Mapping out the punishment strategy space

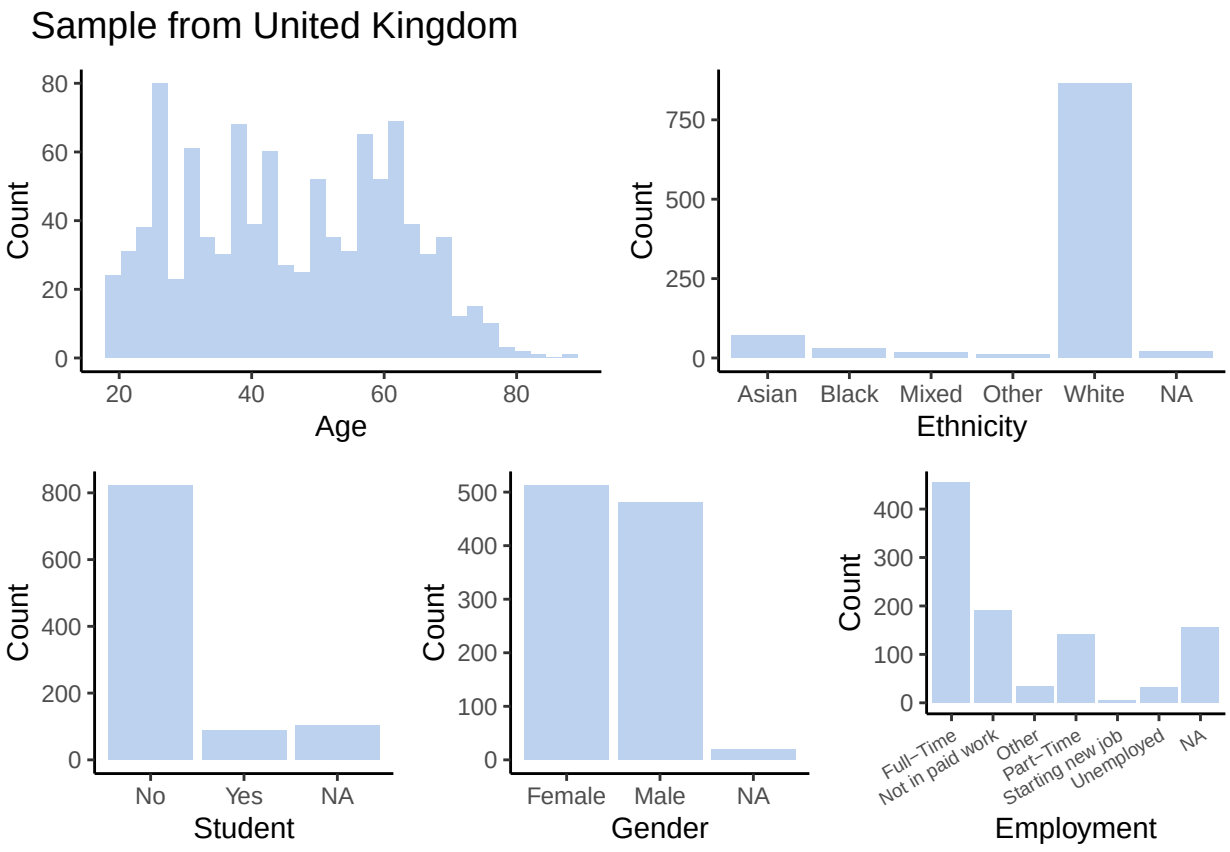
Scott Claessens

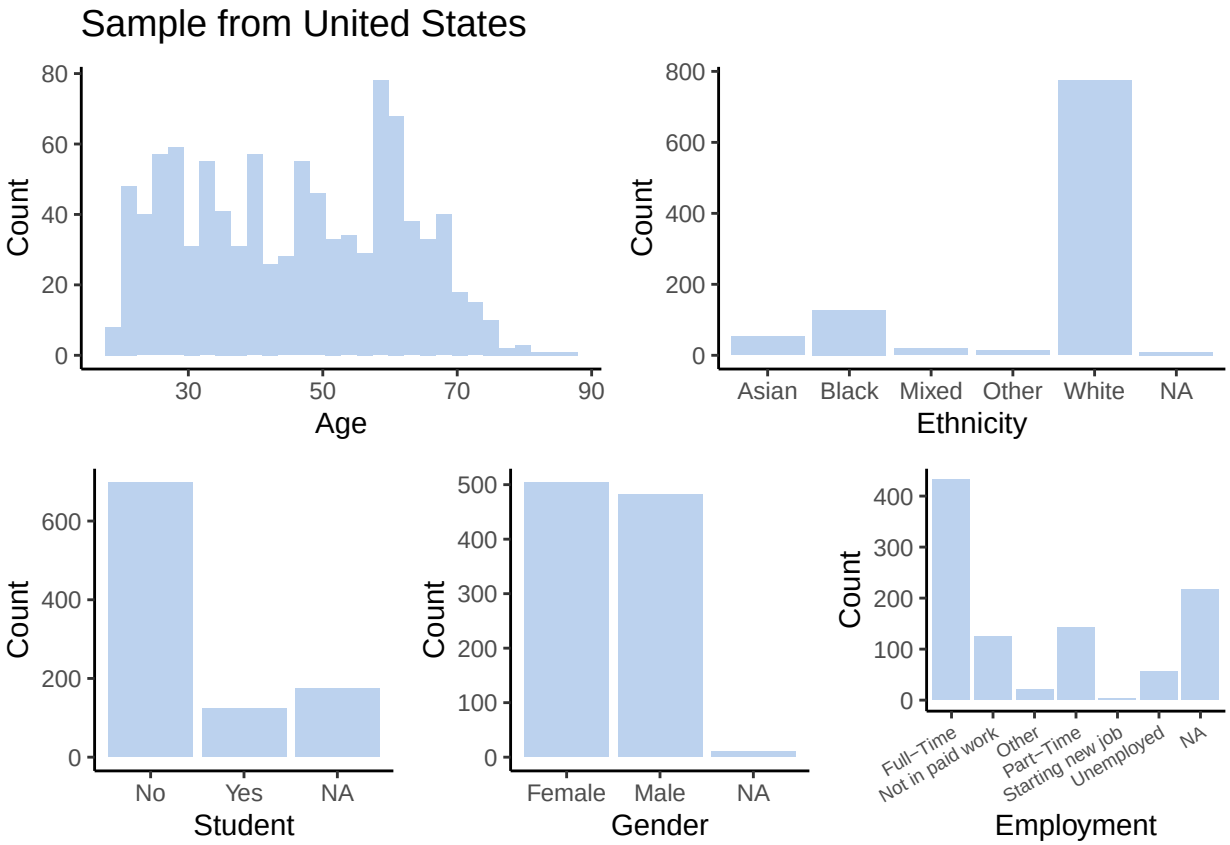
2023-11-22

This document outlines the data exploration and analyses for our project "Mapping out the punishment strategy space".

Sample

After cleaning the data, we have data for 2024 participants from Prolific. Participants are representative samples from both the United Kingdom (n = 1019) and the United States (n = 1005).





Punishment games

We asked participants to respond to six games where they had the opportunity to punish another player for their behaviour. We refer to these games as follows:

- No Disadvantageous Inequity 1
- No Disadvantageous Inequity 2
- No Disadvantageous Inequity 3 (Computer)
- No Disadvantageous Inequity 4 (1:1 Fee Fine Ratio)
- Disadvantageous Inequity
- Third-Party

In each game, participants could punish (1) when the other player chose to "take" and (2) when the other player did nothing. For more details about these games (e.g. exact payoff structures), see preregistration.

Comprehension

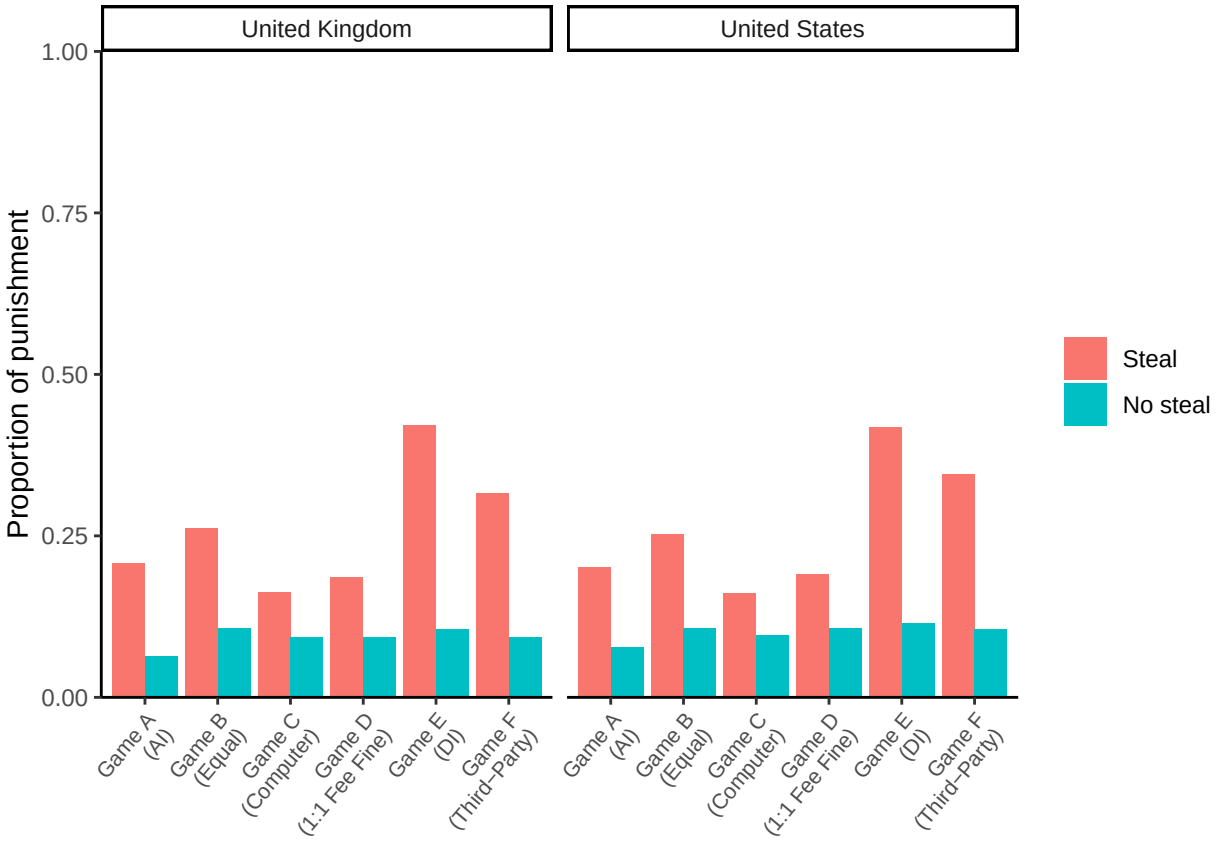
Answers to the comprehension questions revealed that participants were able to understand the payoff structure of all six punishment games. Here are the comprehension rates in both countries:

Game	United Kingdom	United States
Game A (AI)	0.96	0.94

Game	United Kingdom	United States
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

Punishment decisions

We can plot the proportion of participants who decided to punish in each game.



The pattern is very similar in both countries. Participants appear more likely to punish if the other player took, compared to when they did nothing. Participants were most likely to punish when the other player took in the disadvantageous inequity game and in the third-party game.

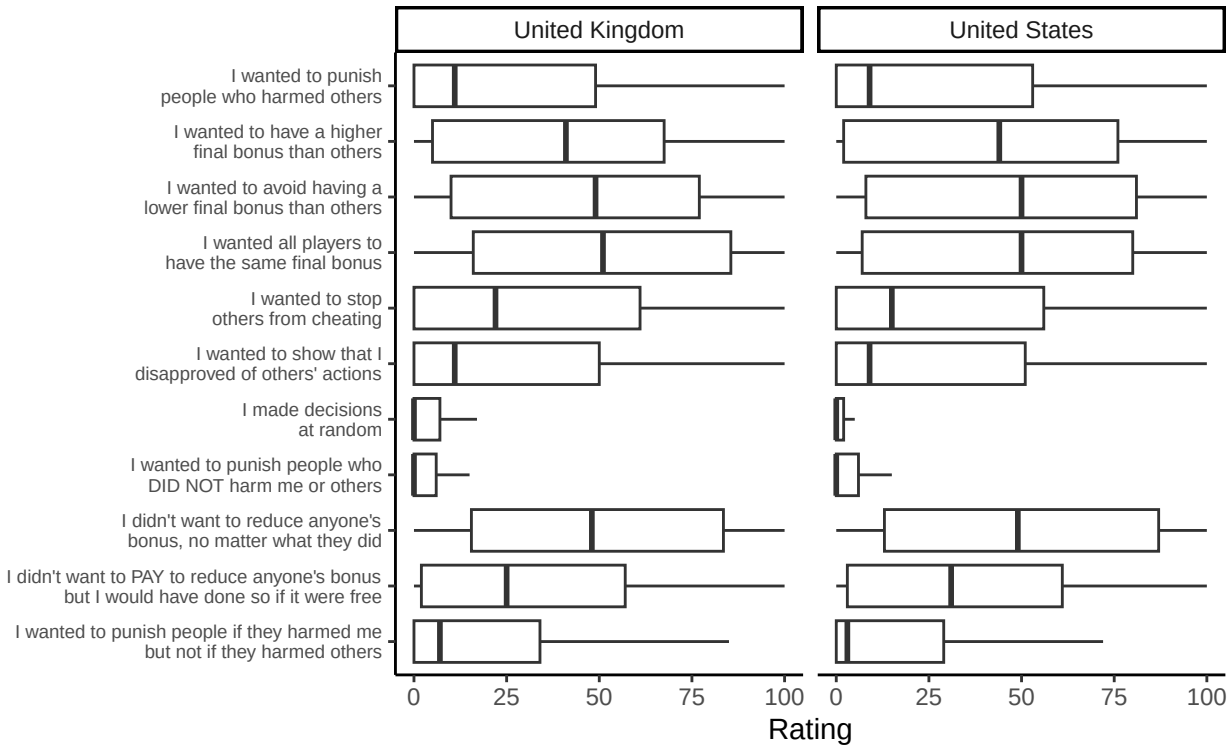
Reasons given for punishing in the games

At the end of the survey, we asked participants why they decided to punish (if they ever did). First, we allowed them to provide an open-ended answer to this question. The following wordclouds summarise frequently used words in these open-ended answers.

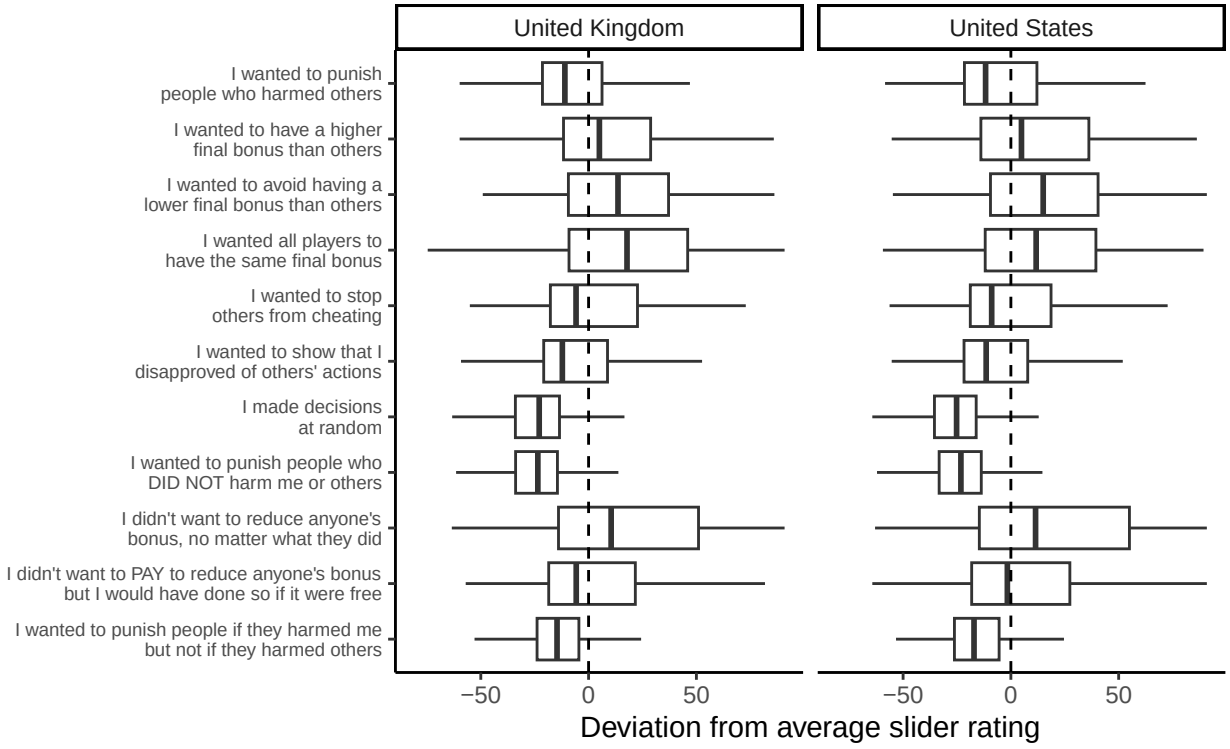
Here is the wordcloud for the United Kingdom sample:

Here is the wordcloud for the United States sample:

Second, we gave participants a number of sliders on which they could rate how strongly they followed different approaches to the games. Below are the raw distributions from these slider scales (0 = statement does not apply, 100 = statement does apply).



We can also visualise these distributions as deviations from participant's average ratings across all sliders.



In both countries, participants reported being especially motivated by equality, avoiding disadvantageous inequity, and seeking advantageous inequity. People also expressed that they never punished.

Frequencies of punishment strategies

Before data collection, we posited ten different strategies that might underlie people's punishment behaviour in the games:

- Competitive
- Avoid disadvantageous inequity
- Egalitarian
- Seek advantageous inequity
- Retributive
- Deterrent
- Norm-enforcing
- Antisocial
- Random choice
- Anti-punish

We pre-registered predictions for how these strategies would behave in the different games.

Counts and proportions from raw data

As a first step, we can look to see the proportion of participants who fitted these strategy predictions *exactly* across all games.

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

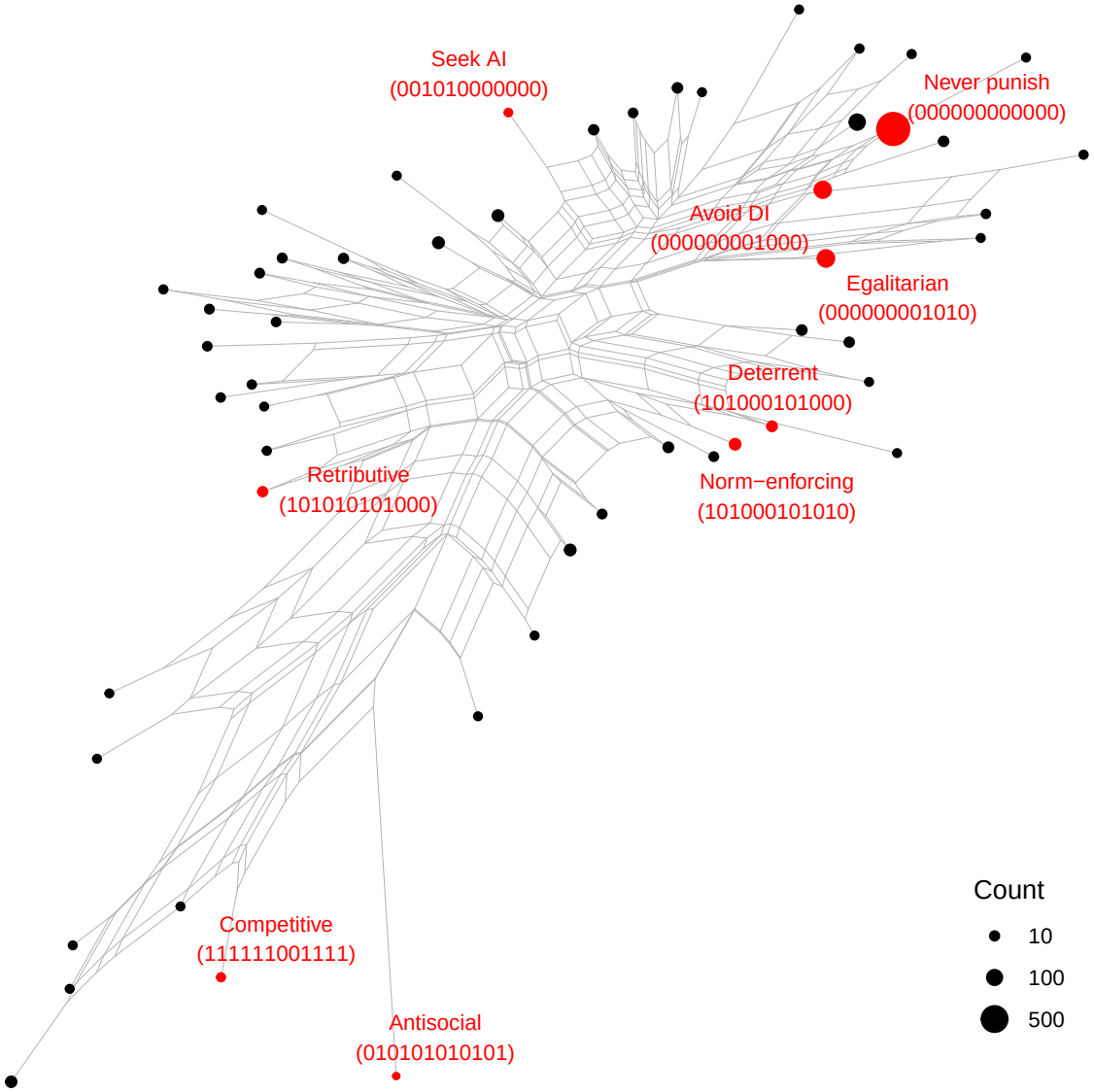
Many participants, denoted by N/A, were unable to be classified into a strategy (i.e. their pattern of behaviour across all games did not fit any of our strategy predictions). However, of the participants who could be classified, most followed the "anti-punish" strategy by never punishing. The next most common strategies were the "egalitarian" and "avoid disadvantageous inequity" strategies.

We can also look at the most common behavioural patterns across all six games (twelve punishment decisions in total). Below, we summarise the 25 most common patterns of behaviour as strings of twelve 0s and 1s, indicating whether participants did (1) or did not (0) punish in each game, with a brief explanation of this pattern and its frequency in the dataset.

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

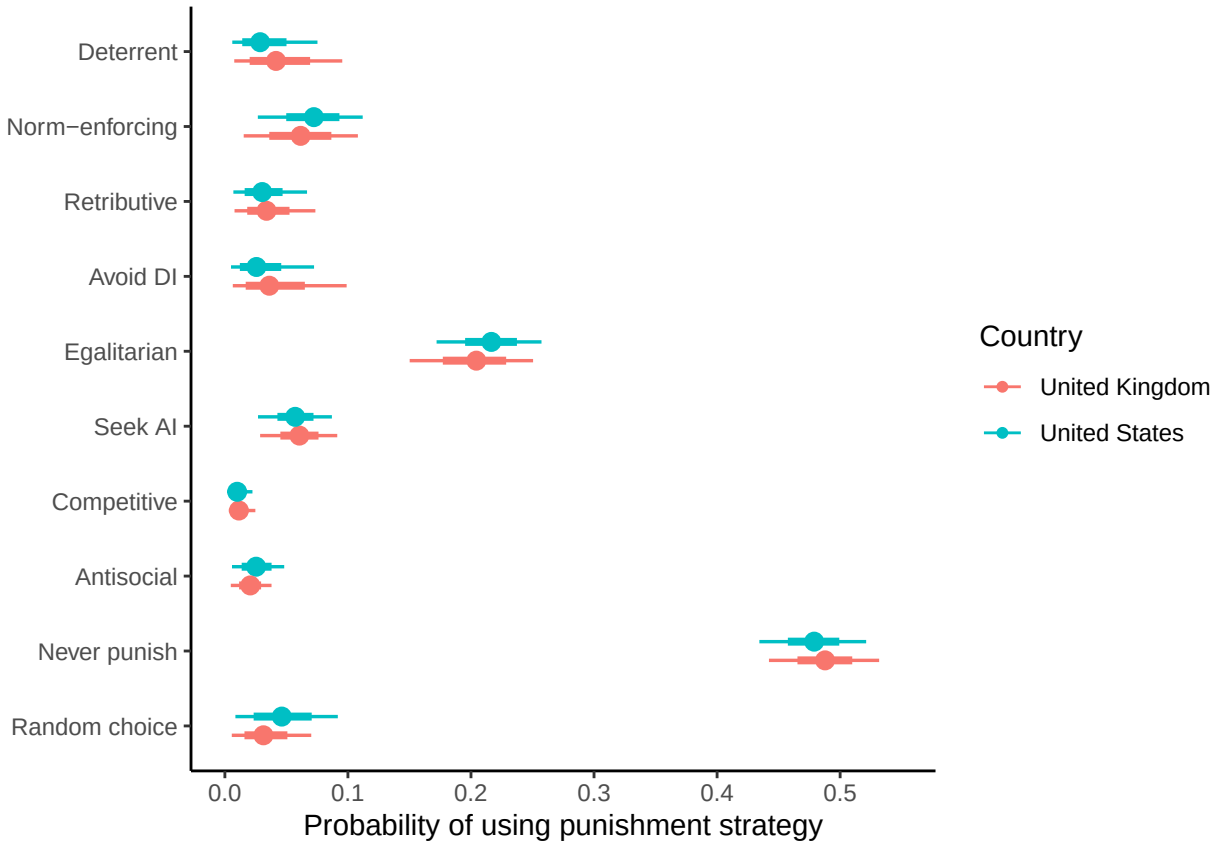
Pattern	Explanation	N	Prop	N	Prop
001000001000	Punish when take in Games B and E	14	0.014	11	0.011
101000001010	Punish when take in Games A, B, E, and F	11	0.011	4	0.004
100000000000	Punish when take in Game A	10	0.010	2	0.002
000000100000	Punish when take in Game D	9	0.009	3	0.003
001000001010	Punish 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

These behavioural patterns can be visualised, along with their frequency of usage, on a splits graph. This graph plots the distance between the difference strategies as proportional to the number of substitutions required to get from one to another. The top of the graph captures the less punitive strategies, and the strategies become more punitive towards the bottom of the graph. We only include behavioural patterns followed by at least three participants.



Bayesian modelling

We can also estimate the frequencies of different strategies using a Bayesian approach. We construct a model that contains our *apriori* predictions for the different punishment strategies, and the feed the model our raw data to estimate the relative probabilities of following each strategy. In the model, we assume that participants sometimes make errors in converting their strategy into behaviour (5% error rate), which could explain why many of the participants were unable to be classified into a strategy type in our raw counts and proportions above. The Bayesian model is fitted in the probabilistic programming language Stan.



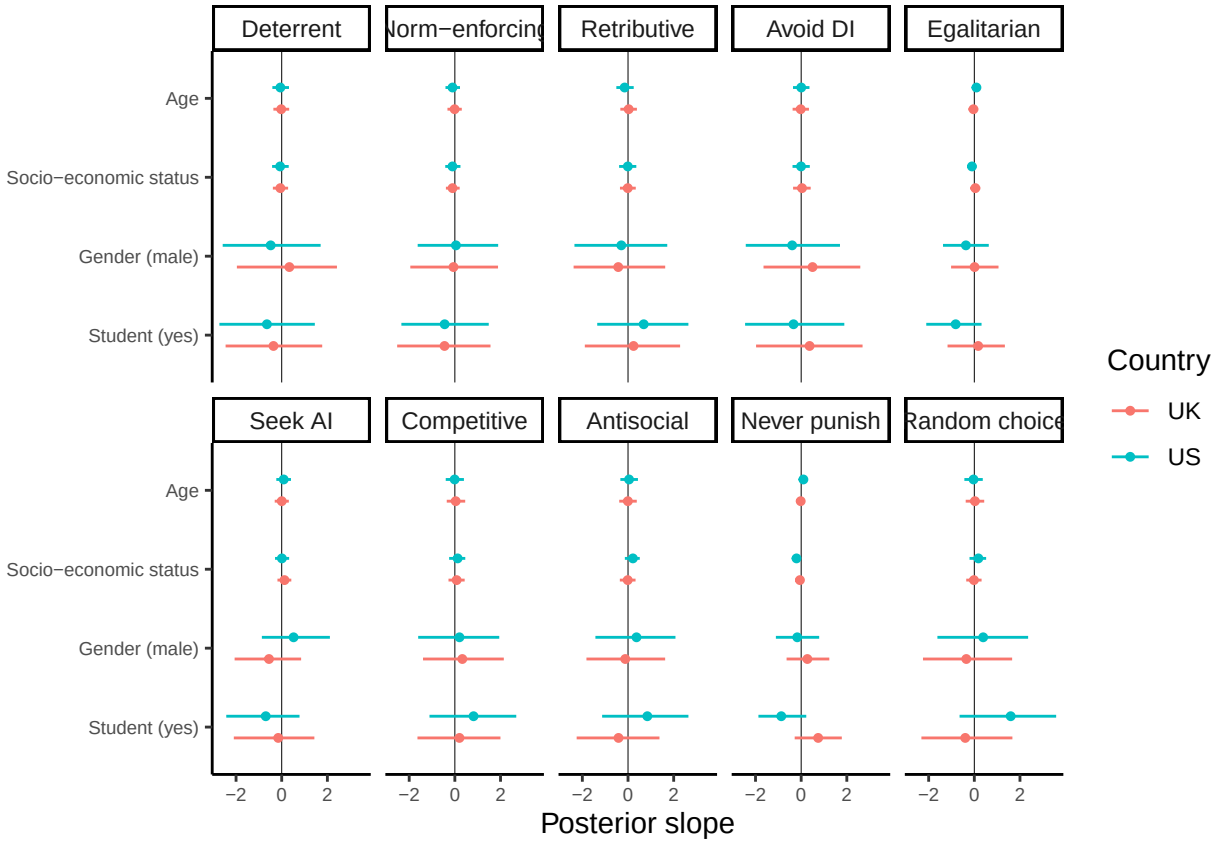
The pattern is similar in both countries. Taking advantage of all available data, the model suggests that "anti-punish" is the most common strategy. Of the punishment strategies, "egalitarian" is the most common. All other strategies have median posterior probabilities less than 10%. "Competitive" and "antisocial" punishment strategies are the most unlikely.

Why is egalitarian the winner in this model, considering that it wasn't an obvious winner in the raw counts? One explanation is that many participants punished only in the third-party punishment game (see table above). The model does not include this pattern of behaviour as an explicit strategy. This pattern of behaviour is one substitution away from egalitarian, but two substitutions away from avoid DI, so it upweights egalitarian accordingly.

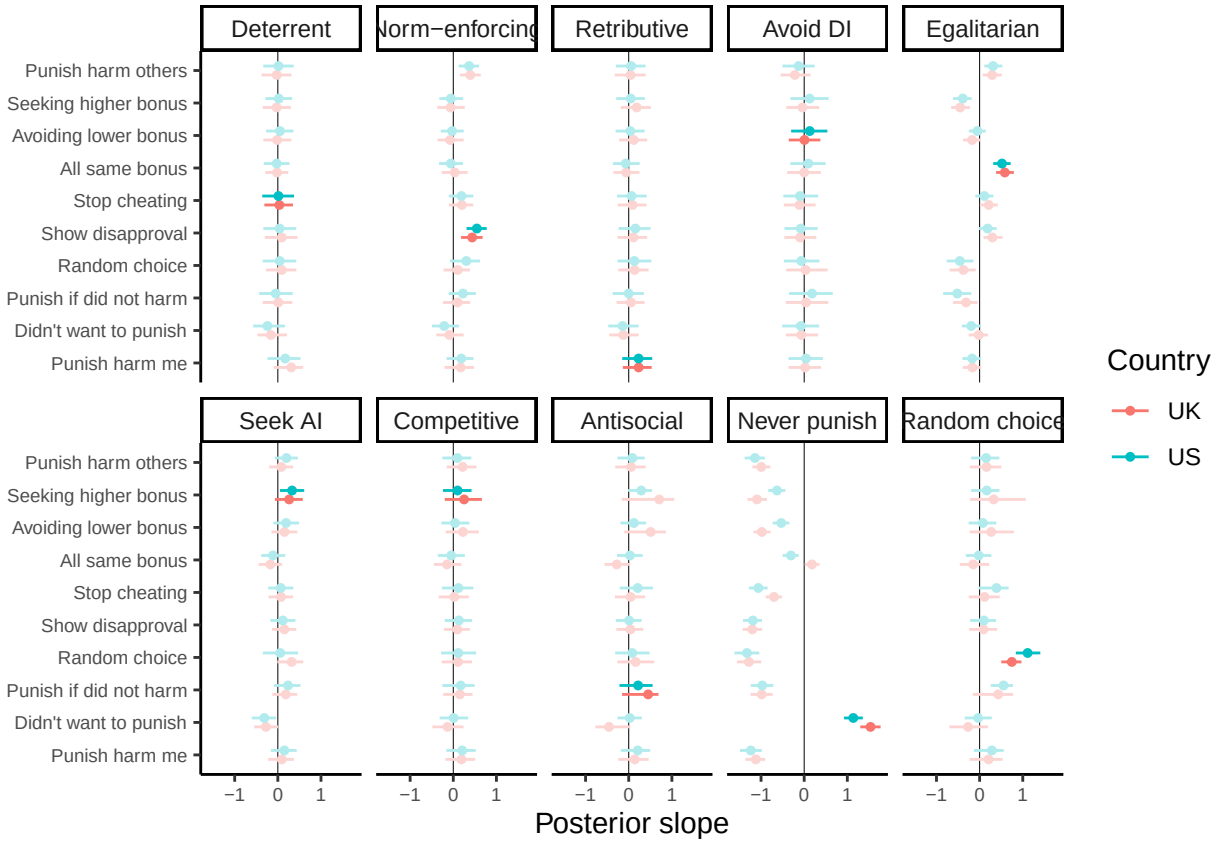
Predicting strategy usage

We then fit a series of models predicting the usage of different strategies from a variety of predictors. In each plot that follows, we show the posterior slopes from several models that include different variables as predictors of all ten strategies in both countries. Each predictor is included separately, apart from SDO and RWA where we include both in the model simultaneously to control for one another.

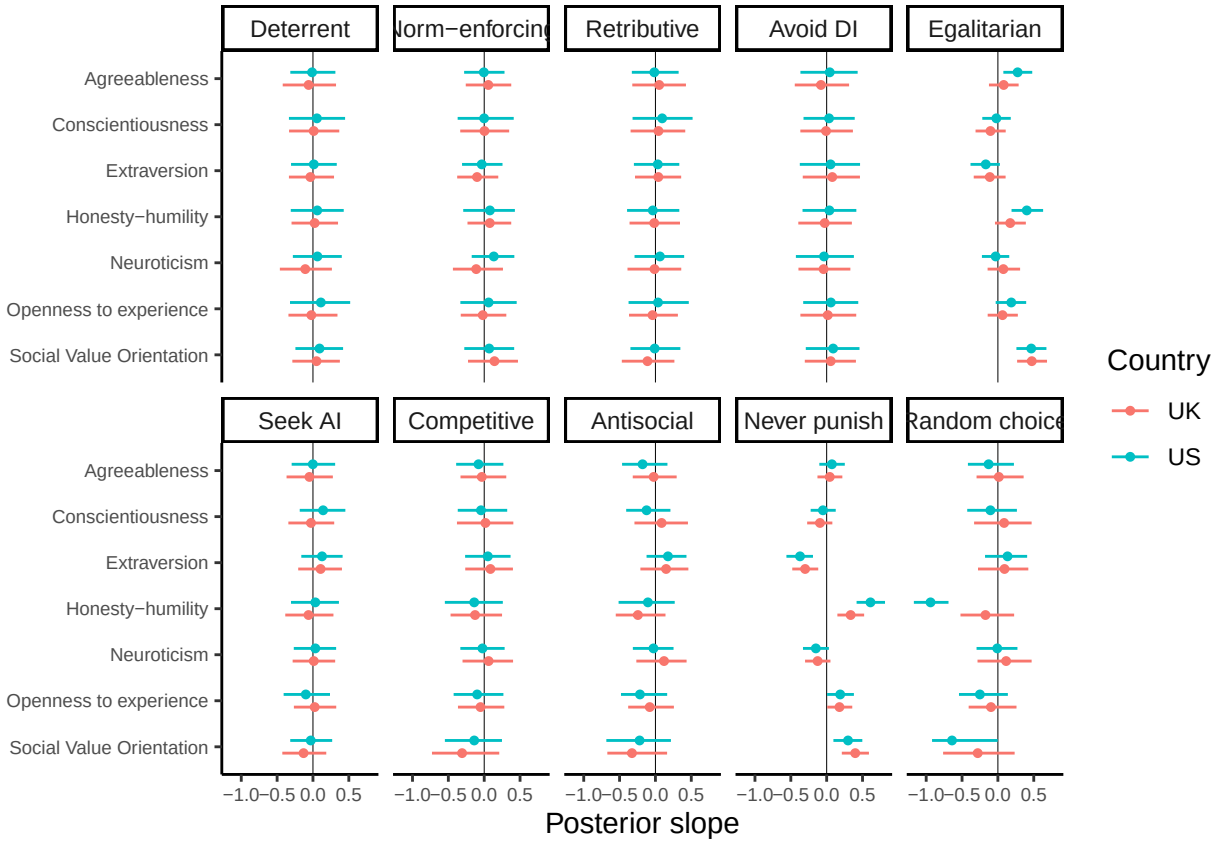
Demographics



Self-ratings



Big-6 personality and SVO



Politics and religion