Why do people punish? Evidence for a range of strategic concerns

Scott Claessens*[1], Quentin D. Atkinson[1], & Nichola Raihani[1,2]

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

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

* Correspondence concerning this article should be addressed to Scott Claessens, Level 2, Building 302, 23 Symonds Street, Auckland, New Zealand. E-mail: scott.claessens@gmail.com

This working paper has not yet been peer-reviewed.

Abstract

Costly punishment is thought to be one of the key mechanisms sustaining cooperation in humans. However, the motives for punitive behaviour remain unclear. Punishment is often assumed to be motivated by a desire to convert free riders into cooperators, but it is also consistent with a host of other functions, such as levelling payoffs or increasing one's relative position. We used a suite of six economic games to tease apart the different motives for punishment. Across representative samples from the United Kingdom and the United States, we estimated the frequency of different punishment strategies in the population, finding that egalitarian motives for punishment are more common than behaviour-change motives. Moreover, different punishment strategies were differentially predicted by personality, social preferences, political ideology, and religious views. Self-reports of behaviour in the games suggested that people have some degree of insight into their punishment strategy. These findings highlight the multipurpose nature of human punishment.

Keywords: punishment; cooperation; economic games

Word count: 5185 words

Why do people punish? Evidence for a range of strategic concerns

Introduction

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]) in the future. Punishment therefore 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].

In humans, many studies of punishment have been carried out in laboratory settings using economic games[6–15]. In these games, participants are usually given a sum of money that they can use to invest in collective action or to help others. Alternatively, participants can 'cheat' by keeping the money for themselves or by exploiting the contributions of others. 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[12], that they enjoy punishing[16], and that they frequently, though not always[17], punish cheating or exploitative co-players[10,11].

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[6,7,15]. The effect that the threat of punishment has on cooperation is also evident in the higher contributions typically observed in the Ultimatum Game (where punishment is possible) compared to the structurally-similar Dictator Game (where it is not)[18]. This typical cooperation-enhancing effect of punishment has also been observed across societies[7], leading some to suggest that costly punishment has played a key role in the cultural evolution of cooperation in humans[19–22].

Nevertheless, it remains unclear whether individuals playing economic games use punishment as a behaviour-change tool to enforce cooperation or as a means to achieve other ends. Some have argued that punishment is primarily used to shape future behaviour, either to deter personal harm[3,9,23] or to uphold normative standards of cooperative behaviour[20,21,24–27]. 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[15]. 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 a host of other reasons why people may want to punish in economic games. Punishers might be motivated by a desire for retribution rather than deterrence, punishing in proportion to the amount of harm that was personally caused[28]. Punishment might be driven by concerns about relative payoffs, such as disadvantageous inequity aversion (i.e., avoiding having less than others[15,29]) and/or general egalitarian preferences (i.e., wanting all participants to receive the same payoffs[30]). 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[15].

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 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[12,29,31,32], on the very last round of repeated games[33], or in games where the target never learns about the punishment[34]; punishing those who did not cheat or who over-contributed to collective action (antisocial punishment[17,35]); punishing in scenarios where they were not personally harmed (third-party punishment[36]); and punishing in scenarios where disparities in payoffs did not arise from participants' actions[30,37,38].

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[15]. 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[39]. Some individuals may use punishment as a behaviour shaping tool, for example, while others may use it to reduce or reverse payoff differentials.

This insight raises several underexplored questions. First, which punishment strategies are more frequent in human populations? Second, what traits predict adherence to a particular strategy? Previous work has reported that personality is related to cooperative behaviour[40] and demographics, political ideology, and religiosity are related to punitive behaviour[41], but no research has related these variables to specific punishment strategies. Third, do people have insight into their own punishment strategy? Previous work has argued that people are often unaware of the underlying function of their punitive behaviour, yet they feel compelled to enact it anyway[28,42].

Here, we aim to delineate nine possible punishment strategies by asking whether people punish in a manner consistent with a specific strategy and, if so, what other characteristics (personality, social preferences, political orientation) predict the use of different punishment strategies. Table 1 summarises the potential functions for costly punishment in the economic games that we considered, and the behavioural strategies they predict. Note that Table 1 is not an exhaustive list of all possible punishment strategies: we do not include reputational functions of punishment in this table, such as signalling trustworthiness[4,43–46], because our focus is on punishment strategies in anonymous economic games without reputational incentives (but see ref[47]).

Building on previous designs[29,31,48,49], 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 carefully 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. We then combine these behavioural patterns with data on demographics, personality, social preferences, political ideology, religiosity, and self-reported strategy usage.

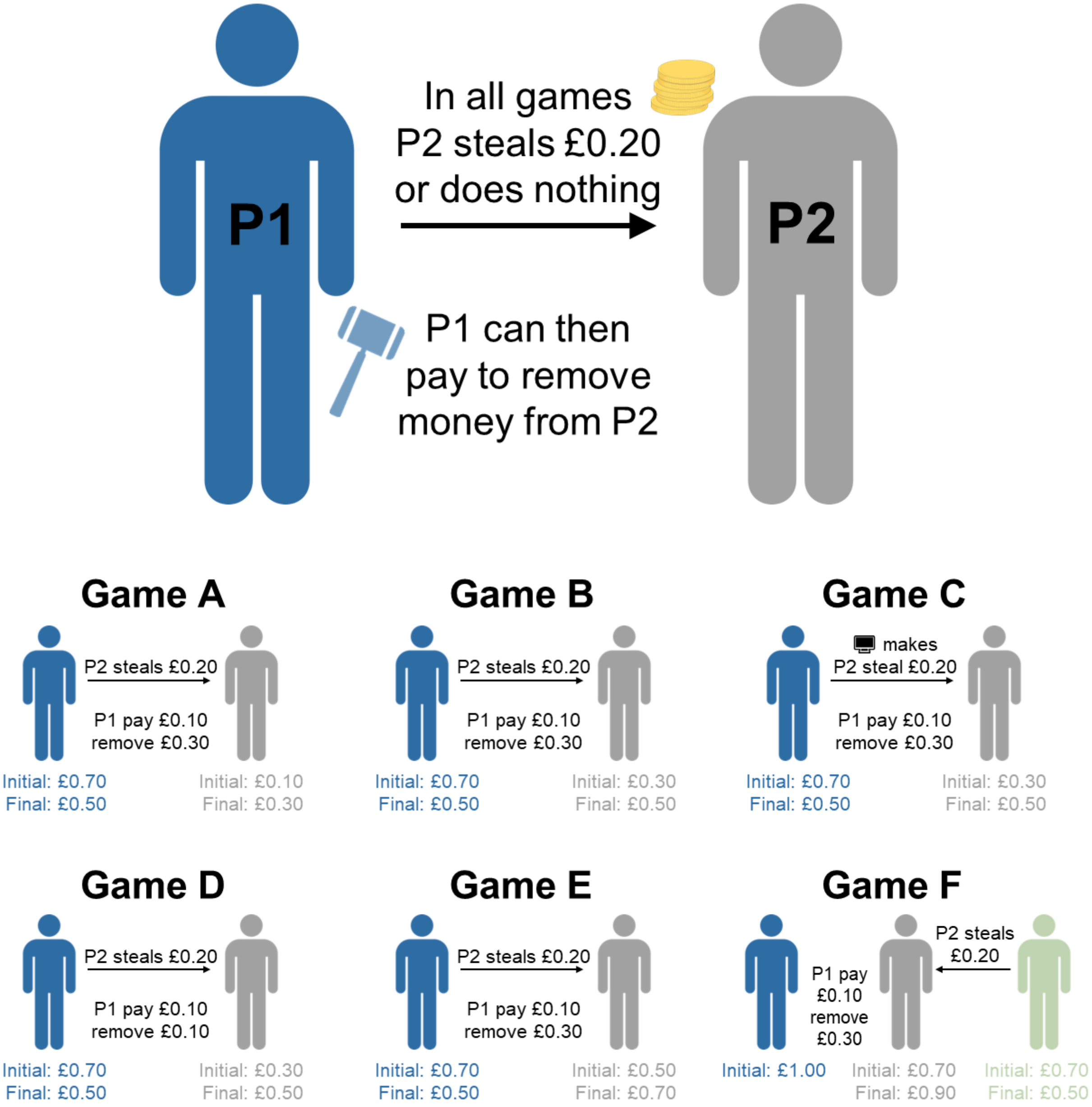


Figure 1. *Visual summary of the six economic games.* In all games, Player 2 either steals £0.20 from Player 1 (the focal player) or does nothing. Player 1 is then given the option to punish by paying a certain amount of money to remove money from Player 2 (this money is destroyed). The six games are variants on this general setup, creating situations where (A) Player 2 is still worse off by 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 by stealing, and (F) Player 2 steals instead from a third-party.

Table 1

Summary of the different functions for punishment and the behavioural strategies they predict. 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. Green ticks reflect decisions to punish, red 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.

		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]	
Function	Behavioural strategy	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
Retributive	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

Results

The overall pattern of punitive behaviour in the six economic games was in line with previous research and very similar across both countries (Figure 2). Participants were generally more likely to punish targets who stole from another individual compared to targets who did not steal (multilevel logistic regression; $b = 1.93$, standard error = 0.27, $p < .001$). Participants were also more likely to punish when targets' stealing behaviour generated inequalities, specifically in Games E and F ($b = 2.42$, SE = 0.44, $p < .001$).

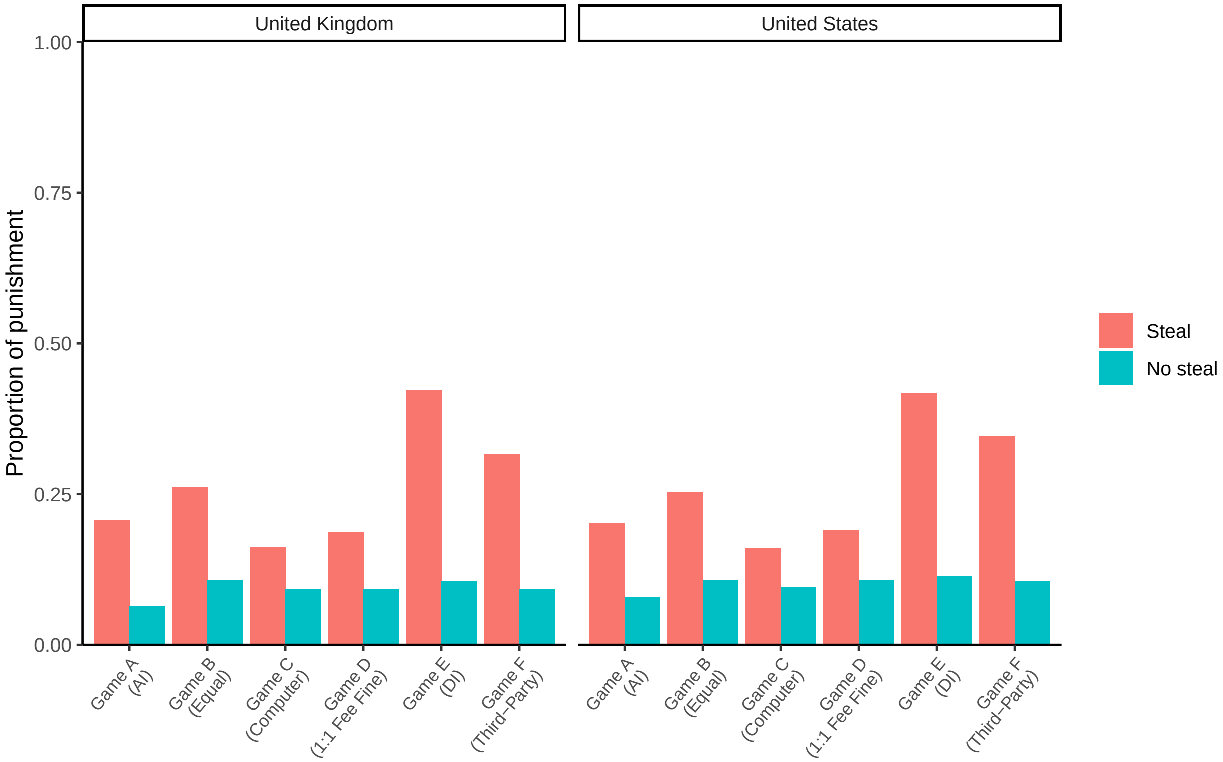


Figure 2. *Overall pattern of punitive behaviour across all six economic games, split by country.* AI = advantageous inequity, DI = disadvantageous inequity.

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

Table 2
Counts and proportions of participants following each punishment strategy exactly, split by country. N/A implies that participants were unable to be classified exactly into any of the punishment strategies.

	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
Retributive	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

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 retributive strategy, and the competitive strategies (seek AI, 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 S1 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). Punishment in this game is consistent with an egalitarian motive, as stealing produces unequal outcomes, but third-party punishment here is also consistent with norm-enforcing and competitive motives (see Table 1). Other common 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 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 AI 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 S1).

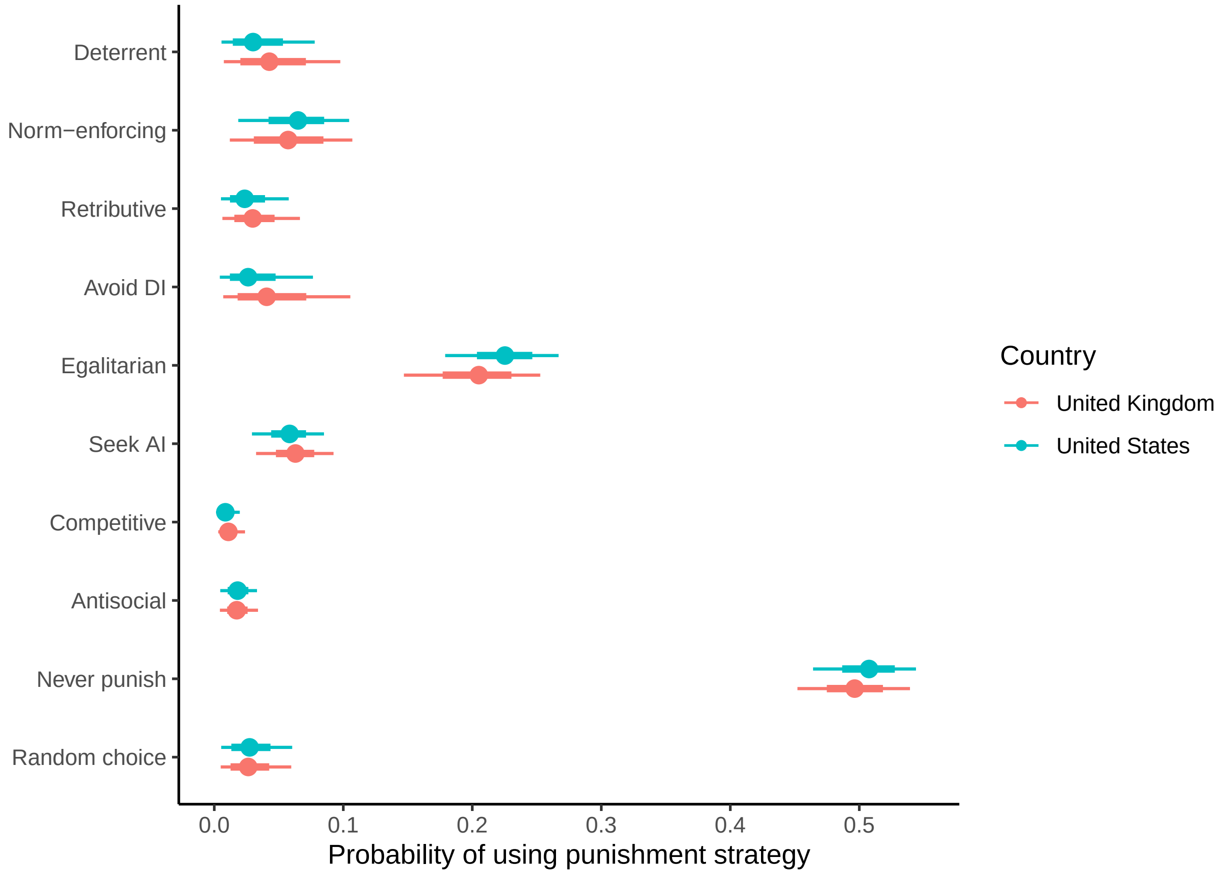


Figure 3. *Posterior estimates of the probabilities of following different punishment strategies from the Bayesian latent state model.* The model assumes an implementation error rate of 5%. Points represent posterior medians, line ranges represent 50% and 95% credible intervals.

Next, we explored which traits predicted adherence to different 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 S2 and S3). 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 -0.02]) but this effect was small.

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 (Figure 4). Participants higher in SVO 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 S4).

Political and religious variables were also associated with punishment strategy (Figure 5). These effects 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 participants were less likely to follow the never punish strategy. This general pattern of results was replicated with the full dataset (Supplementary Figure S5).

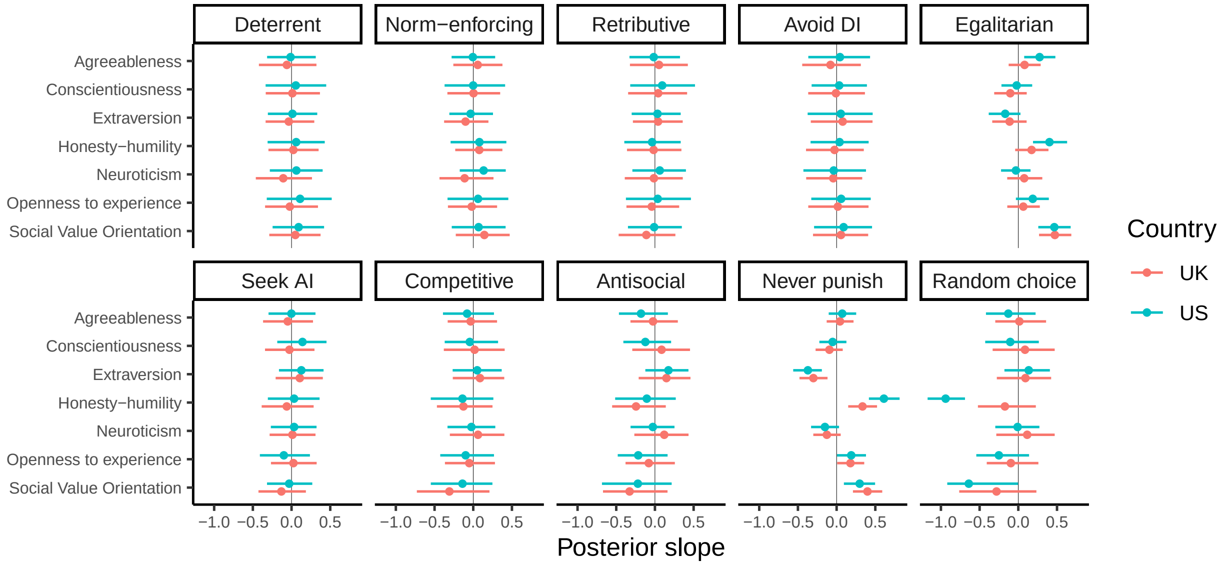


Figure 4. *Posterior slopes from Bayesian latent state models including Big-6 personality dimensions and Social Value Orientation.* Each row represents a separate model. All models assume an implementation error rate of 5%. Points represent posterior medians, line ranges represent 95% credible intervals.

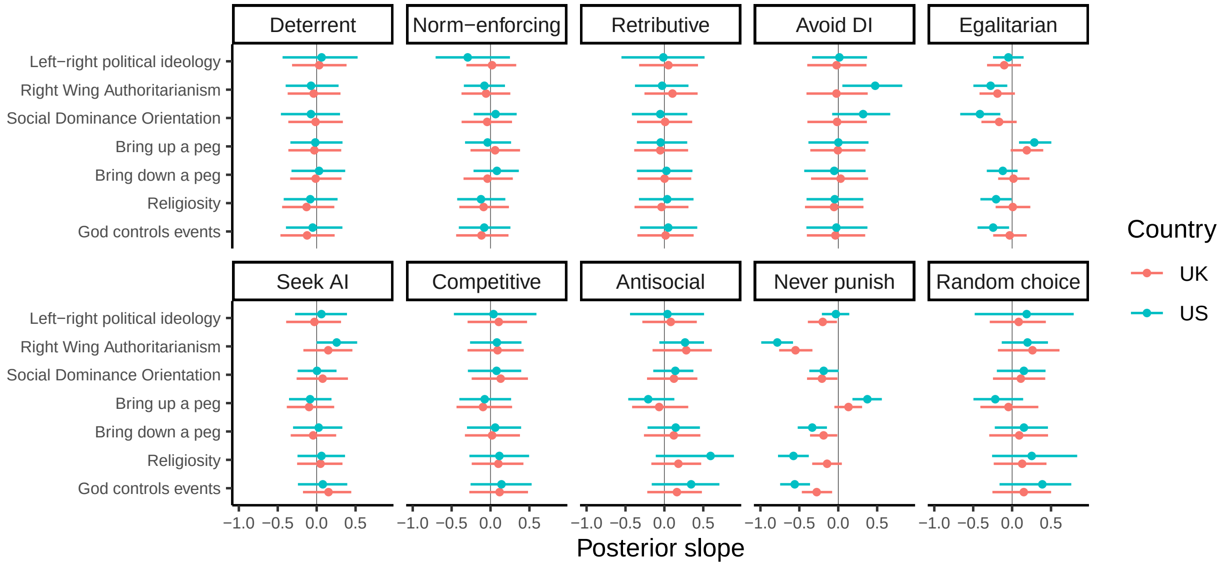


Figure 5. *Posterior slopes from Bayesian latent state models including political ideology, views about social inequality, and religiosity.* Each row represents a separate model aside from Social Dominance Orientation and Right Wing Authoritarianism, which control for one another within the same model.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

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 (see Supplementary Figures S6 and S7 for the distribution of responses to self-report questions). Figure 6 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 S8).

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[15]. Punishment is more akin to a swiss army knife than a hammer, used by some to enforce norms of cooperation and by others 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[28], 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

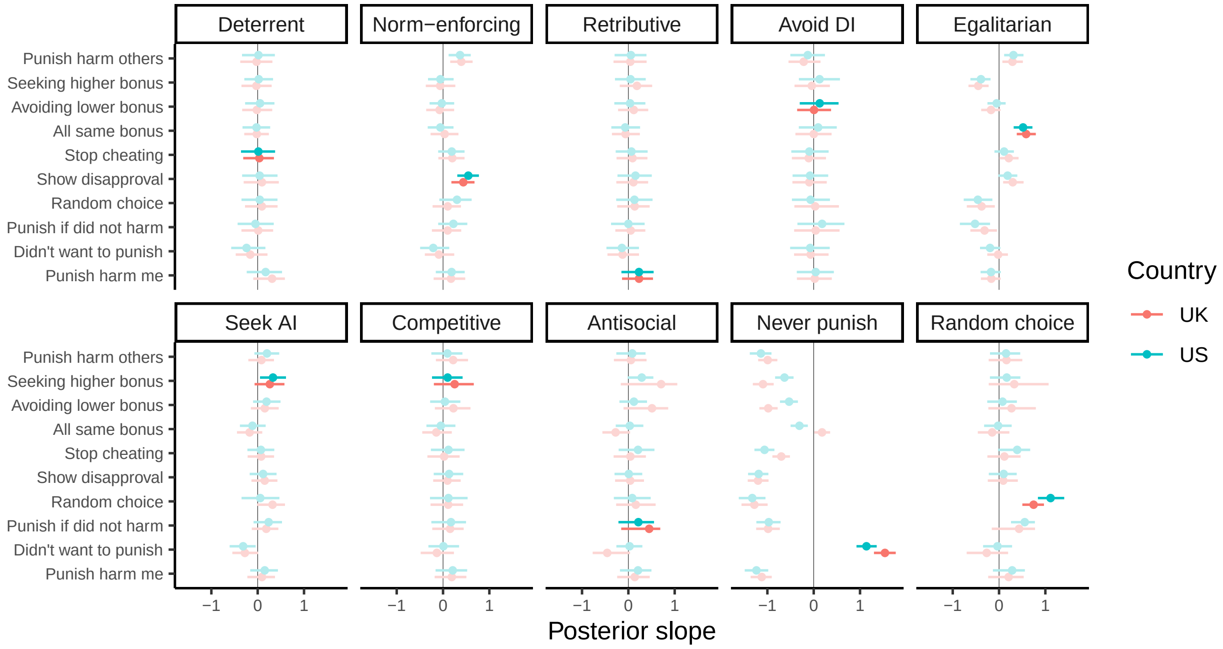


Figure 6. *Posterior slopes from models including self-reported strategy usage.* Each row represents a separate model. All models assume an implementation error rate of 5%. Highlighted estimates represent combinations where the self-report question matched the behavioural strategy. Points represent posterior medians, line ranges represent 95% credible intervals.

States.

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[29,30]. Personality and social preference variables mapped onto these strategies in expected ways. Traits associated with other-regarding concern, such as SVO and honesty-humility, predicted following the egalitarian strategy, whereas religious and conservative individuals were less likely to follow this strategy, especially in the United States. Moreover, participants following the egalitarian strategy were able to self-report this strategy, though the same was not true for the avoid disadvantageous inequity strategy.

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 (Supplementary Figures S6 and S7). Regarding the predictors of these strategies, we found that demographic, personality, political, and religious variables tended to be unrelated to deterrent and norm-enforcing punishment strategies. We also found that participants had insight into 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[28].

Other punitive strategies were less common in our dataset, but some were more prevalent than others. For example, participants were more likely to use punishment to seek advantageous inequity than to exclusively harm those who did not steal (i.e., antisocial punishment). The existence of the "seek AI" strategy in our dataset supports the claim that punishment can also be used as a tool to increase one's own relative position[15]. While generally rare, we found that this motive for punishment was more common among authoritarian participants, at least in the United States, potentially providing an explanation for why peer punishment has been found to be more common among conservatives in previous work[41]. Moreover, the fact that no participants in our sample punished non-stealing across all games suggests that antisocial punishment does not function to harm cooperators specifically, as has been previously suggested[17]. Instead, antisocial punishment appears to be motivated by improving one's relative position in general, which is in line with work showing that antisocial punishment disappears with a 1:1 fee-fine ratio[35].

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 will punish to change the behaviour of cheats. But in reality, people may be choosing the punishment option to achieve a variety of different goals. The targets of punishment in these studies are likely well aware that punishment could be levied for these different reasons and this knowledge may impact their responses. For example, if cheating targets interpret punishment as serving a competitive motive, it may elicit retaliation rather than encourage cooperation[9,15,50]. This might help to explain the mixed findings in the field as to whether punishment actually motivates cheating targets to cooperate in the future[15].

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[51,52] and in small-scale societies[53]. One reason that people may be averse to peer punishment is that, due to its multipurpose nature, it may be interpreted as a competitive challenge by targets and trigger retaliation[15]. In situations that lack clear institutional norms to legitimise punishment, such as our economic games and some situations in the real world, people might abstain from peer punishment to avoid such retaliation, regardless of whether retaliation is actually possible. By contrast, institutionalised punishment in small-scale societies often functions to compensate victims adequately while limiting the potential for feuds and cycles of retaliation[54,55]. Future research should uncover whether people are more willing to punish in these conventionalised contexts.

There are several limitations with our study design that can guide future research. First, we used one-shot economic games to measure punishment strategies, which may have led us to underestimate behaviour-change strategies like deterrence. Our inclusion of Game C somewhat mitigated this issue by manipulating whether stealing behaviour was intentional vs. unintentional and thus whether there was any behaviour to be deterred. Due to limits on our within-subjects design and the complexity of the strategy space, it was not feasible for us to expand our study to include additional contexts to elicit behaviour-change strategies (e.g., repeated games, games where targets are not made aware of the punishment, games where targets can retaliate). Future work could study these contexts separately.

A second limitation is that some strategies required more punishment than others to be met. 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). Strategies thus differed in how "expensive" they were to implement, perhaps explaining why some strategies were more common than others. This issue is largely unavoidable in our design since strategies, by their very nature, differ in how punitive they are. To partially mitigate this issue, we employed the strategy method to incentivise participants, such that payoffs were calculated from a randomly chosen game instead of summed across all games.

Another limitation is that our results may be contingent on the particular suite of anonymous stealing games that we used. With our anonymous design, we were unable to study other potential reputational strategies underlying punishment, such as signalling trustworthiness[47]. Moreover, stealing may be evaluated differently to other forms of cheating, such as not contributing to public goods, and other negative behaviours, such as lying or breaking taboos. Finally, we were unable to include all permutations of situational features in the games (e.g., second-party vs. third-party, equal vs. unequal) making it difficult to interpret some patterns of behaviour. For example, many participants punished only in the third-party game (Game F), but it is not clear whether these participants were driven by the third-party nature of the game or simply by the fact that stealing in that game generated inequality. Future work should determine whether different reputational contexts, target behaviours, and combinations of situational features elicit different punishment strategies.

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[39], 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.

Methods

Ethical approval

Ethical approval was granted by the University College London Ethics Board (project: 3720/002). The study was performed in accordance with all the relevant guidelines and regulations. Informed consent was obtained from all participants prior to the study.

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 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 (see Supplementary Figure S9 for sample characteristics).

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 (see Supplementary Figure S10 for sample characteristics).

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. 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.

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 S2).

Survey questions. In a follow-up survey, we collected the following data on participants (for wordings of all questions, see Supplementary Table S3):

- *Demographics.* In the survey, we collected information on participants' education level and self-reported socio-economic status (MacArthur ladder[56]). We also collected additional demographic data from Prolific (e.g., age, gender, student status).

- *Personality.* We used the Mini-IPIP scale[57] to measure the Big 6 personality dimensions of agreeableness, conscientiousness, extraversion, honesty-humility, openness to experience, and neuroticism (four items each).
- *Social Value Orientation.* We used the Social Value Orientation Slider Measure to measure other-regarding preferences[58]. 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[58].
- *Political ideology.* We included several measures of political ideology, including left-right conservatism, Social Dominance Orientation[59] (eight items), and Right Wing Authoritarianism[60] (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[61].

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 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 S4). We used the strategy method to incentivise the economic games, choosing a random game to determine bonus payment. 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 decision). 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 decision).

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[62]). In 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:

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

The conditional probabilities Pr(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 green ticks in Table 1 with a conditional probability of $1 - \delta$ and coding red crosses with a conditional probability of $0 + \delta$. The random choice 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:

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

We estimated the posterior distributions of these models using Hamiltonian Monte Carlo as implemented in Stan version 2.26.1[63]. 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 S11.

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 S12).

Reproducibility

All data and code are accessible on GitHub: https://github.com/ScottClaessens/punishStrategies. All analyses were conducted in R version 4.2.1[64]. Visualisations were created with the *ggplot2*[65] and *cowplot*[66] R packages. We used the *targets*[67] R package to create a reproducible data analysis pipeline and the *papaja*[68] R package to reproducibly generate the manuscript.

Acknowledgements

This work was supported by a Royal Society of New Zealand Catalyst Leaders Grant to Q.D.A and N.R. (ref: ILFUOA2002).

Author Contributions

All authors conceptualised the research, designed the study, and developed the surveys. N.J. conducted data collection on Prolific. S.C. conducted all analyses and visualisation of the data. All authors wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

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

Code Availability

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

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

7. 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).

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

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

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

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

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

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

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

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

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

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

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

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

20. 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).

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

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

23. 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).

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

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

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

27. 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).

28. 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).

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

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

31. 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).

32. 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).

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

34. 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).

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

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

37. 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).

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

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

40. Thielmann, I., Spadaro, G. & Balliet, D. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin* **146**, 30–90 (2020).

41. 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).

42. Nisbett, R. E. & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**, 231–259 (1977).

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

44. 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).

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

46. 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).

47. 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).

48. 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).

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

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

51. 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).

52. 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).

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

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

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

56. 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).

57. 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).

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

59. 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).

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

61. 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).

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

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

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

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

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

67. 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).

68. Aust, F. & Barth, M. *papaja: Prepare reproducible APA journal articles with R Markdown.* (2022).

Supplementary Material

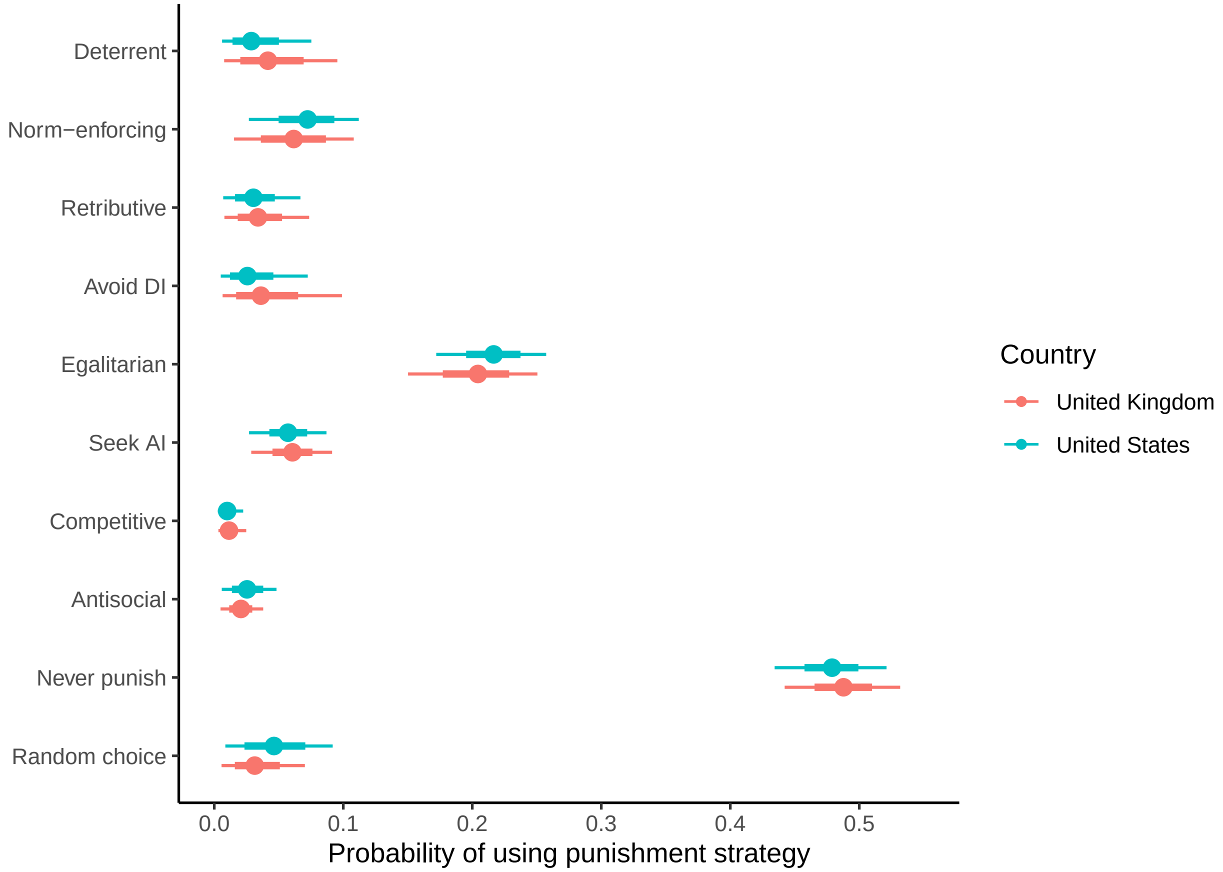
Why do people punish? Evidence for a range of strategic concerns

Scott Claessens[1], Quentin D. Atkinson[1], Nichola Raihani[1,2]

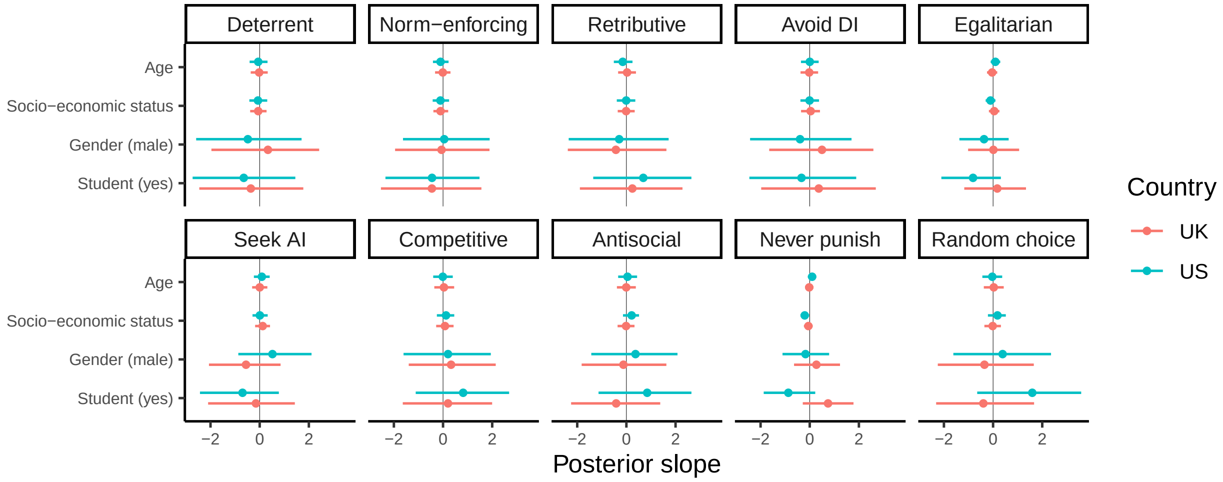
[1] School of Psychology, University of Auckland, Auckland, New Zealand

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

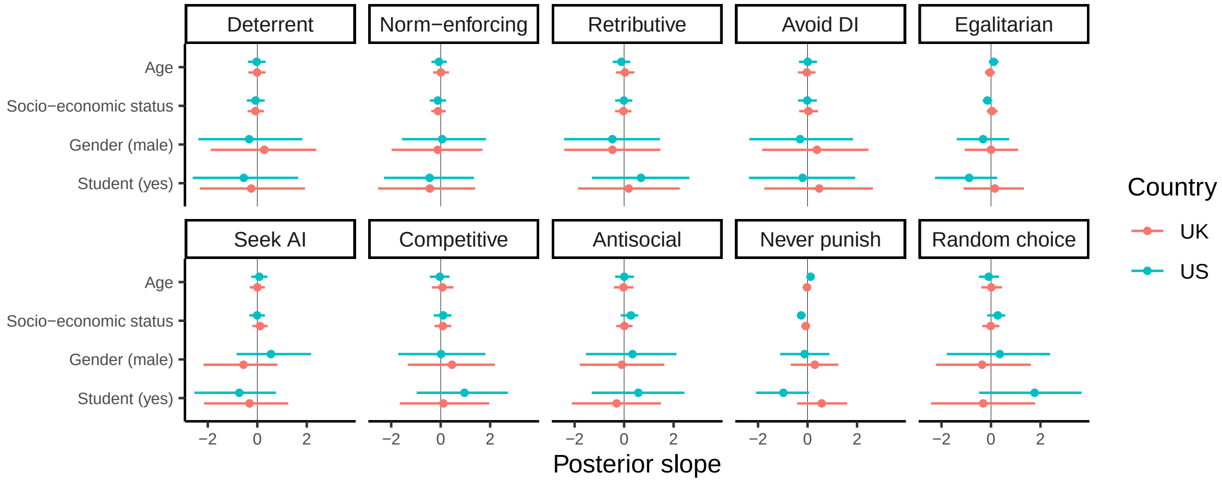
Supplementary Figures



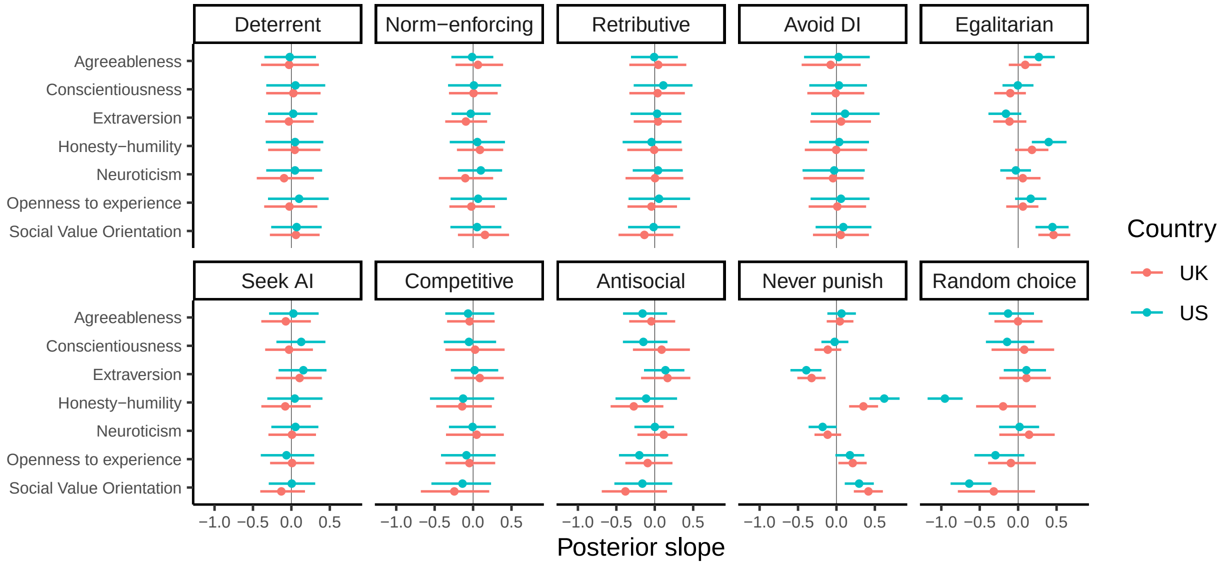
Supplementary Figure S1. Posterior estimates of the probabilities of following different punishment strategies from the Bayesian latent state model fitted to the full dataset without pre-registered exclusions. The model assumes an implementation error rate of 5%. Points represent posterior medians, line ranges represent 50% and 95% credible intervals.



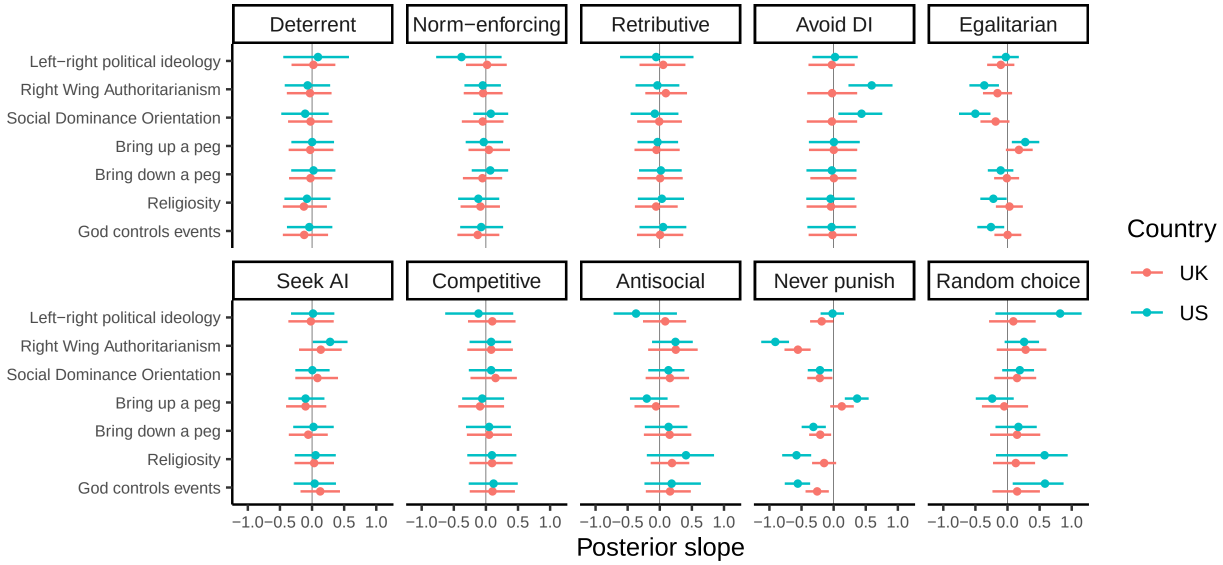
Supplementary Figure S2. *Posterior slopes from models including age, socio-economic status, gender, and student status, fitted to the subsetted dataset with pre-registered exclusions.* Each row represents a separate model. Points represent posterior medians, line ranges represent 95% credible intervals.



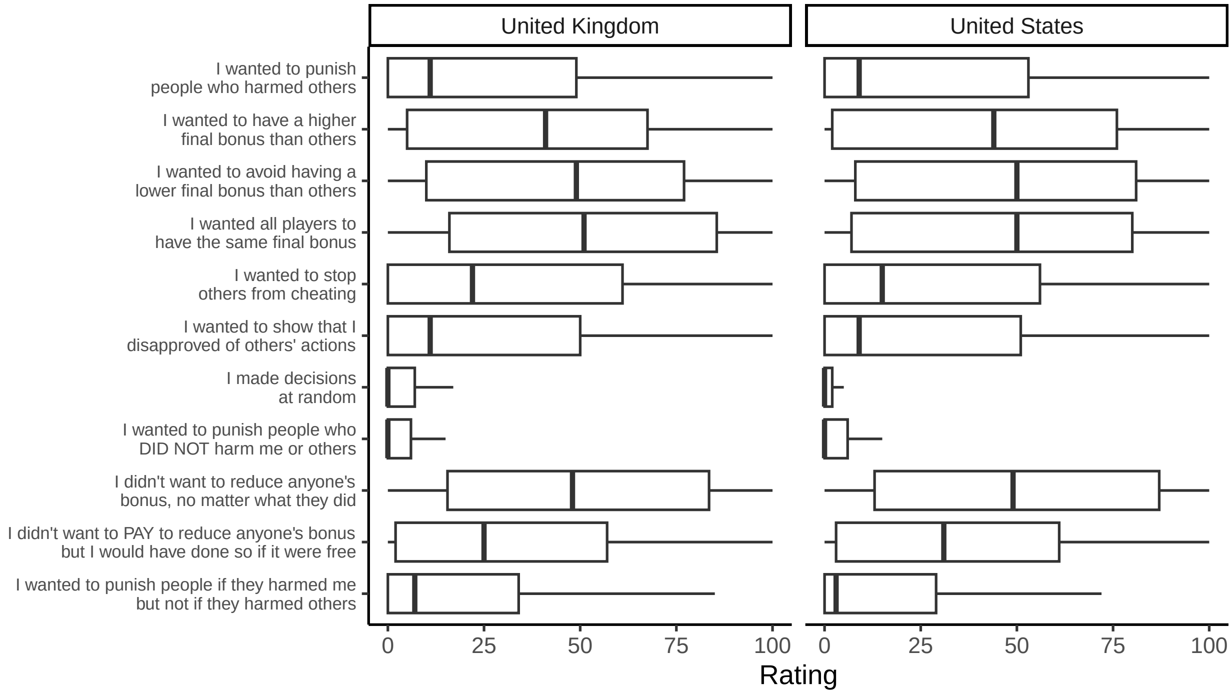
Supplementary Figure S3. Posterior slopes from models including age, socio-economic status, gender, and student status, fitted to the full dataset without pre-registered exclusions. Each row represents a separate model. Points represent posterior medians, line ranges represent 95% credible intervals.



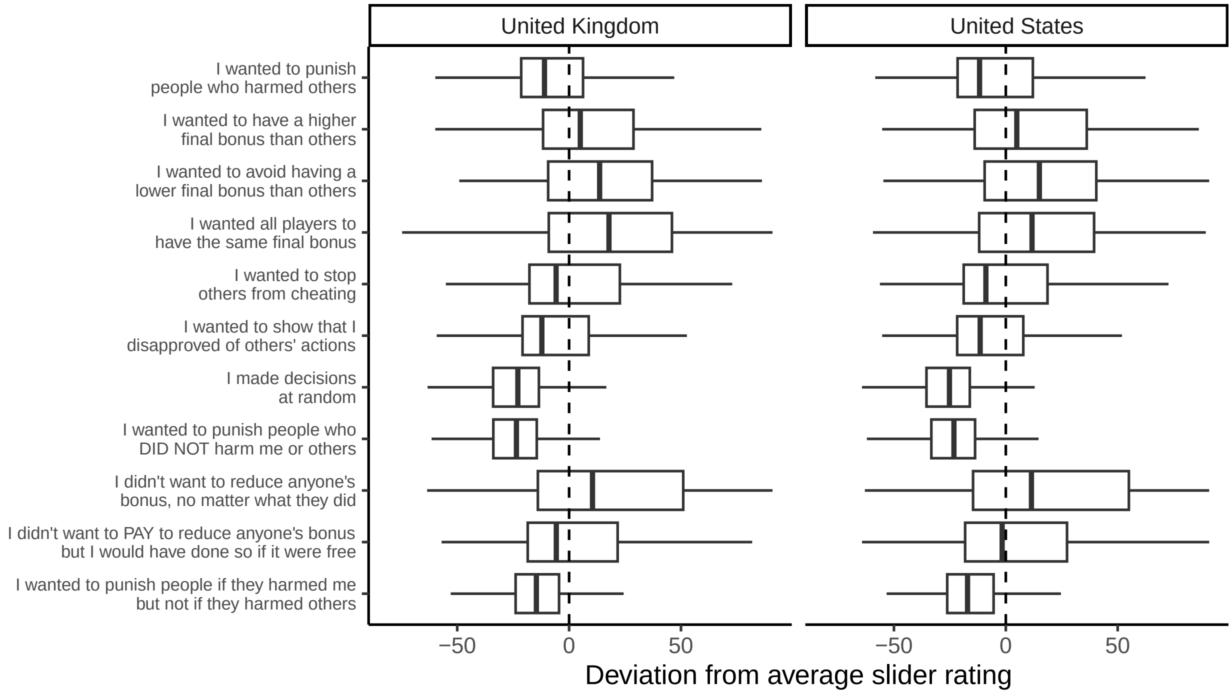
Supplementary Figure S4. *Posterior slopes from models including Big-6 personality dimensions and Social Value Orientation, fitted to the full dataset without pre-registered exclusions.* Each row represents a separate model. Points represent posterior medians, line ranges represent 95% credible intervals.



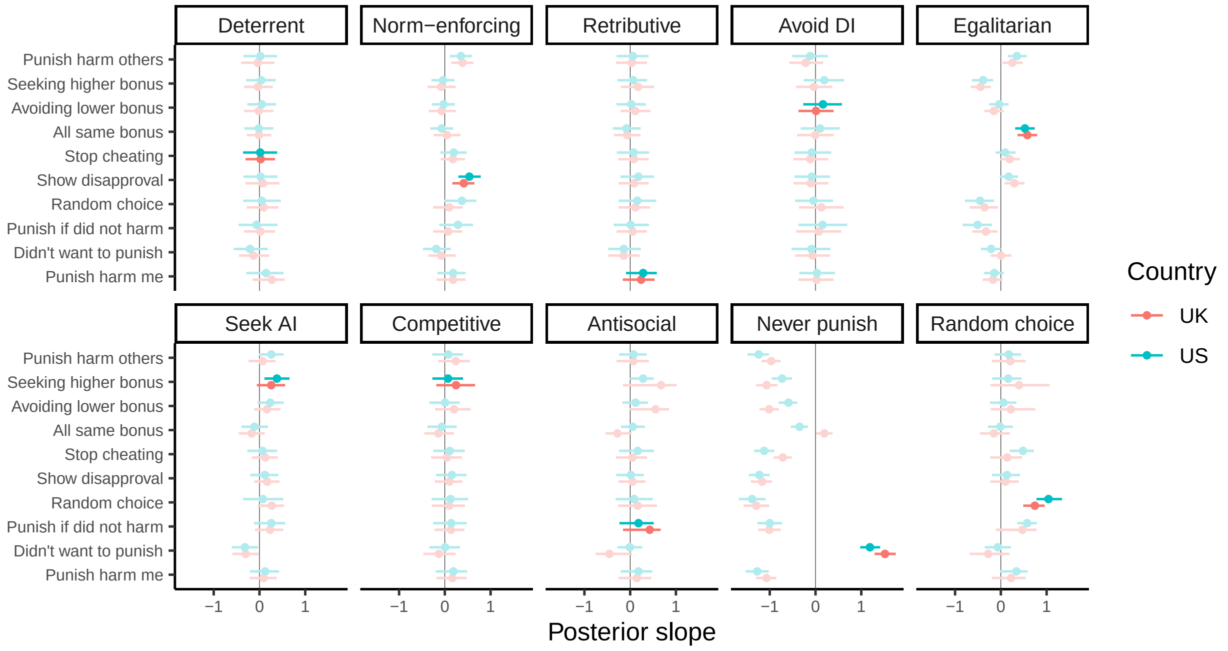
Supplementary Figure S5. *Posterior slopes from models including political ideology, views about social inequality, and religiosity, fitted to the full dataset without pre-registered exclusions.* Each row represents a separate model aside from Social Dominance Orientation and Right Wing Authoritarianism, which control for one another within the same model. Points represent posterior medians, line ranges represent 95% credible intervals.



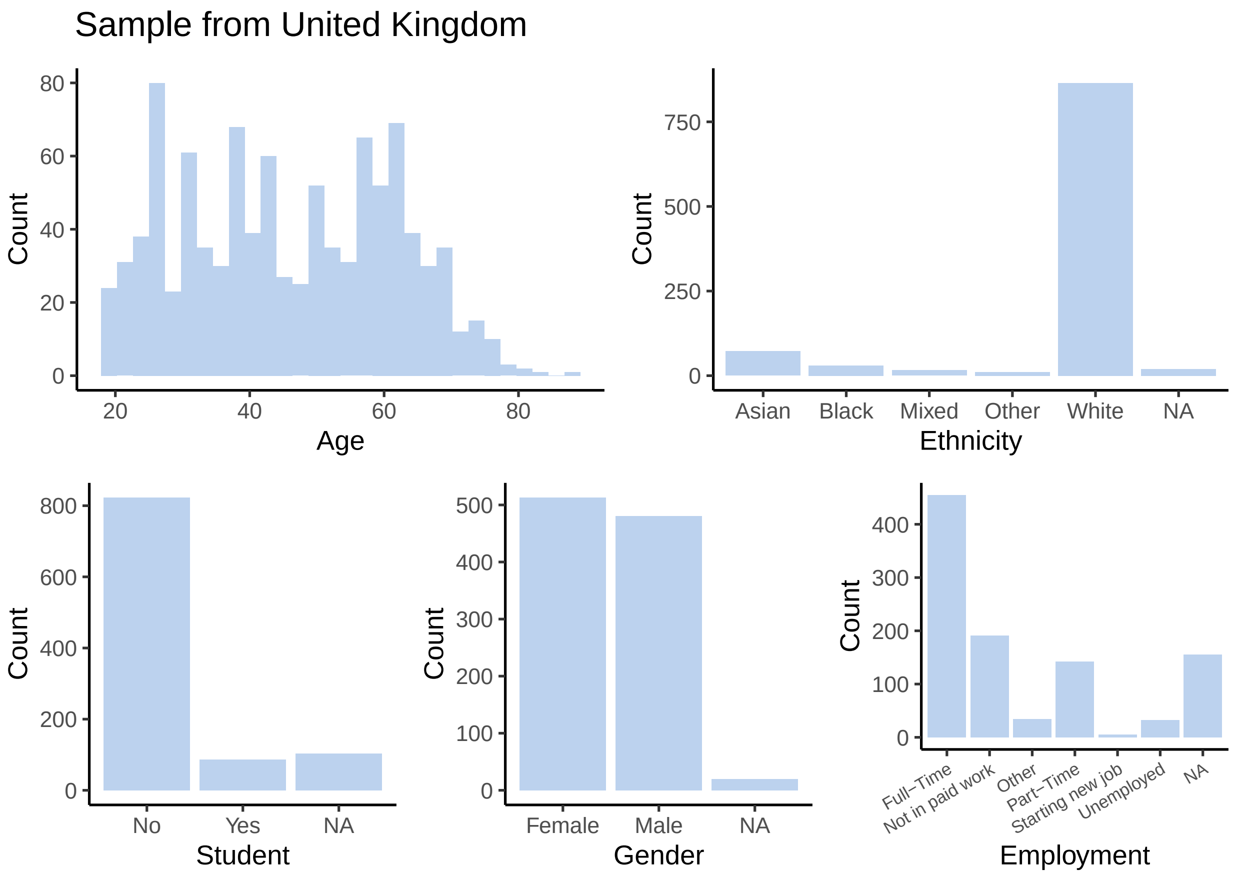
Supplementary Figure S6. Boxplots showing the distribution of responses to each self-report question about the reasons for participants' behaviour in the games. Boxplots represent medians and interquartile ranges.



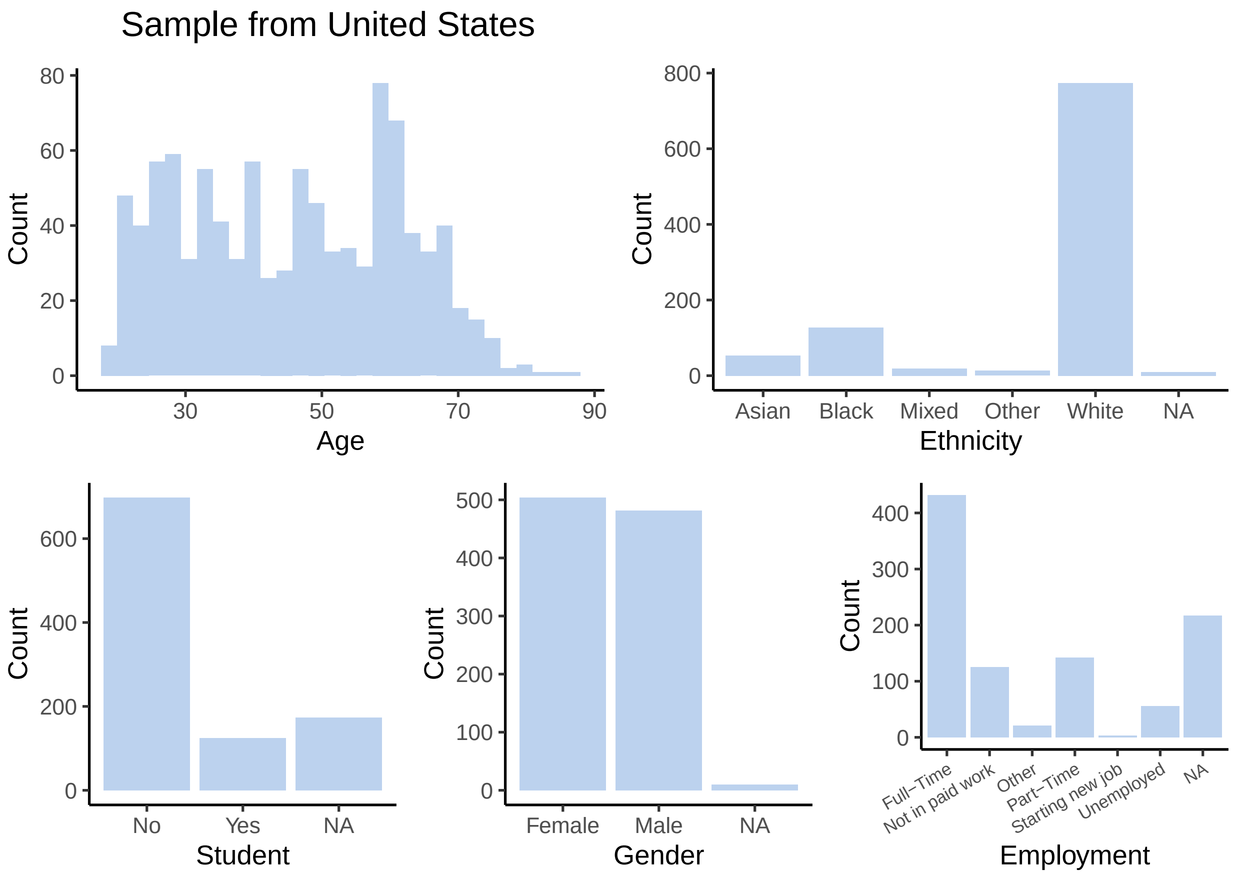
Supplementary Figure S7. Boxplots showing the distribution of responses to each self-report question about the reasons for participants' behaviour in the games, presented as deviations from participants' average rating across all questions. Boxplots represent medians and interquartile ranges.



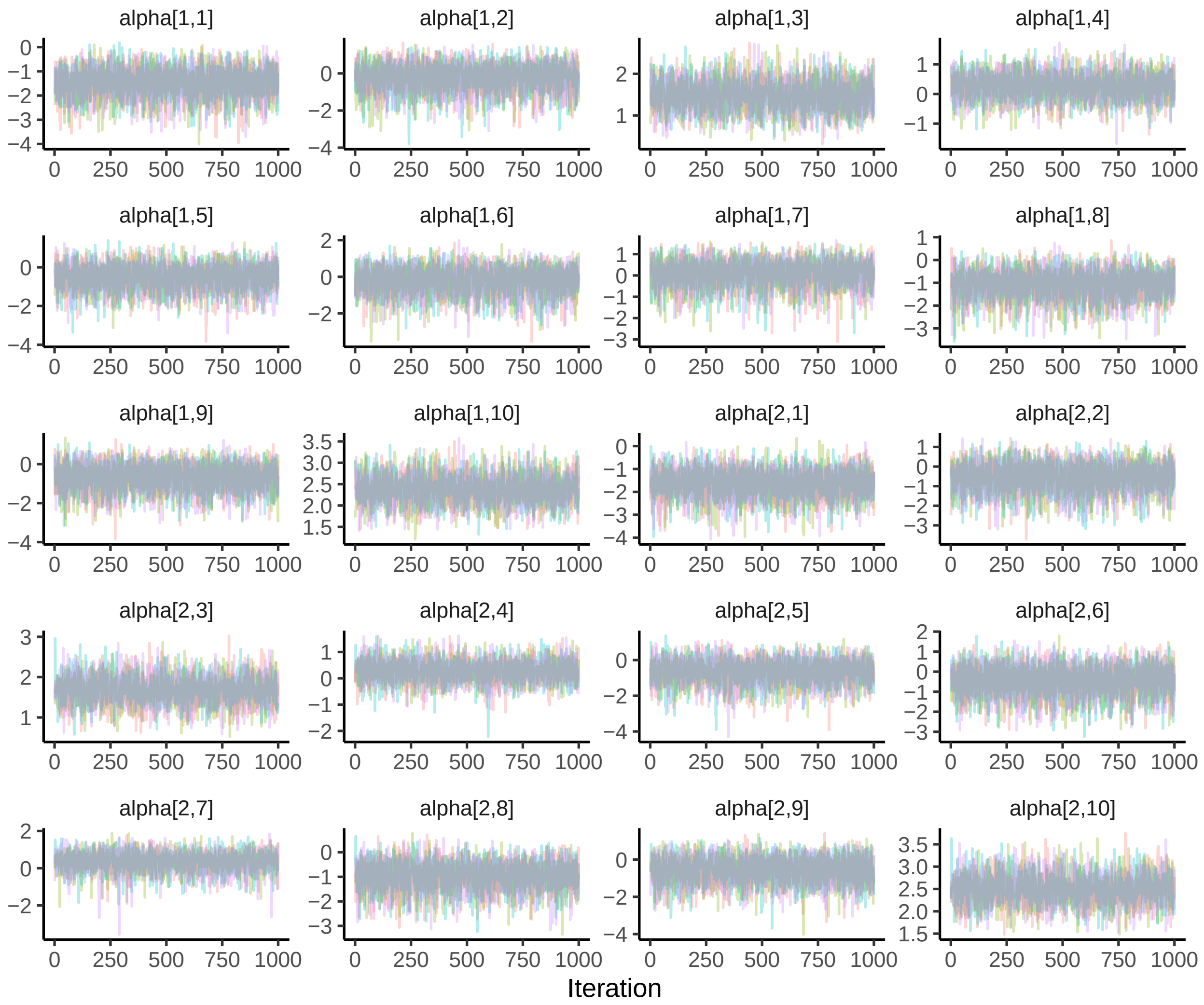
Supplementary Figure S8. *Posterior slopes from models including self-reported strategy usage, fitted to the full dataset without pre-registered exclusions.* Each row represents a separate model. Highlighted estimates represent combinations where the self-report slider matched the behavioural strategy. Each strategy had an associated self-report slider except for the competitive strategy. Points represent posterior medians, line ranges represent 95% credible intervals.



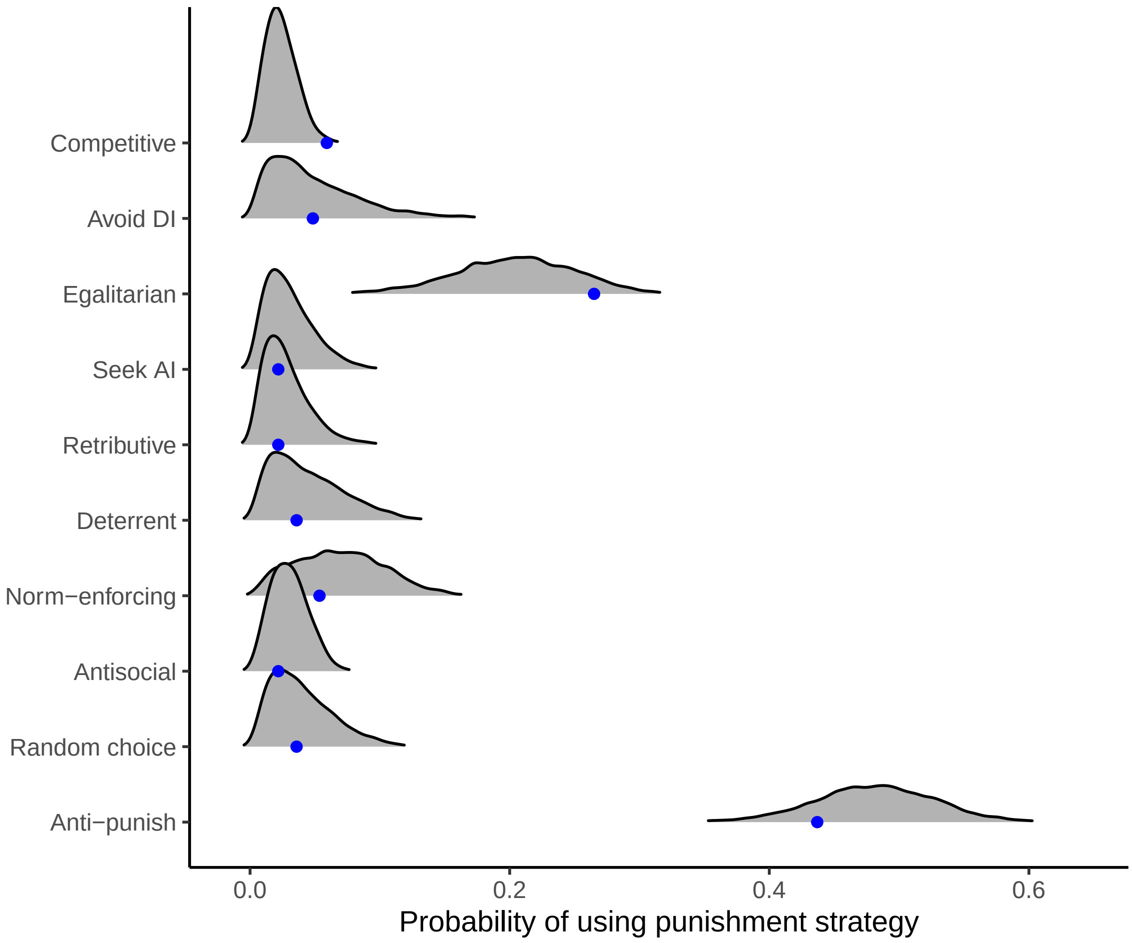
Supplementary Figure S9. *Sample characteristics in the United Kingdom.*



Supplementary Figure S10. *Sample characteristics in the United States.*



Supplementary Figure S11. *Trace plots for parameter values from the Bayesian latent state model fitted to data with exclusions.*



Supplementary Figure S12. Results of Bayesian latent state model fitted to simulated data (n = 100) with known strategy frequencies in the population. Blue points represent known strategy frequencies, grey densities represent posterior estimates of strategy frequencies.

Supplementary Tables

Supplementary Table S1

Counts and proportions of the 25 most common patterns of punitive behaviour across all twelve decisions, split by country. Binary strings represent punishment (1) or no punishment (0) in each decision, aligning with the order of game decision columns in Table 1.

		United Kingdom (N = 1014)		United States (N = 996)	
Pattern	Explanation	N	Prop	N	Prop
000000000000	*Never punish strategy (exact)*	426	0.420	447	0.449
000000001000	*Avoid DI strategy (exact)*	67	0.066	62	0.062
000000001010	*Egalitarian strategy (exact)*	65	0.064	71	0.071
000000000010	Punish when take in Game F	55	0.054	49	0.049
001000001000	Punish when take in Games B and E	14	0.014	11	0.011
101000001010	Punish when take in Games A, B, E, and F	11	0.011	4	0.004
100000000000	Punish when take in Game A	10	0.010	2	0.002
000000100000	Punish when take in Game D	9	0.009	3	0.003
001000001010	Punish when take in Games B, E, and F	9	0.009	17	0.017
101000101000	*Deterrent strategy (exact)*	9	0.009	6	0.006
101010101010	Punish when take in all games	9	0.009	15	0.015
101000101010	*Norm-enforcing strategy (exact)*	8	0.008	16	0.016
001000000000	Punish when take in Game B	7	0.007	4	0.004
001010101000	Punish when take in Games B, C, D, and E	7	0.007	0	0.000
100000001000	Punish when take in Games A and E	6	0.006	5	0.005
101000001000	Punish when take in Games A, B, and E	6	0.006	7	0.007
101010101000	*Retributive strategy (exact)*	6	0.006	5	0.005
111111111111	Always punish	6	0.006	16	0.016
000000101000	Punish when take in Games D and E	5	0.005	2	0.002
000000101010	Punish when take in Games D, E, and F	5	0.005	3	0.003
101010001010	Punish when take in all games except Game D	5	0.005	2	0.002
001000101000	Punish when take in Games B, D, and E	4	0.004	2	0.002
001000101010	Punish when take in Games B, D, E, and F	4	0.004	6	0.006
101000000000	Punish when take in Games A and B	4	0.004	2	0.002
101010001000	Punish when take in Games A, B, C, and E	4	0.004	0	0.000

Supplementary Table S2

Wordings for 11 self-report slider questions asking participants to report the reasons for their behaviour in the six games. Participants were prompted with the following text: “We would now like you to answer a few questions about your main motivation in the games. Please answer truthfully - there is no right or wrong answer and your first answer is probably best. Please rate the extent to which the following statements apply to your decisions to reduce or not to reduce other players’ bonuses in the games.”

Slider	Wording
1	I wanted to punish people who harmed others
2	I wanted to have a higher final bonus than others
3	I wanted to avoid having a lower final bonus than others
4	I wanted all players to have the same final bonus
5	I wanted to stop others from cheating
6	I wanted to show that I disapproved of others’ actions
7	I made decisions at random
8	I wanted to punish people who DID NOT harm me or others
9	I didn’t want to reduce anyone’s bonus, no matter what they did
10	I didn’t want to PAY to reduce anyone’s bonus but I would have done so if it were free
11	I wanted to punish people if they harmed me but not if they harmed others

Supplementary Table S3

Wordings for survey questions in the study.

Measure	Wording	Scale
Demographics	What is your highest level of education?	
	Where would you place yourself on this ladder? Please indicate which number on the rung best represents where you stand at this time in your life, relative to other people in your country	
	Please could you tell us roughly how many years have you lived in your current country of residence?	
Big 6 Extraversion	I am the life of the party	1-7
	I don't talk a lot (reversed)	1-7
	I keep in the background (reversed)	1-7
	I talk to a lot of different people at parties	1-7
Big 6 Agreeableness	I sympathise with others' feelings	1-7
	I am not interested in other people's problems (reversed)	1-7
	I feel others' emotions	1-7
	I am not really interested in others (reversed)	1-7
Big 6 Conscientiousness	I get chores done right away	1-7
	I like order	1-7
	I make a mess of things (reversed)	1-7
	I often forget to put things back in their proper place (reversed)	1-7
Big 6 Neuroticism	I have frequent mood swings	1-7
	I am relaxed most of the time (reversed)	1-7
	I get upset easily	1-7
	I seldom feel blue (reversed)	1-7
Big 6 Openness to experience	I have a vivid imagination	1-7
	I have difficulty understanding abstract ideas	1-7
	I do not have a good imagination (reversed)	1-7
	I am not interested in abstract ideas (reversed)	1-7

Table S3 continued

Measure	Wording	Scale
Big 6 Honesty-humility	I feel entitled to more of everything (reversed)	1-7
	I deserve more things in life (reversed)	1-7
	I would like to be seen driving around in a very expensive car (reversed)	1-7
	I would get a lot of pleasure from owning expensive luxury goods (reversed)	1-7
Social Value Orientation	Please indicate how you would like to distribute money between yourself and the other player	9 choices
Left-right political ideology	Political views are often organised on a single scale from left to right. For example, in the United States, the Democratic Party is described as more to the left and the Republican Party is described as more to the right. If you had to place your political views on this left-right scale, generally speaking, where would you put yourself?	0-100 slider
Social Dominance Orientation	An ideal society requires some groups to be on top and others to be on the bottom	1-7
	Some groups of people are simply inferior to other groups	1-7
	No one group should dominate in society (reversed)	1-7
	Groups at the bottom are just as deserving as groups at the top (reversed)	1-7
	Group equality should not be our primary goal	1-7
	It is unjust to try to make groups equal	1-7
	We should do what we can to equalize conditions for different groups (reversed)	1-7
	We should work to give all groups an equal chance to succeed (reversed)	1-7
Right Wing Authoritarianism	It's great that many young people today are prepared to defy authority (reversed)	1-9

Table S3 continued

Measure	Wording	Scale
	What our country needs most is discipline, with everyone following our leaders in unity	1-9
	God's laws about abortion, pornography, and marriage must be strictly followed before it is too late	1-9
	There is nothing wrong with premarital sexual intercourse (reversed)	1-9
	Our society does NOT need tougher government and stricter laws (reversed)	1-9
	The facts on crime and the recent public disorders show we have to crack down harder on troublemakers, if we are going to preserve law and order	1-9
Views on social inequality	I would like to bring the people above me on the ladder down a peg or two	1-7
	I would like to bring the people below me on the ladder up a peg or two	1-7
Religious views	How religious are you?	1-5
	It is likely that God, or some other type of spiritual non-human entity, controls the events in the world	1-7

Supplementary Table S4

Proportions of correct answers to comprehension questions for all six economic games, split by country.

Game	United Kingdom	United States
Game A (AI)	0.96	0.94
Game B (Equal)	0.95	0.93
Game C (Computer)	0.95	0.95
Game D (1:1 Fee-Fine)	0.95	0.94
Game E (DI)	0.96	0.94
Game F (Third-Party)	0.95	0.94