individual differences in motives for costly punishment

Scott Claessens[1,2*], Quentin D Atkinson[2], and Nichola J Raihani[2,3]

[1]School of Psychology, University of Kent, Canterbury, United Kingdom

[2]School of Psychology, University of Auckland, Auckland, New Zealand

[3]Department of Experimental Psychology, University College London, London, United Kingdom

*Correspondence concerning this article should be addressed to Scott Claessens, School of Psychology, University of Kent, Canterbury, United Kingdom, Email: scott.claessens@gmail.com

ARTICLE IN PRESS

Author Note

Scott Claessens https://orcid.org/0000-0002-3562-6981

Quentin D Atkinson https://orcid.org/0000-0002-8499-7535

Nichola J Raihani https://orcid.org/0000-0003-2339-9889

This manuscript was published as a pre-print on PsyArXiv under a CC BY 4.0 license: https://doi.org/10.31234/osf.io/ys6rm Word count (excluding abstract, figures, tables, and references) = 6687 words.

ARTICLE IN PRESS

Abstract

Costly punishment is thought to be a key mechanism sustaining human cooperation. However, the motives for punitive behaviour remain unclear. Although often assumed to be motivated by a desire to convert cheats into cooperators, punishment is also consistent with other functions, such as levelling payoffs or improving one's relative position. We used six economic games to tease apart different motives for punishment and to explore whether different punishment strategies were associated with personality variables, political ideology, and religiosity. We used representative samples from the United Kingdom and the United States (N = 2010) to estimate the frequency of different punishment strategies in the population. The most common strategy was to never punish. For people who did punish, strategy use was more consistent with egalitarian motives than behaviour-change motives. Nevertheless, different punishment strategies were also associated with personality, social preferences, political ideology, and religiosity. Self-reports of behaviour in the games suggested that people have some insight into their punishment strategy. These findings highlight the multipurpose nature of human punishment and show how the different motives underpinning punishment decisions are linked with core character traits.

Keywords: punishment, cooperation, economic games

ARTICLE IN PRESS

ARTICLE IN PRESS

Individual differences in motives for costly punishment

Humans cooperate on a scale that is unparalleled in the animal kingdom. One mechanism thought to sustain this level of cooperation is costly punishment, whereby individuals harm others at a personal cost[1], ostensibly encouraging cooperative behaviour from the target or bystanders[2–4]. Punishment offers a route to maintaining or increasing cooperation by changing the payoff structure of social interactions such that it no longer pays to cheat or exploit social partners[1,5]. Yet, despite its theoretical importance, the question of why people choose to punish others is still hotly contested[6]. In this study, we use a battery of economic games to disentangle the different motives underpinning punishment and explore how these motives vary across individuals.

In humans, punishment is often studied in laboratory settings using economic games[6–15]. In some of these games, participants are given a sum of money that they can use to invest in collective action or to help others; conversely, they can ‘cheat’ by keeping the money for themselves or by exploiting the contributions of others. In other games, participants are endowed with sums of money and can then ‘cheat’ by destroying a partner’s payoffs or stealing from them[16,17]. Punishment is introduced into such games by giving participants the option to pay a small ‘fee’ to impose a greater ‘fine’ on their co-players. Several lines of experimental evidence indicate that people use this punishment option[13], that they enjoy punishing[18], and that they frequently, though not always[19], punish those who cheat, free-ride, or steal[11,12].

Evidence from these experiments suggests that the threat of costly punishment plays an important role in promoting human cooperation. People tend to cooperate more in games where punishment is possible compared to those where it is not[6–8]. Similarly, people make higher offers in the Ultimatum Game (where punishment is possible) compared to the structurally-similar

ARTICLE IN PRESS

ARTICLE IN PRESS

Dictator Game (where it is not)[20]. This cooperation-enhancing effect of punishment has also been observed across societies[8], leading some to suggest that costly punishment has played a key role in the cultural evolution of cooperation in humans[21–24].

Nevertheless, it remains unclear whether individuals playing economic games use punishment as a behaviour-change tool to enforce cooperation or to achieve other ends. Some have argued that punishment is primarily used to shape behaviour, either to deter personal harm[3,10,25] or to uphold normative standards of cooperative behaviour[22,23,26–29]. But while the *threat* of punishment can have a cooperation-enhancing effect, the *enactment* of this punishment does not consistently deter targets from cheating in the future[6]. This calls into question whether punishment primarily operates as a behaviour-change tool or whether it is used to achieve other goals.

Beyond behaviour shaping concerns, there are other reasons why people punish in economic games. Punishers might be motivated by a desire for revenge rather than deterrence, punishing in response to harm that was personally incurred[30]. Punishment might be driven by concerns about relative payoffs, such as disadvantageous inequity aversion (i.e., avoiding having less than others)[6,17] and/or general egalitarian preferences (i.e., wanting all participants to receive the same payoffs)[31]. Such concerns about relative payoffs may be activated when participants earn less than cheats in economic games or when there are income disparities in these settings. People might also use punishment for competitive purposes, seeking advantageous inequity for themselves (i.e., having more than others) and/or improving their relative position[6].

Common economic game designs have been unable to tease apart these different motives for punishment because participants who interact with cheats in these games experience both losses *and* lower relative payoffs. The typical 1:3 fee-fine ratio of punishment in economic

ARTICLE IN PRESS

ARTICLE IN PRESS

games compounds this issue. With this setup, people can simultaneously use punishment to reciprocate losses, to deter others from cheating, and to reduce or reverse disparities in payoffs between themselves and targets. To add to this complexity, it is evident that people use punishment in seemingly disparate ways: punishing when no behaviour change is possible, such as in one-shot games[13,17,32,33], on the very last round of repeated games[34], or in games where the target never learns about the punishment[35]; punishing those who did not cheat or who over-contributed to collective action (sometimes called 'antisocial punishment')[19,36]; punishing in scenarios where they were not personally harmed (third-party punishment)[37]; and punishing in scenarios where disparities in payoffs did not arise from participants' actions[31,38,39].

The general conclusion from this research is that there is no one unifying function of costly punishment in humans. Instead, punishment should be thought of as a flexible behavioural tool that serves a variety of functions that are not mutually exclusive[6]. Due to its multipurpose nature, we should therefore expect variation in punishment strategies in the population, much like the observed variation in social learning strategies[40]. Some individuals may use punishment as a behaviour shaping tool, for example, while others may use it to reduce or reverse payoff differentials. This raises the underexplored question of which punishment strategies are more frequent in human populations.

Here, we aim to delineate nine possible punishment strategies by asking whether people punish in a manner consistent with a specific strategy across a suite of economic games. Table 1 summarises the potential strategies for costly punishment in the economic games that we considered, and the behavioural patterns they predict. These proximate strategies are consistent with different ultimate functions of punishment that have been emphasised in the literature, such as changing how others behave towards the punisher[1,15], changing relative payoffs between the

ARTICLE IN PRESS

ARTICLE IN PRESS

punisher and the target[6,41], and improving cooperation within the cultural group such that the group is more successful than other groups[22]. Note that Table 1 is not an exhaustive list of all possible punishment strategies: we do not include proximate punishment strategies that serve reputational functions, such as signalling trustworthiness[4,42–45], because our focus is on punishment strategies in anonymous economic games without reputational incentives (but see ref[46]).

Building on previous designs that have used one-shot economic games to explore behaviour-shaping and egalitarian motives for punishment[17,32,35,47,48], we employ a suite of one-shot economic games where individuals are given the opportunity to punish targets at a personal cost (Figure 1). In each game, targets either steal from another individual or do nothing. Representative samples of participants from the United Kingdom (*n* = 1014) and the United States (*n* = 996) completed all six games on the online platform Prolific. We designed the suite of games to tease apart the proposed punishment strategies in Table 1, such that each strategy predicts a different pattern of behaviour across all the games (see Methods for more detail about the six games). We use the resulting behavioural patterns to discern which punishment strategy participants are employing.

ARTICLE IN PRESS

We also explore whether individual differences in personality, religiosity, political ideology, and demographics are associated with variation in punishment strategies. Prior work has shown that punitive behaviour in economic games is predicted by personality, social preferences, religiosity, and political ideology. For example, people who are more agreeable or more prosocial are less likely to punish in economic games[49–51], whereas political conservatives and those higher on measures of right-wing authoritarianism express more punitive sentiment[52,53]. Other associations are less clear: for example, religiosity can show a positive,

ARTICLE IN PRESS

negative, or no association with punitive behaviour[54–56]. However, because previous studies have not distinguished between different punishment strategies, the proximate motivations underlying these trait correlations remain unclear. We aim to fill this gap by linking our economic game data to a host of individual difference measures collected in a separate survey. Although we pre-registered our study design and analysis plan (see Methods), we note that we did not pre-register any directional hypotheses for these exploratory analyses.

Finally, we ask whether people have insight into their own punishment strategy. Prior work has argued that people are often unaware of the underlying function of their punitive behaviour, yet they feel compelled to enact it anyway[30]. We assess this claim by asking people to explicitly report the reasons for their behaviour in the games, exploring whether these self-reported reasons align with patterns of punitive behaviour.

ARTICLE IN PRESS

Methods

Ethical approval

This research was approved by the UCL Department of Psychology Ethics Committee (project: 3720/002) and ratified by the University of Auckland Human Participants Ethics Committee. The study was performed in accordance with all the relevant guidelines and regulations. Informed consent was obtained from all participants prior to the study and participation was voluntary.

Pre-registration

We pre-registered the study on the Open Science Framework before collecting data in the United Kingdom (11th November 2022; https://osf.io/k75fc). We submitted another pre-registration before collecting data in the United States (20th June 2023; https://osf.io/q4hdy). In the pre-registrations, we outlined our study design, exclusion criteria, and analysis plan. As the

ARTICLE IN PRESS

study was exploratory, we did not pre-register any explicit hypotheses. We did not deviate from the pre-registrations.

Exclusion criteria

We pre-registered that we would exclude participants who failed any of the attention checks, sped through the surveys (i.e., two standard deviations below the median duration), or flatlined (i.e., provided identical responses to matrix questions). We also stated that we would exclude data for particular games if participants failed the comprehension question for that game. We followed our pre-registered plan of conducting analyses with and without these exclusions (analyses without exclusions are reported in the Supplementary Material).

Participants

We collected a representative sample of 1019 participants from the United Kingdom through the online platform Prolific (https://www.prolific.com/). All of these participants completed the economic games and 973 returned to complete the follow-up survey a week later (95% retention rate). After exclusions, we were left with 1014 participants overall (513 female, 481 male, 20 unspecified gender; 72 Asian, 30 Black, 16 Mixed, 865 white, 11 Other, 20 unspecified ethnicity; mean age = 45.78; see Supplementary Figure 1).

We later collected a representative sample of 1005 participants from the United States through Prolific. All of these participants completed the economic games and 957 returned to complete the follow-up survey (95% retention rate). After exclusions, we were left with 996 participants overall (504 female, 482 male, 10 unspecified gender; 53 Asian, 127 Black, 19 Mixed, 774 white, 13 Other, 10 unspecified ethnicity; mean age = 45.65; see Supplementary Figure 2).

ARTICLE IN PRESS

ARTICLE IN PRESS

Materials

Economic games

In the first part of the study, participants completed six economic games, each with slight variations. In all games, there are multiple players and the participant takes the role of P1. P2 either (a) steals £0.20 from another player and adds it to their payoff or (b) does nothing. Following the strategy method, participants responded to each of these possible decisions. For each of these cases, participants are asked whether they would like to pay money to reduce P2's payoff. Games A-E have two players and Game F has three players.

The six games are as follows (variations bolded; see Figure 1 for a visual representation of the games):

1. *Game A (Advantageous Inequity)*. P1 starts with £0.70 and P2 starts with £0.10. P2 is given the option to either steal £0.20 from P1 or do nothing. P1 can then pay £0.10 to reduce P2's payoff by £0.30.
2. *Game B (Equal)*. P1 starts with £0.70 and P2 starts with **£0.30**. P2 is given the option to either steal £0.20 from P1 or do nothing. P1 can then pay £0.10 to reduce P2's payoff by £0.30.
3. *Game C (Computer)*. P1 starts with £0.70 and P2 starts with £0.30. Participants are told that **"the computer will decide"** whether P2 steals £0.20 from P1 or does nothing. P1 can then pay £0.10 to reduce P2's payoff by £0.30.
4. *Game D (1:1 Fee-Fine)*. P1 starts with £0.70 and P2 starts with £0.30. P2 is given the option to either steal £0.20 from P1 or do nothing. P1 can then pay £0.10 to reduce P2's payoff by **£0.10**.
5. *Game E (Disadvantageous Inequity)*. P1 starts with £0.70 and P2 starts with **£0.50**. P2 is given the option to either steal £0.20 from P1 or do nothing. P1 can then pay £0.10 to reduce P2's payoff by £0.30.
6. *Game F (Third-Party)*. P1 starts with £1.00, P2 and P3 start with £0.70. P2 is given the option to either steal £0.20 **from P3** or do nothing. P1 can then pay £0.10 to reduce P2's payoff by £0.30.

We delineate nine punishment strategies that can be isolated based on players' decisions across the six economic games (Table 1). In existing literature, it is often unclear whether punishment decisions are sensitive to the previous actions of others or to payoff sensitivity. This

ARTICLE IN PRESS

is because (*i*) players' decisions typically introduce payoff differentials in economic games and (*ii*) where fine > fee, punishment tools can be used to harm co-players and to change relative payoffs[6]. Of the nine strategies we define, *deterrent*, *norm-enforcing*, *revenge* and *antisocial* are strategies that are only sensitive to the actions of others. Conversely, *avoid disadvantageous inequity*, *egalitarian*, *seek advantageous inequity* and *competitive* are strategies that are sensitive to payoffs. *Never punish* is an unconditional strategy (i.e., no sensitivity to actions or payoffs).

Our strategy definitions sometimes differ from how these strategies are defined and discussed in the literature (e.g., antisocial punishment). This is essential because previous definitions permit many possible motives, while we are aiming to isolate punishment motives. We define our strategies as follows:

- *Deterrent*: Deter others who have harmed you from harming you again in the future (when the target steals in Games A, B, D, and E).
- *Norm-enforcing*: Punish to enforce a shared anti-harm norm and encourage future norm compliance (when the target steals in Games A, B, D, E and F). Though we cannot measure future norm compliance in our one-shot games, these non-repeated (and 'stranger design') games have been used to identify norm-enforcing motives in many previous studies[11,12].
- *Revenge*: Punish if doing so harms another who harmed you (when the target steals in Games A-E).
- *Avoid disadvantageous inequity*: Punish if doing so avoids disadvantageous inequity for the self (when the target steals in Game E).
- *Egalitarian*: Punish if doing so makes payoffs for all more equal (when the target steals in Games E and F) (c.f. ref[31]).
- *Seek advantageous inequity*: Punish if doing so produces advantageous inequity for the self (when the target steals in Games B and C). This is one possible motive underlying what others have previously called 'antisocial punishment' or 'spite'; here we isolate it.
- *Competitive*: Punish if doing so improves your relative position (all games except Game D).
- *Antisocial*: Punish exclusively those who do not cause harm (punish non-stealing target in all games). This definition differs from some previous definitions of antisocial punishment, but this is essential to isolate the full range of motives that could be driving punishment and which are not typically disentangled in other settings.
- *Never punish*: A person who never incurs a personal cost to punish a co-player (does not punish in any game).

ARTICLE IN PRESS

For each game, participants saw the game instructions and answered a comprehension question before providing their decisions. After completing all the games, participants were asked to give an open-ended response explaining their behaviour in the games, and then responded to several slider questions capturing the different reasons for their decisions (for full wordings, see Supplementary Table 1).

Survey questions

In a follow-up survey, we collected the following data on participants (for wordings of all questions, see Supplementary Table 2):

- *Demographics*. In the survey, we collected information on participants' education level and self-reported socio-economic status (MacArthur ladder[57]). We also collected additional demographic data from Prolific (e.g., age, gender, student status).
- *Personality*. We used the Mini-IPIP scale[58] to measure the Big 6 personality dimensions of agreeableness (α = 0.83), conscientiousness (α = 0.75), extraversion (α = 0.83), honesty-humility (α = 0.77), openness to experience (α = 0.79), and neuroticism (α = 0.79). Four items were used for each personality dimension.
- *Social Value Orientation*. We used the Social Value Orientation Slider Measure to measure other-regarding preferences[59]. Across fifteen items, participants made decisions on how to allocate different amounts of money between themselves and another anonymous individual. From these decisions, we calculated participants' Social Value Orientation "angle" as a measure of their other-regarding preference, following the steps outlined in ref[59].
- *Political ideology*. We included several measures of political ideology, including left-right conservatism, Social Dominance Orientation[60] (α = 0.91; eight items), and Right Wing Authoritarianism[61] (α = 0.82; six items). We also probed participants' views on social inequality by asking them whether they would like to bring people above (below) them on the MacArthur socio-economic status ladder down (up) a peg or two.
- *Religious views*. We asked participants how religious they consider themselves and whether they believe that God or another spiritual non-human entity controls the events in the world[55].

ARTICLE IN PRESS

Procedure

We began data collection in the United Kingdom on 28th November 2022, with participants returning to complete the follow-up survey on 5th December 2022. We then ran a

ARTICLE IN PRESS

second wave of data collection in the United States on 20th June 2023, with participants returning to complete the follow-up survey on 27th June 2023. Our surveys were designed through the online survey platform Qualtrics (https://www.qualtrics.com/).

In the initial games survey, participants completed all six economic games in a random order, with punishment decisions (whether to punish a stealing target and whether to punish a target who did nothing) randomised within games. Responses to comprehension questions suggested that participants understood the six economic games (Supplementary Table 3). In order to partially mitigate the fact that some punishment strategies predict more punishment than others and are thus more "expensive" to implement (Table 1), we chose a random game to determine bonus payments rather than summing participants' earnings across all games[62]. After all games, 62% of participants stated that they believed that their decisions had real consequences for others.

In the follow-up survey, participants completed blocks of questions on demographics, personality, Social Value Orientation, political ideology, and religious views in a random order, with questions randomised within blocks. A random decision from the Social Value Orientation Slider Measure was chosen to determine bonus payment.

Participants were paid £1.80 for completing the games survey, plus a bonus payment from the six economic games (between £0.40 – £0.70 depending on their decisions). Participants were paid £1.50 for completing the follow-up survey, plus a bonus payment from the Social Value Orientation Slider Measure (between £0.50 – £0.85 depending on their decisions).

Statistical analysis

We pre-registered that we would use a Bayesian latent state model to infer unobserved punishment strategies from the observed data (for a similar version of this model, see ref[63]). In

ARTICLE IN PRESS

ARTICLE IN PRESS

this model, participants i in countries c make binary punishment decisions across twelve decisions j. We assume that the probability of the observed data $y_{i,j}$ is the weighted average of the probability of the observed data conditional on each of the ten punishment strategies s. From this logic, the model estimates the probability of each strategy p_s. The full model is as follows:

$$\begin{aligned} y_{i,j} &\sim \text{Bernoulli}(\theta_j) \\ \theta_j &= \sum_{s=1}^{10} p_s \Pr(\text{punish}|s,j) \\ p &= \text{softmax}(\alpha_{c[i]}) \\ \alpha_{s,c} &\sim \text{Normal}(0, 1) \end{aligned} \tag{1}$$

The conditional probabilities $\Pr(\text{punish}|s,j)$ are hard coded in the model as outlined in Table 1. We incorporate an implementation error rate δ into these conditional probabilities by coding ticks in Table 1 with a conditional probability of $1 - \delta$ and coding crosses with a conditional probability of $0 + \delta$. We set δ to 0.05 in all models, which is similar to its value when we estimate it as a free parameter in an additional model (median posterior δ = 0.03, 95% CI [0.00 0.06]; Supplementary Figure 3). The random choice strategy is consistently coded with a conditional probability of ½ across all decisions.

To include a categorical predictor in the model, we estimate a different $\alpha_{s,c}$ for each categorical level. To include a continuous predictor x in the model, we include a slope β in the linear model for p:

$$\begin{aligned} y_{i,j} &\sim \text{Bernoulli}(\theta_j) \\ \theta_j &= \sum_{s=1}^{10} p_s \Pr(\text{punish}|s,j) \\ p &= \text{softmax}(\alpha_{c[i]} + \beta_{c[i]} x_i) \\ \alpha_{s,c} &\sim \text{Normal}(0, 1) \\ \beta_{s,c} &\sim \text{Normal}(0, 0.2) \end{aligned} \tag{2}$$

ARTICLE IN PRESS

These models control for multiple comparisons in several ways. First, the models estimate the effects of the predictor on all strategies simultaneously (though different predictors are necessarily included in different models). Second, the models employ strongly regularising priors on the slope parameters $\beta_{s,c}$, which makes our estimates more conservative and less susceptible to random noise[64].

We estimated the posterior distributions of these models using Hamiltonian Monte Carlo as implemented in Stan version 2.32.2[65]. We ran each model for 2000 samples, with 1000 warmup samples. R-hat values and effective sample sizes suggested that all models converged normally. Trace plots are reported in Supplementary Figure 4.

We validated the model by simulating observed data (*n* = 100) from a known frequency of strategies. The model was successfully able to recover the known frequency of strategies from the simulated data (Supplementary Figure 5).

Reproducibility

All data and code are accessible on GitHub[66]: https://github.com/ScottClaessens/punishStrategies. All analyses were conducted in R version 4.4.2[67]. Visualisations were created with the *ggplot2*[68] and *cowplot*[69] R packages. We used the *targets*[70] R package and *quarto*[71] to reproducibly generate the manuscript.

Results

The overall pattern of punitive behaviour in the six economic games was very similar across both countries (Figure 2). Participants were generally more likely to punish targets who stole compared to targets who did not steal (multilevel logistic regression; *b* = 1.93, standard error = 0.27, 95% confidence interval [1.40 2.46], *p* < .001). Participants were also more likely to

ARTICLE IN PRESS

ARTICLE IN PRESS

punish when targets' stealing behaviour generated inequalities, specifically in Games E and F (*b* = 2.42, SE = 0.44, 95% CI [1.56 3.27], *p* < .001).

We classified participants into a particular strategy if their behaviour across all twelve decisions matched our behavioural predictions shown in Table 1 exactly. Table 2 shows the proportion of participants following each strategy, with N/A used to represent participants who did not fit exactly into any particular strategy type. Overall, 59% of our participants could be classified exactly into one of the strategies. The most common strategy in both countries was to never punish across any of the games. The next most common strategies were those that care about minimising payoff differences (avoid disadvantageous inequity, egalitarian). Less common were the behaviour shaping strategies (deterrent, norm-enforcing), the revenge strategy, and the competitive strategies (seek advantageous inequity, competitive). Although participants often punished targets who did not steal in the six games (Figure 2), no participants followed the antisocial strategy by exclusively punishing targets who did not steal across *all* games.

To further investigate the strategies that participants were following, we examined the most common patterns of punitive behaviour across all twelve decisions. Supplementary Table 4 shows the proportion of participants following the 25 most common behavioural patterns, including, where appropriate, the predetermined strategies from Table 1. In both countries, a common pattern of behaviour not captured by any of the strategies was punishing only when the target stole in the third-party game (Game F). Other behavioural patterns not captured by our strategies included punishing whenever the target stole across all games and always punishing in every game irrespective of the targets' behaviour.

While it is useful to look at exact patterns of behaviour, participants may not have implemented their chosen punishment strategy with exact precision. In reality, strategies may

ARTICLE IN PRESS

ARTICLE IN PRESS

have been implemented probabilistically for each punishment decision. There is also the possibility of implementation errors, whereby participants occasionally "slip up" and make decisions that are incongruent with a particular strategy. This may explain why some participants were unable to be classified exactly into a single punishment strategy.

To deal with this complexity and include all observed data in our frequency estimates, we fitted a Bayesian latent state model to the data. This model assumes that the nine strategies in Table 1 (plus a "random choice" strategy that chooses randomly for each decision) are the only latent strategies and that these are instantiated into observed behaviour according to the logic in Table 1 with some probability of implementation error (i.e., an intention to punish is implemented as non-punishment and vice versa). Averaging over all strategies and incorporating the possibility of implementation errors, the model estimates the probability of participants following any particular strategy, given the observed data.

The posterior estimates from the model are presented in Figure 3. The posterior probabilities for each strategy did not differ between the two countries. In both countries, the never punish strategy had the highest probability, followed by the egalitarian strategy. The norm-enforcing and seek advantageous inequity strategies were the next most likely, with higher posterior estimates than the competitive and antisocial strategies. None of the other strategies differed in their posterior estimates. The same general pattern emerged when we analysed the full dataset without pre-registered exclusions (Supplementary Figure 6).

This pre-registered version of the model includes our *a priori* strategies from Table 1, but does not include other strategies that turned out to be common in the raw data. In particular, the strategy that punishes only when the target stole in the third-party game (Game F) is potentially interesting to include, as third-party punishment has been argued to reflect a uniquely-human

ARTICLE IN PRESS

ARTICLE IN PRESS

normative psychology[37]. We therefore ran the model again including this additional strategy (Supplementary Figure 7). The results of this model show more tempered support for the egalitarian strategy, and increased support for both the avoid disadvantageous inequity strategy and the third-party strategy, in line with the raw sample counts. However, this result is difficult to interpret, for a number of reasons. First, the behaviour of this strategy (punishing *only* in third-party contexts and *never* in second-party contexts) does not cohere with previous work showing that people who engage in third-party punishment often also engage in second-party punishment[72]. Second, while Game F is the only third-party game, it is also the game in which stealing results in the most inequality between players (a disparity of £0.40) making it difficult to know what is driving people's behaviour. For these reasons, we did not include the third-party strategy in our subsequent analyses.

After estimating the relative frequencies of different strategies, we explored which traits predicted adherence to the punishment strategies. To answer this question, we included variables capturing demographics, personality, social preferences, political views, and religious views as predictors in our Bayesian latent state model. We included each variable in a separate model, predicting all ten punishment strategies (the nine from Table 1, plus the 'random choice' strategy) simultaneously.

Demographic variables tended to be unrelated to strategy usage: age and gender did not predict adherence to a particular punishment strategy (Supplementary Figures 8 and 9). In the United States, the never punish strategy was slightly more common among participants lower in socio-economic status (median posterior slope = -0.20, 95% CI [-0.38 -0.03]) but this effect was small.

ARTICLE IN PRESS

Conversely, personality and social preferences were linked to variation in punishment strategies. When including the Big-6 personality dimensions and Social Value Orientation (SVO) in the model, we found associations with the egalitarian, never punish, and random choice strategies, with small-to-medium effect sizes (Figure 4). More prosocial participants (those with higher SVO scores) were more likely to follow the egalitarian and the never punish strategies, while those with lower SVO scores were more likely to enact the random choice strategy. The personality dimensions of honesty-humility and openness to experience were both positively associated with following the never punish strategy, while extraversion negatively predicted this strategy. The effects were mostly similar across countries, but occasionally differed: for example, in the United States, but not in the United Kingdom, honesty-humility was positively associated with following the egalitarian strategy and negatively associated with following the random choice strategy. Overall, the same pattern of results emerged when analysing the full dataset without exclusions (Supplementary Figure 10).

ARTICLE IN PRESS

Political and religious variables were also associated with punishment strategy (Figure 5). These effects were small-to-medium in size and tended to be more pronounced in the United States. Controlling for Social Dominance Orientation, American participants higher in Right Wing Authoritarianism were more likely to follow the strategies avoiding disadvantageous inequity and seeking advantageous inequity. Participants who stated that they would like to “bring those below them [on the socio-economic status ladder] up a peg” were more likely to follow the egalitarian strategy, while American participants higher in Social Dominance Orientation, Right Wing Authoritarianism, and believing that God controls events in the world were less likely to follow the egalitarian strategy. In general, religious and conservative

ARTICLE IN PRESS

participants were less likely to follow the never punish strategy. This general pattern of results was replicated with the full dataset (Supplementary Figure 11).

Finally, we asked whether participants had insight into their own punishment strategy. In other words, could participants self-report the strategy that they were following during the games? To answer this question, we included participants' responses to post-game questions about their strategy as predictors in the model. As before, each predictor was included in a separate model, predicting all ten strategies simultaneously.

In general, we found that self-reported strategy usage was positively associated with the behavioural strategy that participants employed, with effects ranging from small to large in size (see Figure 6 and Supplementary Figure 12 for the distribution of responses to self-report questions). Figure 7 shows the relationships between self-report questions and the different punishment strategies, highlighting the combinations where the question matched the behavioural strategy. We found positive relationships between the self-report questions and strategy usage for the norm-enforcing, egalitarian, seek advantageous inequity, never punish, and random choice strategies. The 95% credible intervals for other estimates included zero, though these estimates often trended in a positive direction. The same pattern of results was found when analysing the full dataset without exclusions (Supplementary Figure 13).

ARTICLE IN PRESS

Notably, there were several instances where self-reported strategy usage predicted a *different* behavioural strategy. For example, the random choice strategy is predicted by an intention to "make decisions at random" but also an intention to "punish people who did not harm". One potential explanation for this is that strategies are often motivated by multiple related reasons. Since participants reported their reasons for punishing on several slider scales, they were able to report several related motives for their behaviour. Indeed, self-reported intentions to make decisions at random and punish people who did not harm were positively correlated with one another (ρ = 0.58, 95% CI [0.55 0.61], p < .001). Supplementary Figure 14 reports the full correlation matrix for all self-reported intentions items.

Discussion

Using a suite of economic games measuring punishment in different situations, we have shown that punishment does not serve just one function, but instead is a flexible tool that can be and is used for different purposes[6]. While some use punishment to enforce norms of cooperation, others use it to reduce or even create inequality between individuals. We found that people's punishment strategy can, to some extent, be predicted by individual differences in personality, social preferences, and political and religious views. Moreover, contrary to the view that people are often unable to articulate the reasons for their punitive behaviour[30], people seem to have some degree of insight into the strategy they are using. Despite small differences, these general patterns replicated in samples from both the United Kingdom and the United States, providing further confidence in the results.

ARTICLE IN PRESS

Among the punitive strategies, the most common were particularly sensitive to inequality in payoffs, either from a self-referential perspective (i.e., avoid disadvantageous inequity) or more generally (i.e., egalitarian). This is in line with previous studies which have highlighted inequity aversion as an important motivation for punishment in economic games[17,31,32,38,47,73]. Behaviour shaping strategies, such as deterrence and norm-enforcement, were less common than strategies sensitive to inequality in our set of games. This was reflected both in participants' elicited punishment behaviour (Figure 3) and in their self-reports of their own strategy (Figure 6 and Supplementary Figure 12). Although our design did not explicitly allow for behaviour shaping as the interactions were all one-shot, we did manipulate whether the target's stealing behaviour was intentional or not (Game C), an approach which has been used in previous studies[74,75]. The lower prevalence of behaviour shaping strategies in our study is consistent with prior work showing that punishment often continues to be used when behaviour shaping is impossible, such as when the target will never find out that they have been punished[35] or on the last round of repeated interactions[34]. We also found that participants accurately reported using the norm-enforcing strategy, but not the deterrent strategy. This finding is in line with previous research showing that people struggle to accurately report the deterrent motivations for their punitive behaviour[30].

We defined antisocial punishment as occurring when a participant exclusively punished those who did not steal. There were no participants in our sample who followed this strategy, although some players did punish non-stealing co-players. Harming non-stealing individuals was also consistent with the competitive strategies, which did appear in our sample albeit at low frequencies. Previous work has suggested that another motive underpinning antisocial punishment is the desire to improve one's own relative payoffs[6,36] – though our data partially support this idea, if antisocial punishment was purely motivated by competitive concerns, we would also have expected to observe a reduction in antisocial punishment in Game D, where the fee-fine ratio was 1:1. That we did not suggests that the picture is more complex and that, at least in this study, antisocial punishment may not be attributable to one single motive.

The fact that people use punishment for many different reasons poses problems for the way that punishment is operationalised in classic behavioural economic game studies. In these studies, a common assumption is that participants punish to change the behaviour of cheats[11,12]. But in reality, people may be choosing the punishment option to achieve a variety of different goals. This has implications for how people respond to being targeted by punishers in these

ARTICLE IN PRESS

ARTICLE IN PRESS

games. Targets of punishment in these studies may know that punishment reflects different motives and can respond accordingly. For example, if targets interpret punishment as serving a competitive motive, it may elicit retaliation rather than encourage cooperation[6,10,76]. As punishers' motives must be inferred (and such inferences likely depend on character traits of the target, as well as the context in which punishment occurs), there is likely to be some variation and error in attributing motives to punishers. To the extent that inferred motives affect target responses, this might help to explain the mixed findings in the field as to whether punishment actually motivates cheating targets to subsequently cooperate[6].

It is striking that the most common strategy in our dataset was to never punish. This is partly because punishment in these games imposes an economic cost for no tangible benefit. If the fee-fine ratio had been lower, such that it was cheaper to punish, we may have seen more punishment from participants. Indeed, 72% of participants following the never punish strategy positively stated that they didn't want to pay to reduce anyone's bonus but would have done so if it were free. But the frequency of the never punish strategy perhaps also reflects a more general aversion to peer punishment, an aversion that has been highlighted in both WEIRD (Western, educated, industrialised, rich, and democratic) samples[77,78] and in small-scale societies[79]. One reason that people may be averse to peer punishment is that, because it is a fundamentally harmful act, punishment can reflect badly on the punisher and people might therefore refrain from punishing others to avoid reputational damage[4]. People frequently avoid taking actions that could harm their reputation, even when they don't know if reputation is at stake (as in the one-shot anonymous settings used here). Another reason that people may be averse to peer punishment is that it can trigger retaliation[6]. This may be especially likely in situations that lack clear institutional norms to legitimise punishment, such as our economic games. As with

ARTICLE IN PRESS

ARTICLE IN PRESS

reputation damage, people might abstain from peer punishment to avoid retaliation, regardless of whether retaliation is actually possible. By contrast, institutionalised punishment in small-scale societies often functions to compensate victims while limiting the potential for feuds and cycles of retaliation[80,81]. Future research should uncover whether people are more willing to punish in these conventionalised contexts (e.g., see ref[82]).

Our exploratory investigation of individual differences in punishment strategies sheds light on prior findings. For example, previous studies have found higher levels of punishment among people who are less prosocial, less agreeable, and more politically conservative[49–52]. Our results add further nuance to these previous findings. In our exploratory analyses, participants with prosocial value orientations were indeed more likely to follow the "never punish" strategy – but if they did punish, they were more likely to do so for egalitarian reasons. Agreeable participants in the United States also punished for egalitarian reasons. By contrast, religious and conservative individuals, including individuals higher in SDO and RWA, were less likely to follow the egalitarian strategy and more likely to follow the avoid disadvantageous inequity strategy, especially in the United States. These results suggest that people with different traits may be motivated to punish others for different reasons.

ARTICLE IN PRESS

Limitations

One potential limitation with our design is that some strategies required more punishment than others, meaning that some strategies were more "expensive" to implement. For example, the competitive strategy required punishment in ten of twelve decisions, compared to the avoid disadvantageous inequity strategy which required only one instance of punishment (Table 1). To deal with this, we calculated participant payoffs from a randomly-chosen game instead of accumulating the costs across all games, mitigating concerns about wealth and portfolio effects whereby participants keep track of their overall earnings over the course of the study[62]. But even accounting for this, one could still argue that the difference in overall costs explains why, for example, the competitive strategy is less common in our dataset than the avoid disadvantageous inequity strategy. We do not think this feature of our design is a concern, however, for a number of reasons. First, we are interested in measuring strategies underlying *costly* punishment. Some of these strategies, by their very nature, will be more costly to implement than others. This is reflected in our design. It would not make theoretical sense to use an alternative design where more or less punitive strategies are manipulated to cost participants the same amount. Second, when we asked participants whether they would have punished if it were free to do so, we found low agreement with this statement (Figure 6), suggesting that participants were not particularly sensitive to the costs of punishment. In line with this, when we plot the frequencies of different strategies against their expected costs, we find that cost is not a perfect predictor of strategy frequency: many “cheap” strategies are rare in our data and some “expensive” strategies are quite common (e.g., always punish; Supplementary Figure 15). Third, this argument implies that the high frequency of the never punish strategy is merely an artefact of our design, since it is the only strategy that does not cost anything to implement. But as we have discussed, many other studies have found similar aversions to costly punishment in the lab[17,32,34,43,47] and in the real world[77–79] suggesting that this result is not an artefact of our design.

Another limitation of our study is the focus on samples from the United Kingdom and the United States. Given that previous work has found substantial cross-cultural variation in both the overall prevalence of costly punishment[13] and the prevalence of particular punishment strategies, such as antisocial punishment[19], there are reasons to expect that our results might not generalise to other cultures, such as non-Western and small-scale societies. Nonetheless, it is also notable

ARTICLE IN PRESS

ARTICLE IN PRESS

that we found striking similarities between our two samples, perhaps pointing to some cross-cultural commonalities, at least among these two WEIRD societies[83]. We also note that much previous work on punishment has focused on English-speaking Western cultures, making the strategies we identify particularly relevant to interpreting prior research.

A final potential limitation of our study is that 38% of our participants believed that their decisions in the economic games did not have real monetary consequences for others. We do not think that this is necessarily an issue for interpreting our results, since people do not tend to behave differently in incentivised vs. hypothetical economic games measuring social preferences and punitive behaviour[84]. Nonetheless, future research employing behavioural studies with online samples should monitor participants' belief in experimental paradigms, especially if incentivisation is an important feature of the design.

ARTICLE IN PRESS

Conclusion

In sum, we have shown that while many people choose not to punish peers, those who do are motivated by a variety of different concerns, including behaviour shaping, egalitarianism, and competition. Much like the observed variation in human social learning strategies[40], humans thus also exhibit variation in their punishment strategies. These individual differences map onto personality dimensions, social preferences, political and religious views, and self-reports of behaviour. We hope that future work will continue to unpack the multifaceted nature of human punishment.

ARTICLE IN PRESS

Acknowledgements

This work was supported by a Royal Society of New Zealand Catalyst Leaders Grant to Q.D.A and N.R. (ref: ILFUOA2002). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability Statement

All data used in this study are publicly available on GitHub: https://github.com/ScottClaessens/punishStrategies

Code Availability Statement

All code to reproduce the analyses in this study are publicly available on GitHub: https://github.com/ScottClaessens/punishStrategies

Author Contributions

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: *Scott Claessens***:** conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing – original draft, and writing – review & editing. *Quentin D Atkinson***:** conceptualization, funding acquisition, methodology, supervision, writing – original draft, and writing – review & editing. *Nichola J Raihani***:** conceptualization, funding acquisition, investigation, methodology, supervision, writing – original draft, and writing – review & editing.

ARTICLE IN PRESS

ARTICLE IN PRESS

References

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

2. dos Santos, M., Rankin, D. J. & Wedekind, C. The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences* **278**, 371–377 (2011).

3. dos Santos, M., Rankin, D. J. & Wedekind, C. Human cooperation based on punishment reputation. *Evolution* **67**, 2446–2450 (2013).

4. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends in Ecology & Evolution* **30**, 98–103 (2015).

5. Ostrom, E. *Governing the commons: The evolution of institutions for collective action*. (Cambridge University Press, 1990).

6. Raihani, N. J. & Bshary, R. Punishment: One tool, many uses. *Evolutionary Human Sciences* **1**, e12 (2019).

7. Balliet, D., Mulder, L. B. & Van Lange, P. A. M. Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin* **137**, 594–615 (2011).

8. Balliet, D. & Lange, P. A. M. V. Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science* **8**, 363–379 (2013).

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

10. Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. Winners don’t punish. *Nature* **452**, 348–351 (2008).

11. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *American Economic Review* **90**, 980–994 (2000).

ARTICLE IN PRESS

ARTICLE IN PRESS

12. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).

13. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767–1770 (2006).

14. Nikiforakis, N. & Normann, H.-T. A comparative statics analysis of punishment in public-good experiments. *Experimental Economics* **11**, 358–369 (2008).

15. Raihani, N. J., Thornton, A. & Bshary, R. Punishment and cooperation in nature. *Trends in Ecology & Evolution* **27**, 288–295 (2012).

16. Abbink, K. & Sadrieh, A. The pleasure of being nasty. *Economics Letters* **105**, 306–308 (2009).

17. Raihani, N. J. & McAuliffe, K. Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters* **8**, 802–804 (2012).

18. de Quervain, D. J.-F. *et al.* The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).

19. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).

20. Camerer, C. F. *Behavioral game theory: Experiments in strategic interaction*. (Russell Sage Foundation, 2003).

21. Bowles, S. & Gintis, H. *A cooperative species: Human reciprocity and its evolution*. (Princeton University Press, 2013).

22. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* **100**, 3531–3535 (2003).

23. Chudek, M. & Henrich, J. Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences* **15**, 218–226 (2011).

ARTICLE IN PRESS

ARTICLE IN PRESS

24. Henrich, J. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. (Princeton University Press, 2017).

25. Delton, A. W. & Krasnow, M. M. The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior* **38**, 734–743 (2017).

26. Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nature Human Behaviour* **2**, 458–468 (2018).

27. Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences* **108**, 11375–11380 (2011).

28. Mathew, S. & Boyd, R. The cost of cowardice: Punitive sentiments towards free riders in Turkana raids. *Evolution and Human Behavior* **35**, 58–64 (2014).

29. Richerson, P. *et al.* Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences* **39**, e30 (2016).

30. Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* **83**, 284–299 (2002).

31. Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794–796 (2007).

32. Bone, J. E. & Raihani, N. J. Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior* **36**, 323–330 (2015).

33. Walker, J. M. & Halloran, M. A. Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics* **7**, 235–247 (2004).

ARTICLE IN PRESS

ARTICLE IN PRESS

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

35. Crockett, M. J., Özdemir, Y. & Fehr, E. The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General* **143**, 2279–2286 (2014).

36. Sylwester, K., Herrmann, B. & Bryson, J. J. Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics* **6**, 167–188 (2013).

37. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evolution and Human Behavior* **25**, 63–87 (2004).

38. Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R. & Smirnov, O. The role of egalitarian motives in altruistic punishment. *Economics Letters* **102**, 192–194 (2009).

39. Fowler, J. H., Johnson, T. & Smirnov, O. Egalitarian motive and altruistic punishment. *Nature* **433**, E1 (2005).

40. Molleman, L., Van den Berg, P. & Weissing, F. J. Consistent individual differences in human social learning strategies. *Nature Communications* **5**, 3570 (2014).

41. Price, M. E., Cosmides, L. & Tooby, J. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* **23**, 203–231 (2002).

42. Barclay, P. Reputational benefits for altruistic punishment. *Evolution and Human Behavior* **27**, 325–344 (2006).

43. Batistoni, T., Barclay, P. & Raihani, N. J. Third-party punishers do not compete to be chosen as partners in an experimental game. *Proceedings of the Royal Society B* **289**, 20211773 (2022).

44. Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).

ARTICLE IN PRESS

ARTICLE IN PRESS

45. Jordan, J. J. & Rand, D. G. Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology* **421**, 189–202 (2017).

46. Jordan, J. J. & Rand, D. G. Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology* **118**, 57–88 (2020).

47. Deutchman, P., Bračič, M., Raihani, N. J. & McAuliffe, K. Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior* **42**, 12–20 (2021).

48. Marczyk, J. Human punishment is not primarily motivated by inequality. *PLoS ONE* **12**, e0171298 (2017).

49. Karagonlar, G. & Kuhlman, D. M. The role of social value orientation in response to an unfair offer in the ultimatum game. *Organizational Behavior and Human Decision Processes* **120**, 228–239 (2013).

50. Raihani, N., Martinez-Gatell, D., Bell, V. & Foulkes, L. Social reward, punishment, and prosociality in paranoia. *Journal of Abnormal Psychology* **130**, 177–185 (2021).

51. Roberts, S. C., Vakirtzis, A., Kristjánsdóttir, L. & Havlíček, J. Who punishes? Personality traits predict individual variation in punitive sentiment. *Evolutionary Psychology* **11**, 186–200 (2013).

52. Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B. & Skitka, L. J. Moral punishment in everyday life. *Personality and Social Psychology Bulletin* **44**, 1697–1711 (2018).

ARTICLE IN PRESS

ARTICLE IN PRESS

53. Osborne, D., Costello, T. H., Duckitt, J. & Sibley, C. G. The psychological causes and societal consequences of authoritarianism. *Nature Reviews Psychology* **2**, 220–232 (2023).

54. Galen, L. W., Kurby, C. A. & Fles, E. H. Religiosity, shared identity, trust, and punishment of norm violations: No evidence of generalized prosociality. *Psychology of Religion and Spirituality* **14**, 260 (2022).

55. Laurin, K., Shariff, A. F., Henrich, J. & Kay, A. C. Outsourcing punishment to God: Beliefs in divine control reduce earthly punishment. *Proceedings of the Royal Society B: Biological Sciences* **279**, 3272–3281 (2012).

56. McKay, R., Efferson, C., Whitehouse, H. & Fehr, E. Wrath of God: Religious primes and punishment. *Proceedings of the Royal Society B: Biological Sciences* **278**, 1858–1863 (2011).

57. Adler, N. E., Epel, E. S., Castellazzo, G. & Ickovics, J. R. Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology* **19**, 586–592 (2000).

58. Sibley, C. G. *et al.* The Mini-IPIP6: Validation and extension of a short measure of the Big-Six factors of personality in New Zealand. *New Zealand Journal of Psychology (Online)* **40**, 142 (2011).

59. Murphy, R. O., Ackermann, K. A. & Handgraaf, M. J. J. Measuring social value orientation. *Judgment and Decision Making* **6**, 771–781 (2011).

60. Ho, A. K. *et al.* The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO7 scale. *Journal of Personality and Social Psychology* **109**, 1003 (2015).

ARTICLE IN PRESS

ARTICLE IN PRESS

61. Bizumic, B. & Duckitt, J. Investigating right wing authoritarianism with a Very Short Authoritarianism Scale. *Journal of Social and Political Psychology* **6**, 129–150 (2018).

62. Charness, G., Gneezy, U. & Halladay, B. Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization* **131**, 141–150 (2016).

63. McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan, 2nd edition*. (CRC Press, 2020).

64. Gelman, A. & Tuerlinckx, F. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390 (2000).

65. Stan Development Team. RStan: The R interface to Stan. (2020).

66. Claessens, S. ScottClaessens/punishStrategies. (2025) doi:10.5281/zenodo.17535489.

67. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2022).

68. Wickham, H. *ggplot2: Elegant graphics for data analysis*. (Springer-Verlag New York, 2016).

69. Wilke, C. O. *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. (2020).

70. Landau, W. M. The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software* **6**, 2959 (2021).

71. Allaire, J. J. *et al.* Quarto. (2024) doi:10.5281/zenodo.5960048.

72. Peysakhovich, A., Nowak, M. A. & Rand, D. G. Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications* **5**, 4939 (2014).

ARTICLE IN PRESS

ARTICLE IN PRESS

73. Bone, J. E., McAuliffe, K. & Raihani, N. J. Exploring the motivations for punishment: Framing and country-level effects. *PLoS ONE* **11**, e0159769 (2016).

74. Cushman, F., Dreber, A., Wang, Y. & Costa, J. Accidental outcomes guide punishment in a ‘trembling hand’ game. *PloS ONE* **4**, e6699 (2009).

75. Darley, J. M., Carlsmith, K. M. & Robinson, P. H. Incapacitation and just deserts as motives for punishment. *Law and Human Behavior* **24**, 659–683 (2000).

76. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics* **92**, 91–112 (2008).

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

78. Balafoutas, L., Nikiforakis, N. & Rockenbach, B. Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications* **7**, 13327 (2016).

79. Baumard, N. Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society* **9**, 171–192 (2010).

80. Fitouchi, L. & Singh, M. Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior* **44**, 502–514 (2023).

81. Singh, M. & Garfield, Z. H. Evidence for third-party mediation but not punishment in Mentawai justice. *Nature Human Behaviour* **6**, 930–940 (2022).

82. Molleman, L., Kölle, F., Starmer, C. & Gächter, S. People prefer coordinated punishment in cooperative interactions. *Nature Human Behaviour* **3**, 1145–1153 (2019).

ARTICLE IN PRESS

ARTICLE IN PRESS

83. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behavioral and Brain Sciences* **33**, 61–83 (2010).

84. Amir, O., Rand, D. G. & Gal, Y. K. Economic games on the internet: The effect of $1 stakes. *PloS ONE* **7**, e31461 (2012).

ARTICLE IN PRESS

ARTICLE IN PRESS

Figure Legends

Figure 1

Visual summary of the six economic games. In the games, Player 2 either steals £0.20 from Player 1 (the focal player) or does nothing. Player 1 is then given the option to punish by paying a certain amount of money to remove money from Player 2 (this money is destroyed). The six games are variants on this general setup, creating situations where (A) Player 2 is still worse off after stealing, (B) Player 2 creates equality by stealing, (C) the computer 'decides' whether Player 2 steals, (D) the fee-fine ratio is 1:1, (E) Player 2 is better off after stealing, and (F) Player 2 steals instead from a third-party.

Figure 2

Overall pattern of punitive behaviour across all six economic games, split by country. N = 2010 participants. AI = advantageous inequity, DI = disadvantageous inequity.

Figure 3

Posterior estimates of the probabilities of following different punishment strategies from the Bayesian latent state model. The model assumes an implementation error rate of 5%. Points represent posterior medians, line ranges represent 50% and 95% credible intervals. N = 2010 participants. AI = advantageous inequity, DI = disadvantageous inequity.

Figure 4

Posterior slopes from Bayesian latent state models including Big-6 personality dimensions and Social Value Orientation. Each row represents a separate model. All models assume an implementation error rate of 5%. Points represent posterior medians, line ranges represent 95% credible intervals. N = 2010 participants. AI = advantageous inequity, DI = disadvantageous inequity.

ARTICLE IN PRESS

ARTICLE IN PRESS

Figure 5

Posterior slopes from Bayesian latent state models including political ideology, views about social inequality, and religiosity. Each row represents a separate model aside from Social Dominance Orientation and Right Wing Authoritarianism, which control for one another within the same model. All models assume an implementation error rate of 5%. Points represent posterior medians, line ranges represent 95% credible intervals. N = 2010 participants. AI = advantageous inequity, DI = disadvantageous inequity.

Figure 6

Boxplots showing the distribution of responses to each self-report question about the reasons for participants' behaviour in the games. N = 2010 participants. Boxplots represent medians and interquartile ranges.

Figure 7

Posterior slopes from models including self-reported strategy usage. Each row represents a separate model. All models assume an implementation error rate of 5%. Bolded estimates represent combinations where the self-report question matched the behavioural strategy. Points represent posterior medians, line ranges represent 95% credible intervals. N = 2010 participants. AI = advantageous inequity, DI = disadvantageous inequity.

ARTICLE IN PRESS

ARTICLE IN PRESS

Tables

Table 1

Summary of the different functions for punishment and the behavioural strategies they predict.

		Game A (AI) 70-10		Game B (Equal) 70-30		Game C (Computer) 70-30		Game D (1:1 Fee Fine) 70-30		Game E (DI) 70-50		Game F (Third Party) 70-70 [100]	
Strategy	Behavioural description	Steal 50-30	No steal 70-10	Steal 50-50	No steal 70-30	Steal 50-50	No steal 70-30	Steal 50-50	No steal 70-30	Steal 50-70	No steal 70-50	Steal 50-90 [100]	No steal 70-70 [100]
Deterrent	Punish to deter another who has harmed you from harming you again in the future	✓	x	✓	x	x	x	✓	x	✓	x	x	x
Norm-enforcing	Punish to enforce a shared anti-harm norm and encourage future norm compliance, even amongst third parties	✓	x	✓	x	x	x	✓	x	✓	x	✓	x
Revenge	Punish if doing so harms another who has harmed you	✓	x	✓	x	✓	x	✓	x	✓	x	x	x
Avoid DI	Punish if doing so avoids disadvantageous inequity for self	x	x	x	x	x	x	x	x	✓	x	x	x
Egalitarian	Punish if doing so makes payoffs for all more equal	x	x	x	x	x	x	x	x	✓	x	✓	x
Seek AI	Punish if doing so produces advantageous inequity for self	x	x	✓	x	✓	x	x	x	x	x	x	x
Competitive	Punish if doing so improves your relative position	✓	✓	✓	✓	✓	✓	x	x	✓	✓	✓	✓
Antisocial	Punish exclusively those who do not cause harm	x	✓	x	✓	x	✓	x	✓	x	✓	x	✓
Never punish	Never punish others	x	x	x	x	x	x	x	x	x	x	x	x

ARTICLE IN PRESS

Note. Games A-F are the games employed in the current study (see Methods for more details). In each of the six games, participants are given the opportunity to punish players who 'steal' and those who do not, meaning that participants make twelve punishment decisions in total. Each behavioural strategy implies a unique pattern of punishment across all decisions (see Methods for detailed explanation of strategies). Ticks reflect decisions to punish, crosses reflect decisions to not punish. In column headers, payoffs at the first stage (above) and the second stage (below) are denoted as P1-P2 (or P2-P3 [P1] for Game F) where participants take the role of P1 and P2 is the target of punishment. AI = advantageous inequity, DI = disadvantageous inequity.

ARTICLE IN PRESS

Table 2

Counts and proportions of participants following each punishment strategy exactly, split by country.

	United Kingdom (N = 1014)		United States (N = 996)	
Strategy	N	Prop	N	Prop
Deterrent	9	0.009	6	0.006
Norm-enforcing	8	0.008	16	0.016
Revenge	6	0.006	5	0.005
Avoid DI	67	0.066	62	0.062
Egalitarian	65	0.064	71	0.071
Seek AI	2	0.002	0	0.000
Competitive	3	0.003	1	0.001
Antisocial	0	0.000	0	0.000
Never punish	426	0.420	447	0.449
N/A	428	0.422	388	0.390

Note. N/A implies that participants were unable to be classified exactly into any of the punishment strategies. AI = advantageous inequity, DI = disadvantageous inequity.

ARTICLE IN PRESS

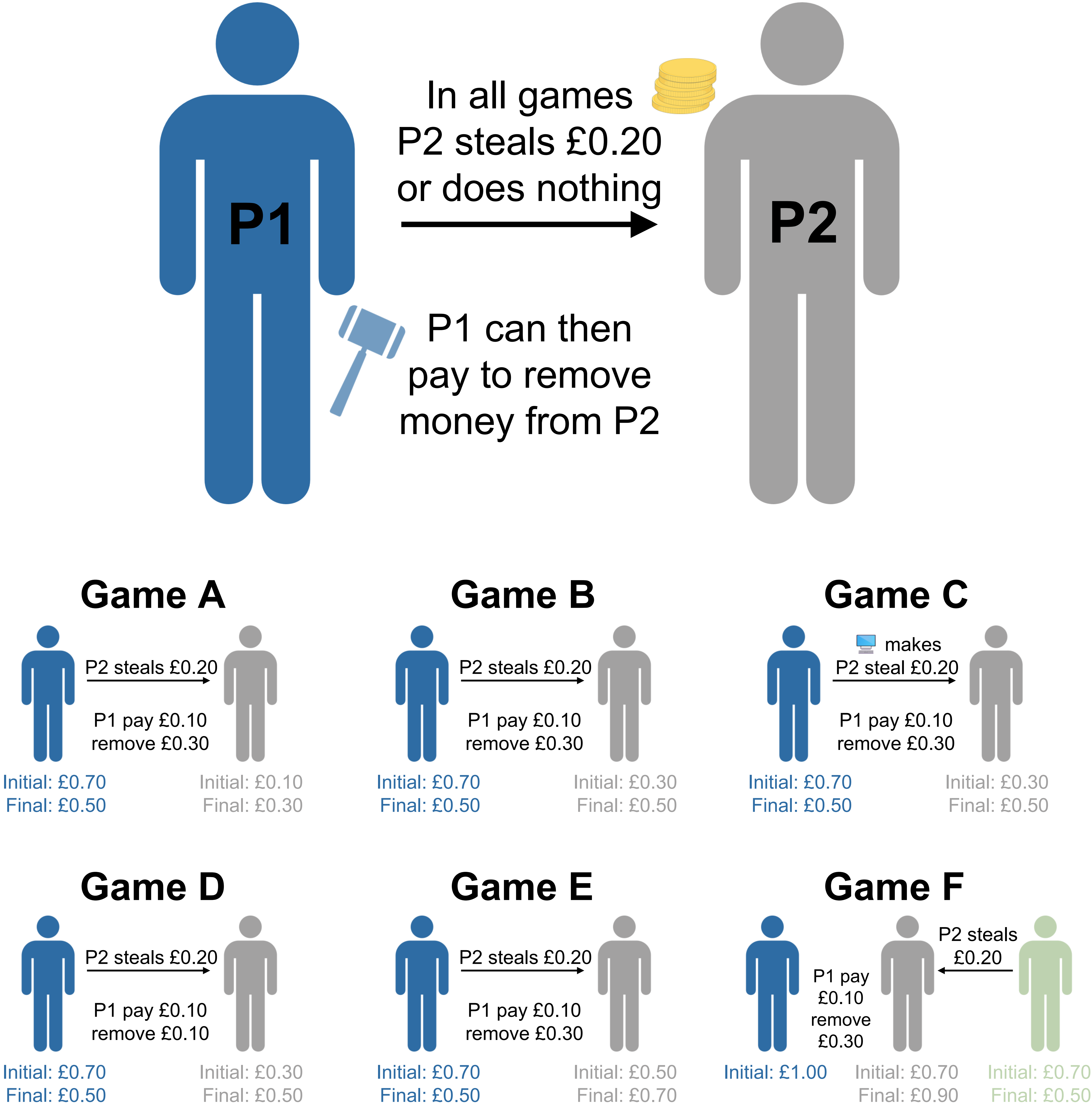
Editor Summary:

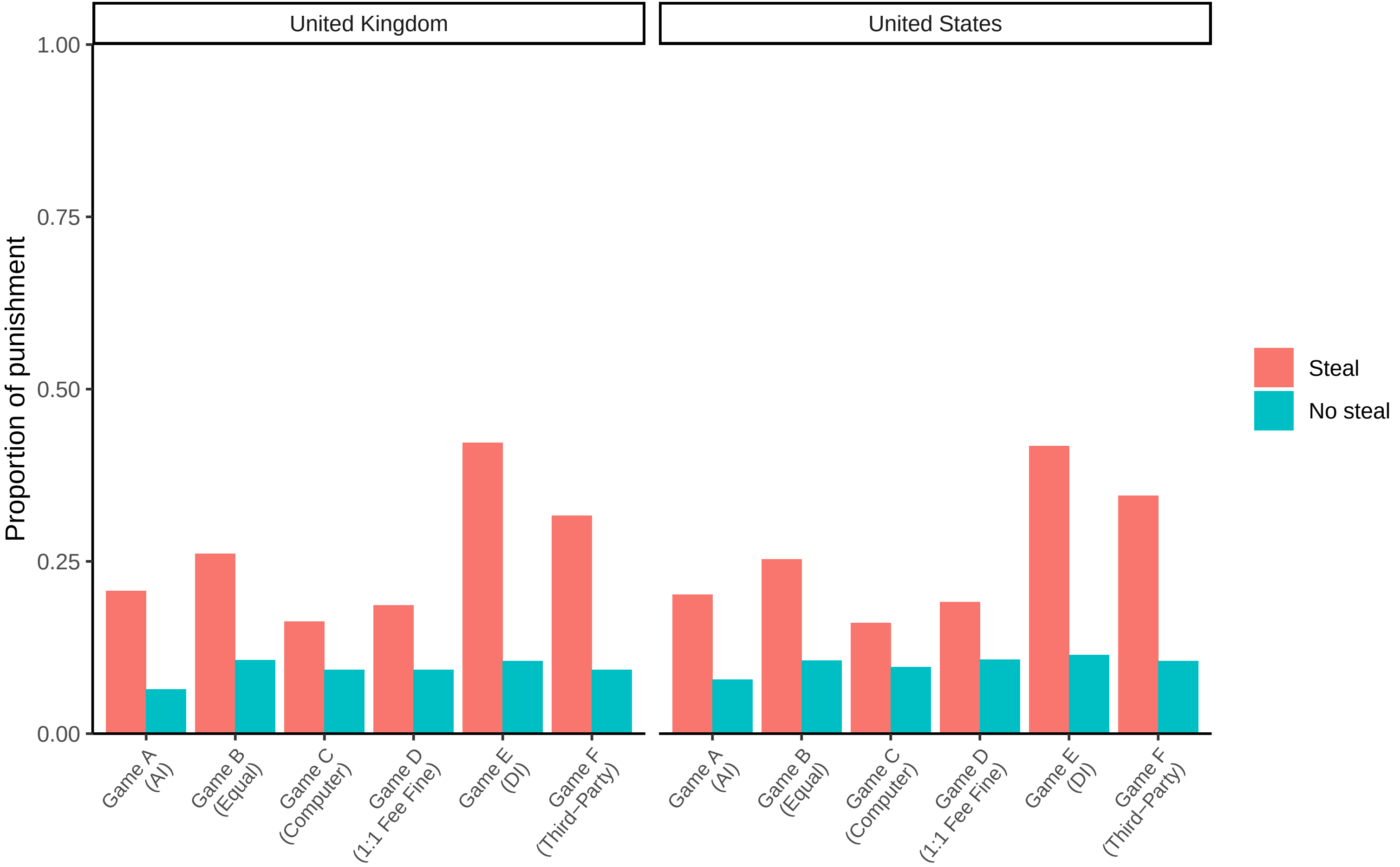
Using behavioural economic games, this study finds that people vary in their motives for punishment. Punishment motives are predicted by personality, social preferences, political and religious views, and self-reported strategies.

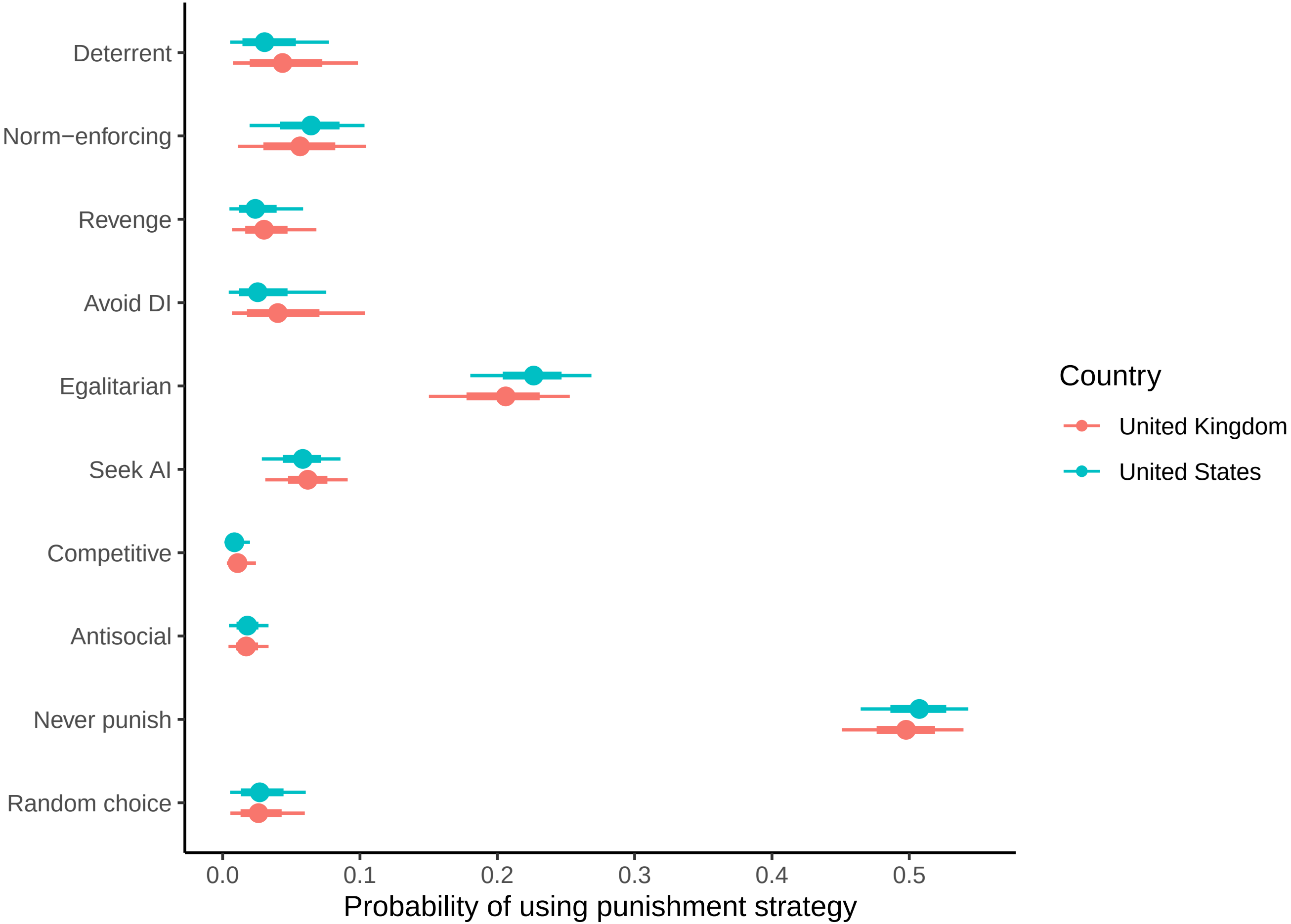
Peer Review Information:

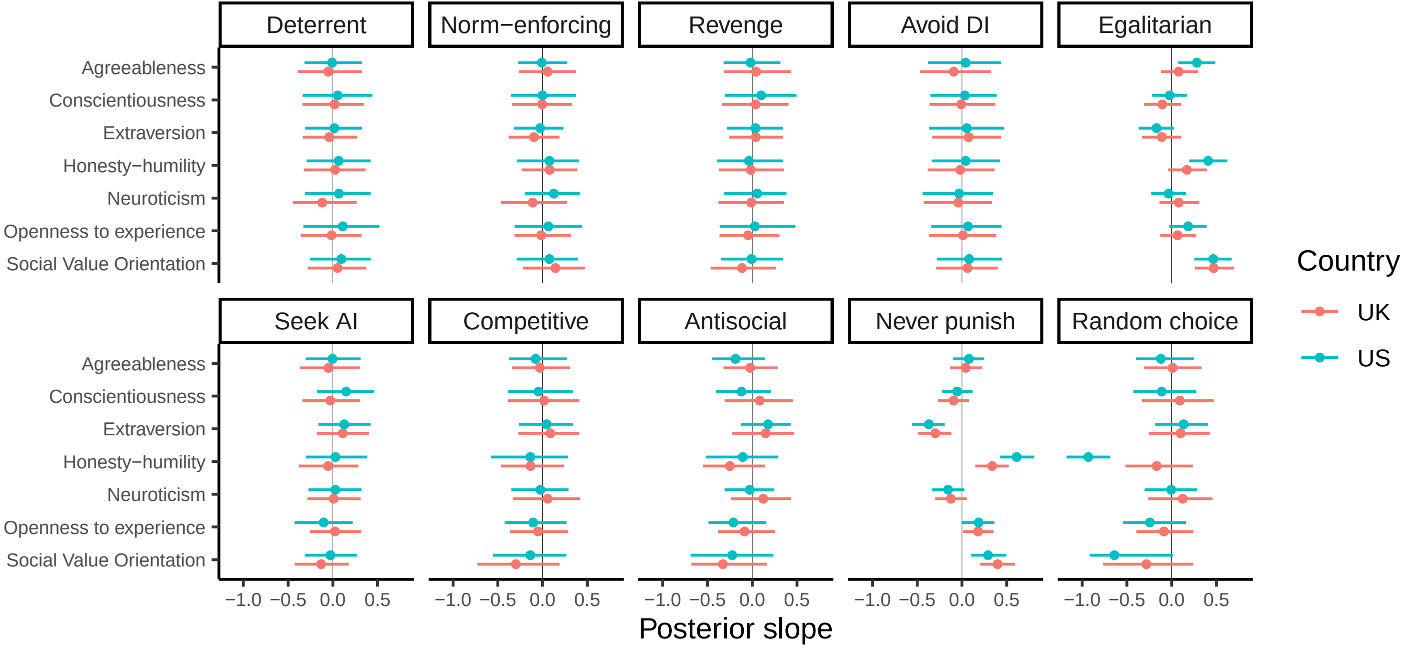
Communications Psychology thanks Alain Schläpfer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available. Primary Handling Editors: Yafeng Pan and Jennifer Bellingtier

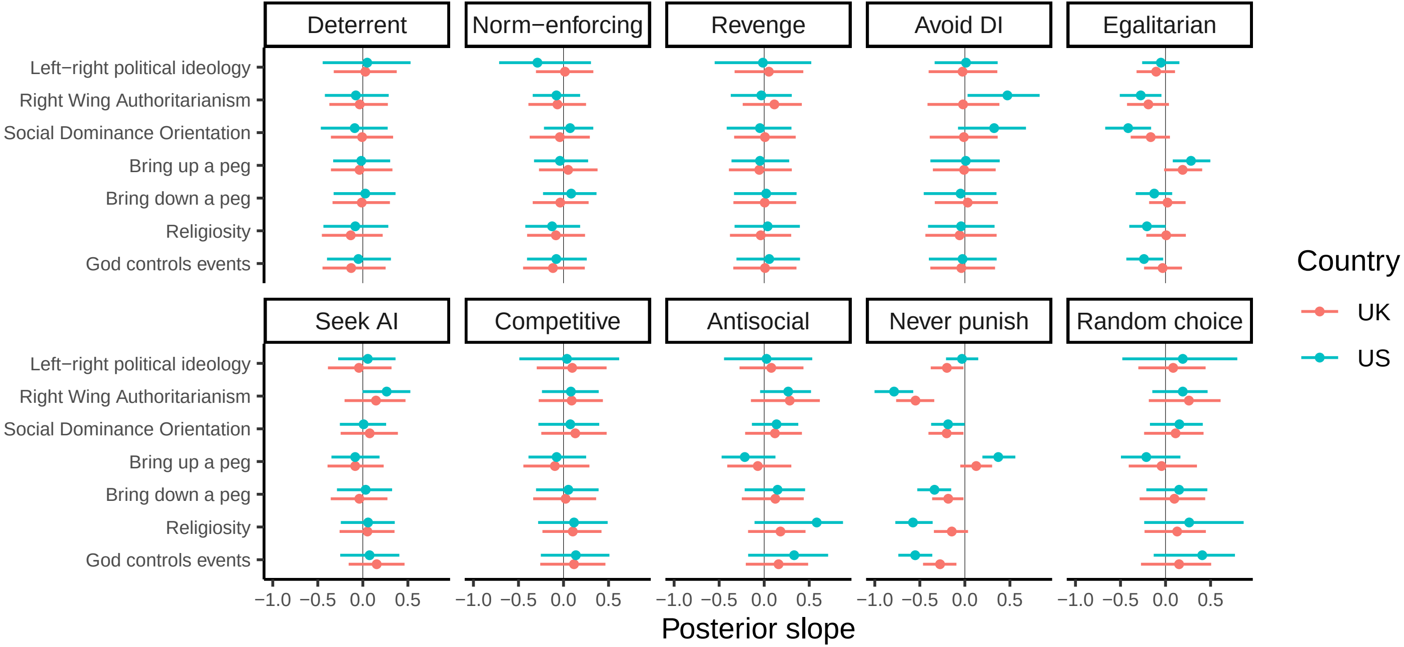
ARTICLE IN PRESS

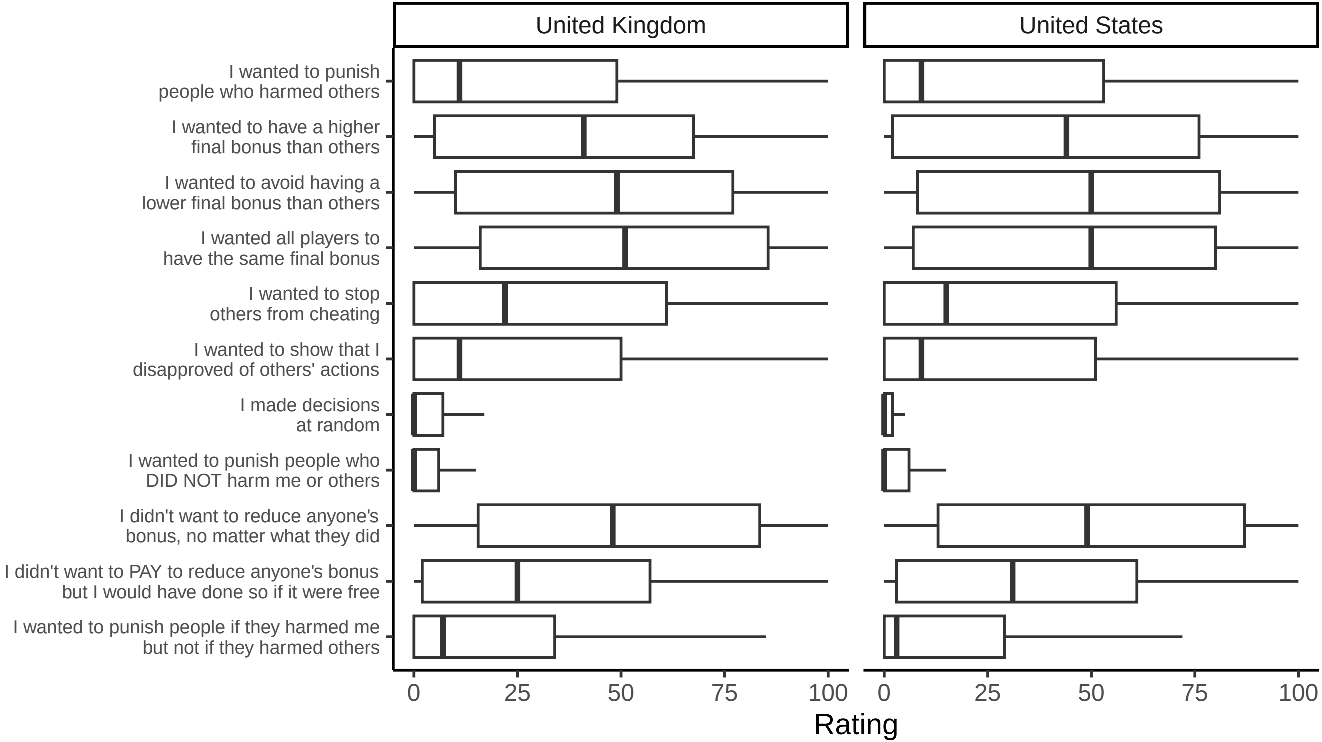

P1
In all games
P2 steals £0.20
or does nothing
P2
P1 can then
pay to remove
money from P2
Game A
P2 steals £0.20
P1 pay £0.10
remove £0.30
Initial: £0.70
Final: £0.50
Initial: £0.10
Final: £0.30
Game B
P2 steals £0.20
P1 pay £0.10
remove £0.30
Initial: £0.70
Final: £0.50
Initial: £0.30
Final: £0.50
Game C
makes
P2 steal £0.20
P1 pay £0.10
remove £0.30
Initial: £0.70
Final: £0.50
Initial: £0.30
Final: £0.50
Game D
P2 steals £0.20
P1 pay £0.10
remove £0.10
Initial: £0.70
Final: £0.50
Initial: £0.30
Final: £0.50
Game E
P2 steals £0.20
P1 pay £0.10
remove £0.30
Initial: £0.70
Final: £0.50
Initial: £0.50
Final: £0.70
Game F
P2 steals
£0.20
P1 pay
£0.10
remove
£0.30
Initial: £1.00
Initial: £0.70
Final: £0.90
Initial: £0.70
Final: £0.50

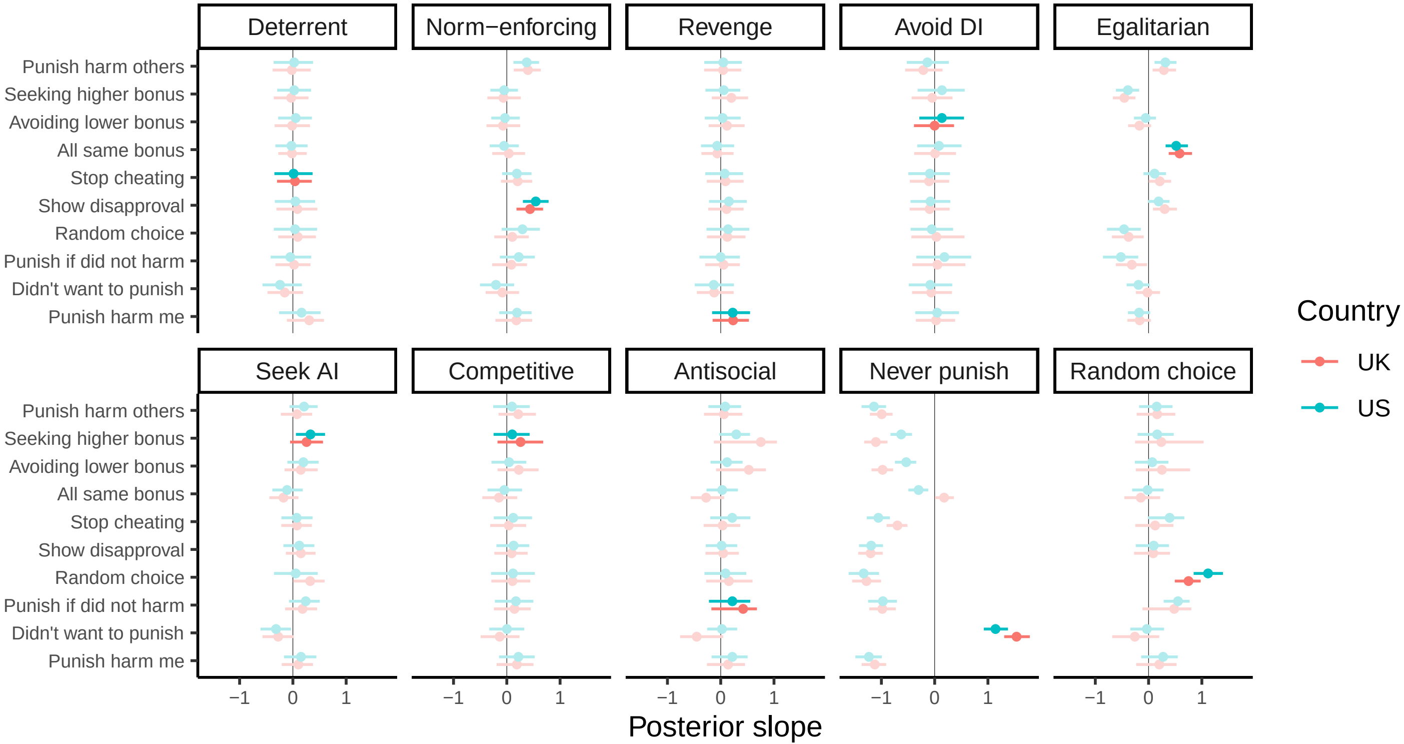

United Kingdom
United States
Proportion of punishment
1.00
0.75
0.50
0.25
0.00
Game A (AI)
Game B (Equal)
Game C (Computer)
Game D (1:1 Fee Fine)
Game E (DI)
Game F (Third–Party)
Steal
No steal


Deterrent
Norm−enforcing
Revenge
Avoid DI
Egalitarian
Seek AI
Competitive
Antisocial
Never punish
Random choice
0.0
0.1
0.2
0.3
0.4
0.5
Probability of using punishment strategy
Country
United Kingdom
United States


Deterrent
Norm−enforcing
Revenge
Avoid DI
Egalitarian
Seek AI
Competitive
Antisocial
Never punish
Random choice
Agreeableness
Conscientiousness
Extraversion
Honesty−humility
Neuroticism
Openness to experience
Social Value Orientation
−1.0
−0.5
0.0
0.5
Posterior slope
Country
UK
US


Deterrent
Norm−enforcing
Revenge
Avoid DI
Egalitarian
Seek AI
Competitive
Antisocial
Never punish
Random choice
Left−right political ideology
Right Wing Authoritarianism
Social Dominance Orientation
Bring up a peg
Bring down a peg
Religiosity
God controls events
−1.0
−0.5
0.0
0.5
Posterior slope
Country
UK
US


United Kingdom
United States
I wanted to punish people who harmed others
I wanted to have a higher final bonus than others
I wanted to avoid having a lower final bonus than others
I wanted all players to have the same final bonus
I wanted to stop others from cheating
I wanted to show that I disapproved of others' actions
I made decisions at random
I wanted to punish people who DID NOT harm me or others
I didn't want to reduce anyone's bonus, no matter what they did
I didn't want to PAY to reduce anyone's bonus but I would have done so if it were free
I wanted to punish people if they harmed me but not if they harmed others
0
25
50
75
100
0
25
50
75
100
Rating


Deterrent
Norm−enforcing
Revenge
Avoid DI
Egalitarian
Seek AI
Competitive
Antisocial
Never punish
Random choice
Punish harm others
Seeking higher bonus
Avoiding lower bonus
All same bonus
Stop cheating
Show disapproval
Random choice
Punish if did not harm
Didn't want to punish
Punish harm me
−1
0
1
Posterior slope
Country
UK
US