Mapping out the punishment strategy space

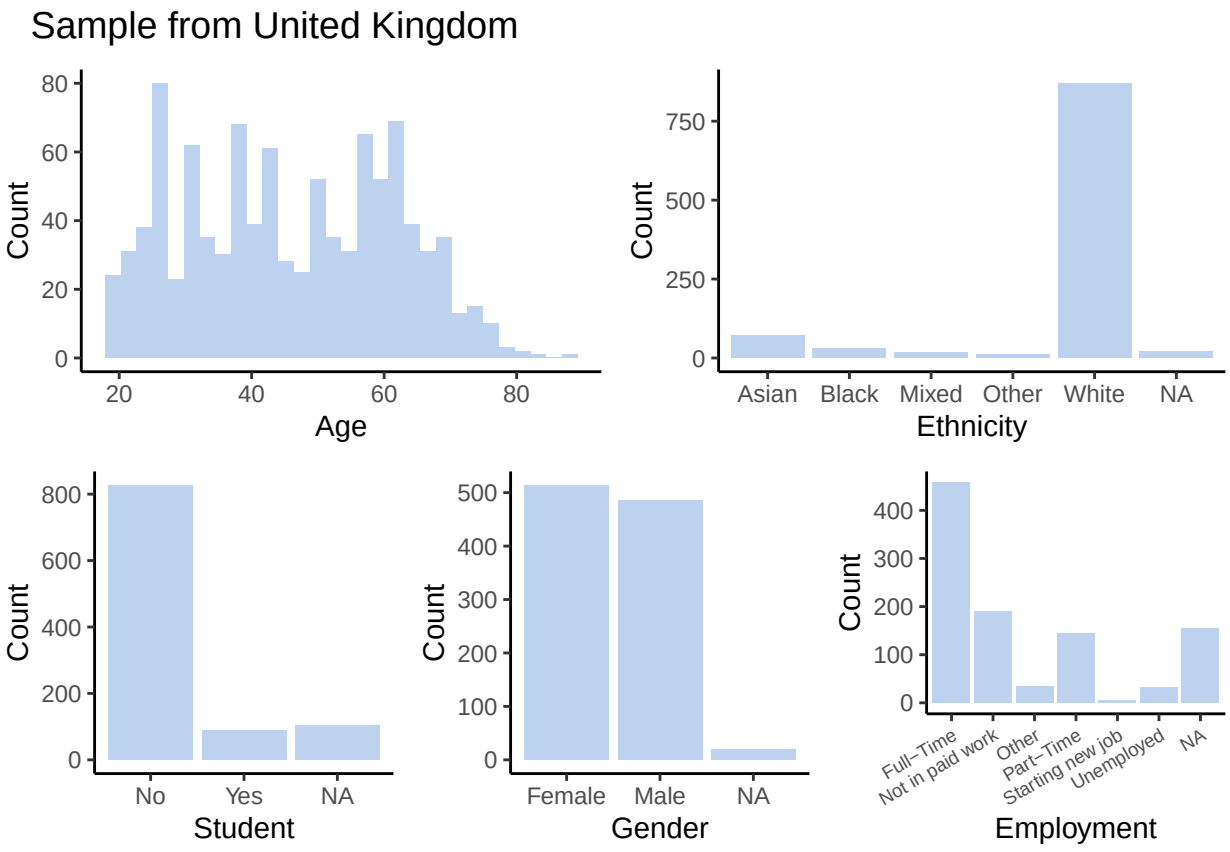
Scott Claessens

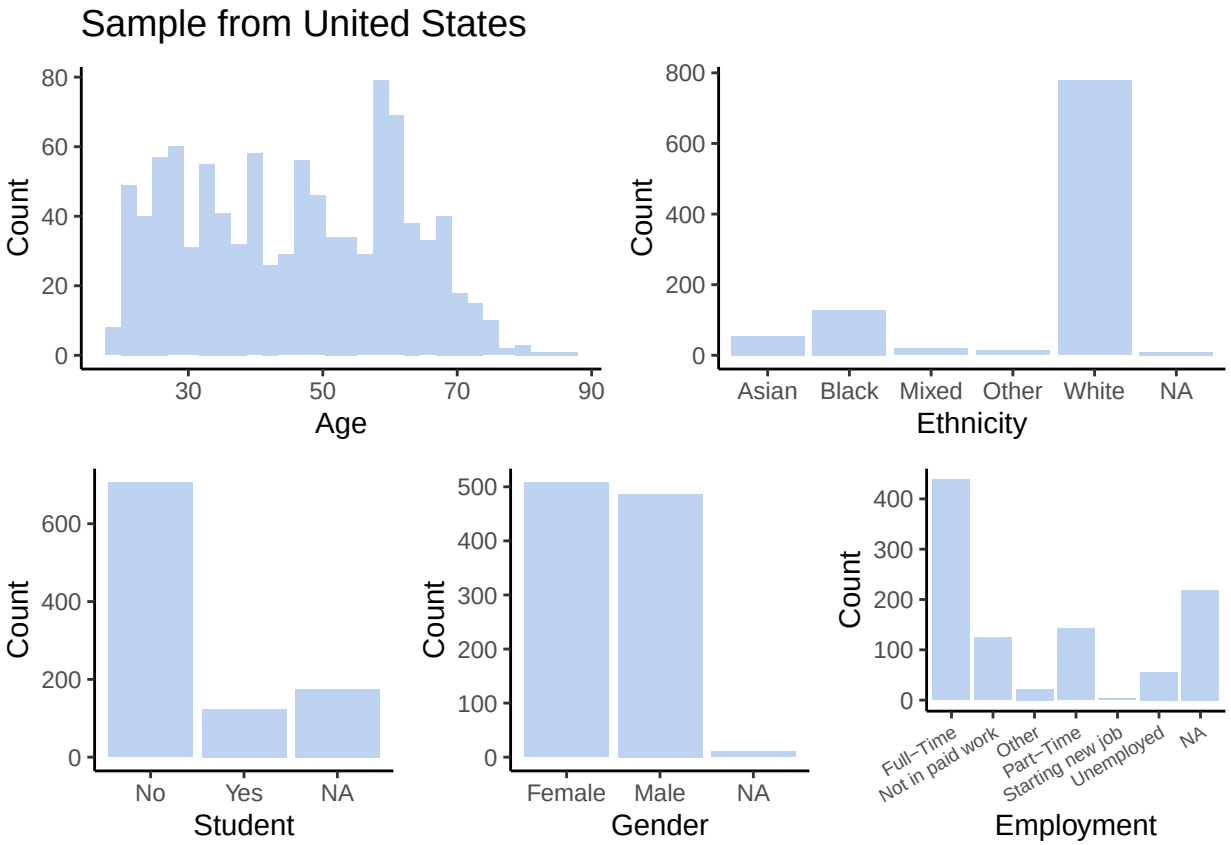
2023-09-08

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

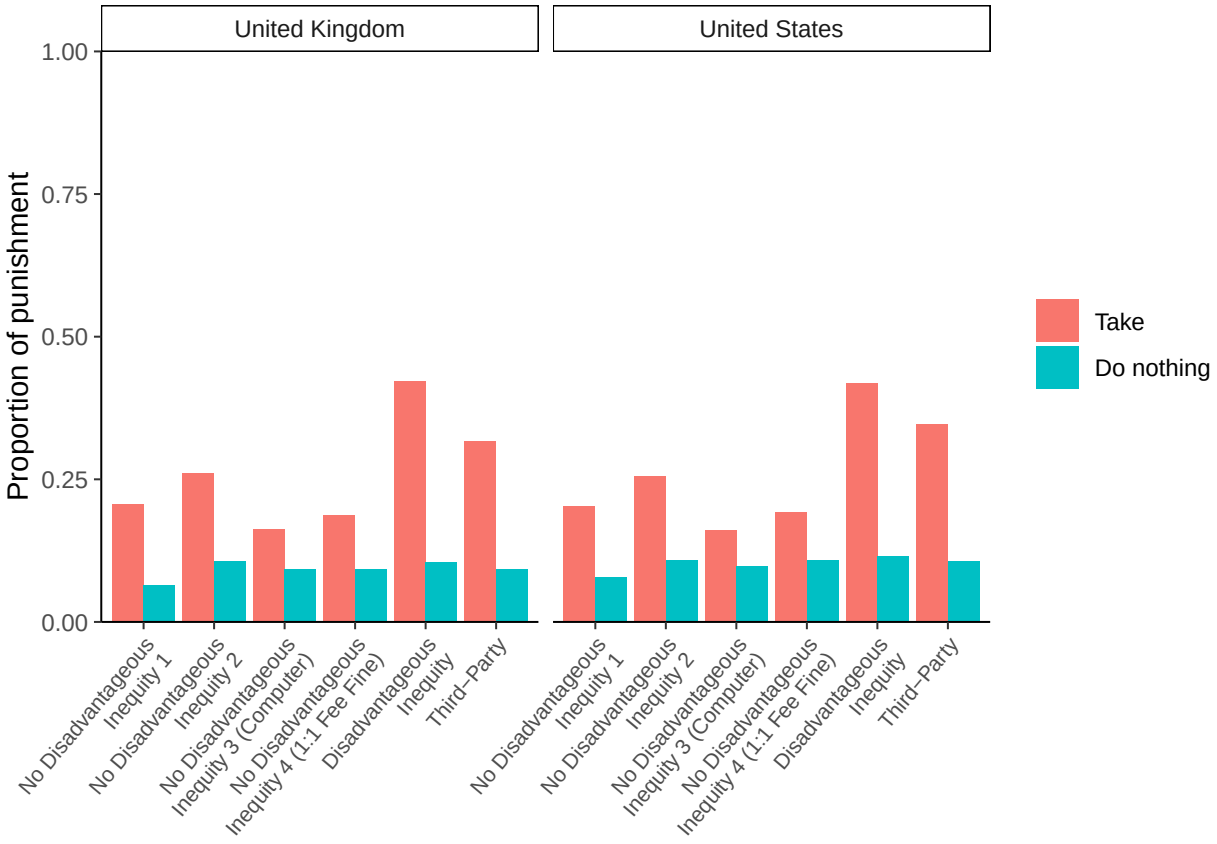
Answers to the comprehension questions revealed that participants were able to understand the payoff structure of all six punishment games. Here are the comprehension rates in both countries:

Game	United Kingdom	United States
No Disadvantageous Inequity 1	0.96	0.94

Game	United Kingdom	United States
No Disadvantageous Inequity 2	0.95	0.93
No Disadvantageous Inequity 3 (Computer)	0.95	0.95
No Disadvantageous Inequity 4 (1:1 Fee Fine)	0.95	0.94
Disadvantageous Inequity	0.96	0.94
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.

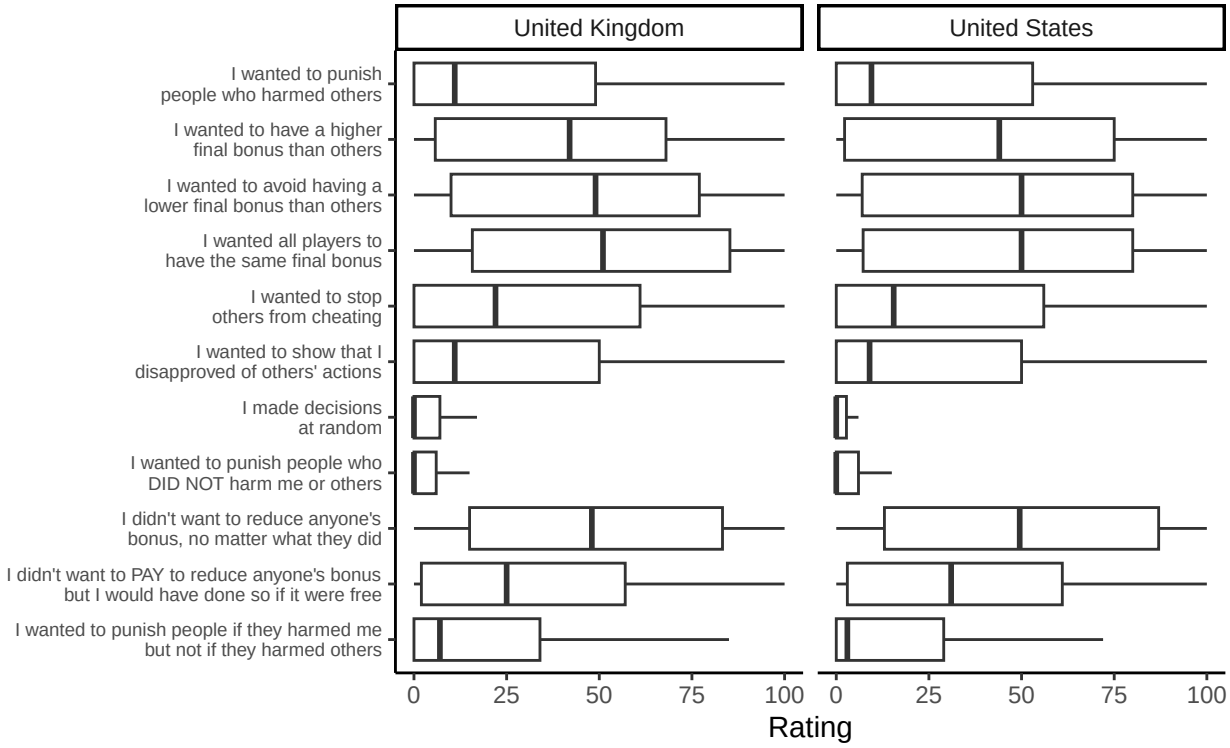
Reasons given for punishing in the games

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.

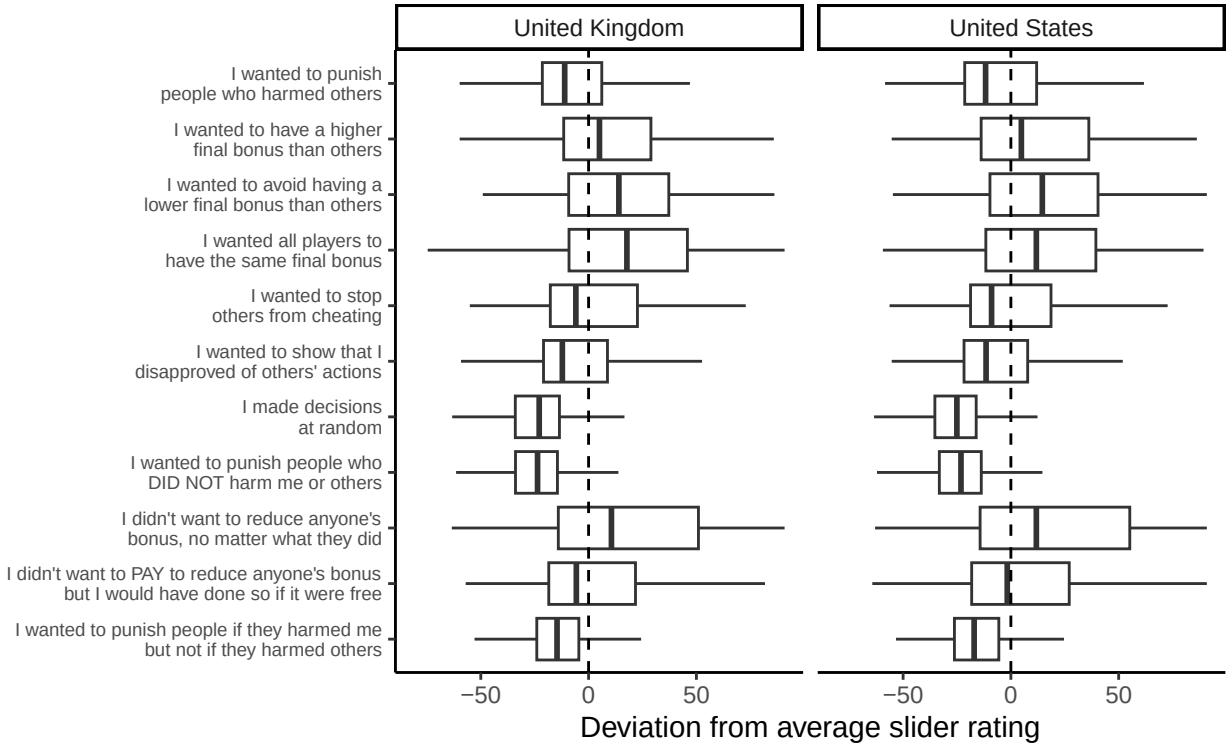
Here is the wordcloud for the United Kingdom sample:

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	Proportion
Anti-punish	878	0.434
N/A	825	0.408
Egalitarian	136	0.067
Avoid disadvantageous inequity	129	0.064
Norm-enforcing	24	0.012
Deterrent	15	0.007
Retributive	11	0.005
Competitive	4	0.002
Seek advantageous inequity	2	0.001

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 assume here that people are following a particular strategy exactly and are not comitting any errors or mistakes. However, it is possible that some participants intended to follow a particular strategy, but made a mistake on a particular decision. We repeat the counts and proportions in the table above, but include behavioural patterns that have one mistake from the exact strategy (note that the proportions will no longer sum to 1).

Strategy	N	Proportion
Avoid DI	1197	0.591
Anti-punish	1165	0.576
Egalitarian	425	0.210
Norm-enforcing	93	0.046

Strategy	N	Proportion
Deterrent	77	0.038
Retributive	63	0.031
Seek AI	18	0.009
Competitive	10	0.005

And including behavioural patterns that have one or two mistakes.

Strategy	N	Proportion
Avoid DI	1457	0.720
Egalitarian	1419	0.701
Anti-punish	1394	0.689
Seek AI	968	0.478
Norm-enforcing	212	0.105
Deterrent	209	0.103
Retributive	154	0.076
Competitive	43	0.021
Antisocial	1	0.000

And finally, including behavioural patterns that have one, two, or three mistakes.

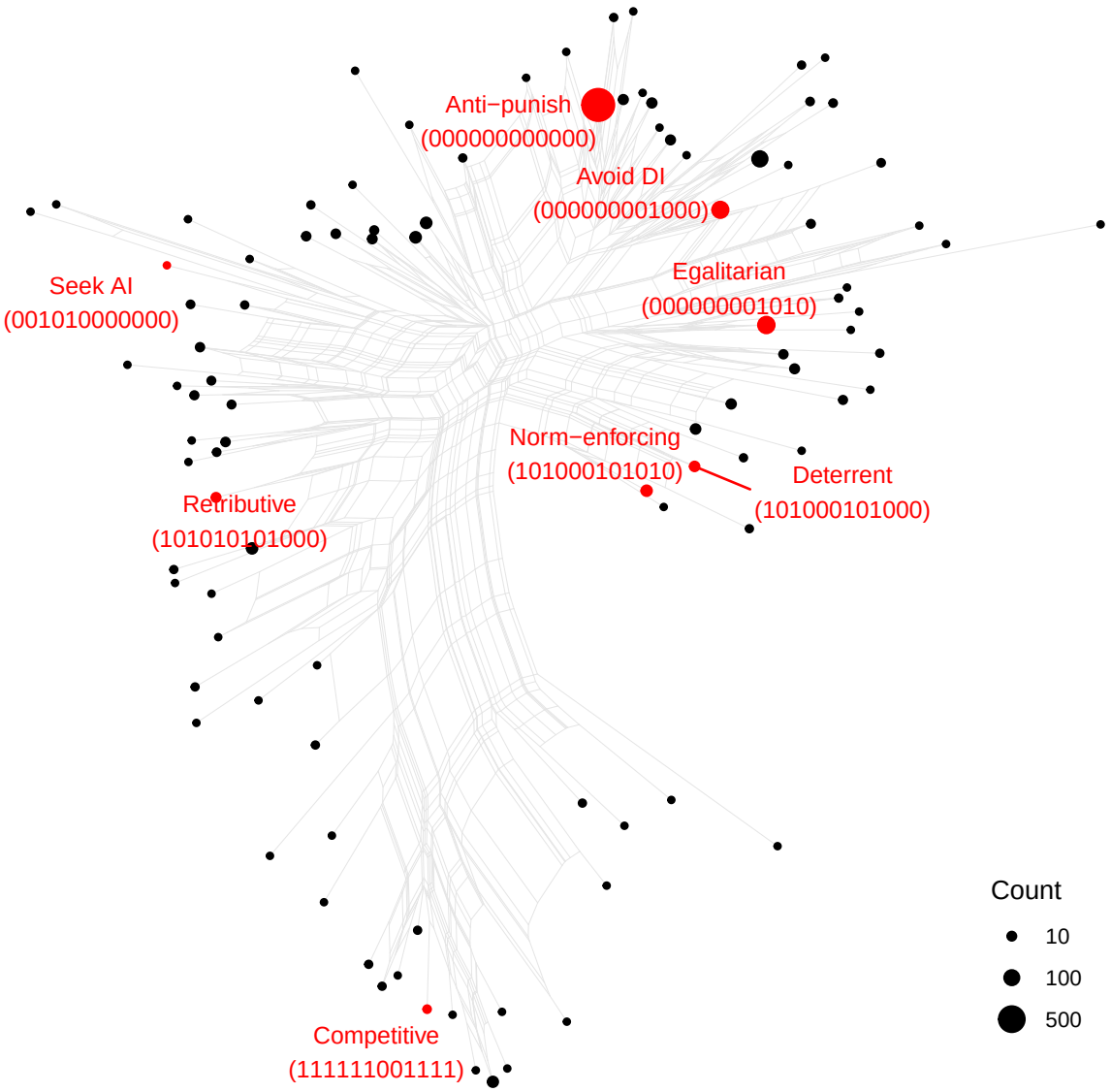
Strategy	N	Proportion
Avoid DI	1594	0.788
Egalitarian	1589	0.785
Anti-punish	1522	0.752
Seek AI	1359	0.671
Deterrent	497	0.246
Norm-enforcing	473	0.234
Retributive	291	0.144
Competitive	79	0.039
Antisocial	7	0.003

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	Proportion
000000000000	*Anti-punish strategy (exact)	878	0.434
000000001010	*Egalitarian strategy (exact)	136	0.067
000000001000	*Avoid DI strategy (exact)	129	0.064
000000000010	Punish when take in 3PP game	106	0.052
001000001010	Punish when take in No DI 2, DI, and 3PP games	26	0.013
001000001000	Punish when take in No DI 2 and DI games	25	0.012
101000101010	*Norm-enforcing strategy (exact)	24	0.012
101010101010	Punish when take in all games	24	0.012
111111111111	Punish when take AND nothing in all games	22	0.011
101000001010	Punish when take in No DI 1, No DI 2, DI, and 3PP games	15	0.007
101000101000	*Deterrent strategy (exact)	15	0.007

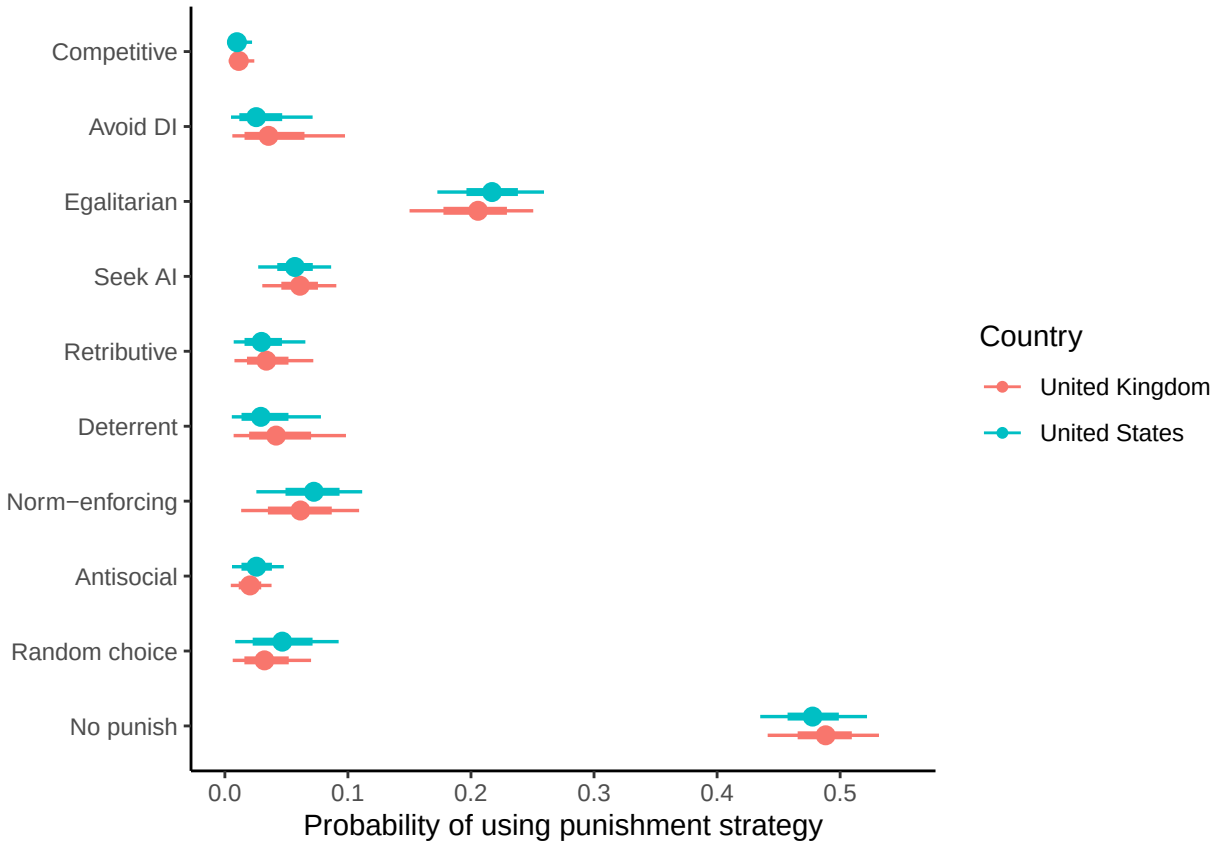
Pattern	Explanation	N	Proportion
000000100000	Punish when take in No DI 4 game	13	0.006
101000001000	Punish when take in No DI 1, No DI 2, and DI games	13	0.006
100000000000	Punish when take in No DI 1 game	12	0.006
001000000000	Punish when take in No DI 2 game	11	0.005
100000001000	Punish when take in No DI 1 and DI games	11	0.005
101010101000	*Retributive strategy (exact)	11	0.005
001000101010	Punish when take in No DI 2, No DI 4, DI, and 3PP games	10	0.005
000000101000	Punish when take in DI and 3PP games	8	0.004
000000101010	Punish when take in No DI 4, DI, and 3PP games	8	0.004
101010001010	Punish when take in all games except No DI 4	8	0.004
001010101000	Punish when take in No DI 2, No DI 3, No DI 4, and DI games	7	0.003
100000001010	Punish when take in No DI 1, DI, and 3PP games	7	0.003
000010001010	Punish when take in No DI 3, DI, and 3PP games	6	0.003
001000101000	Punish when take in No DI 2, No DI 4, and DI games	6	0.003

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 two 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.

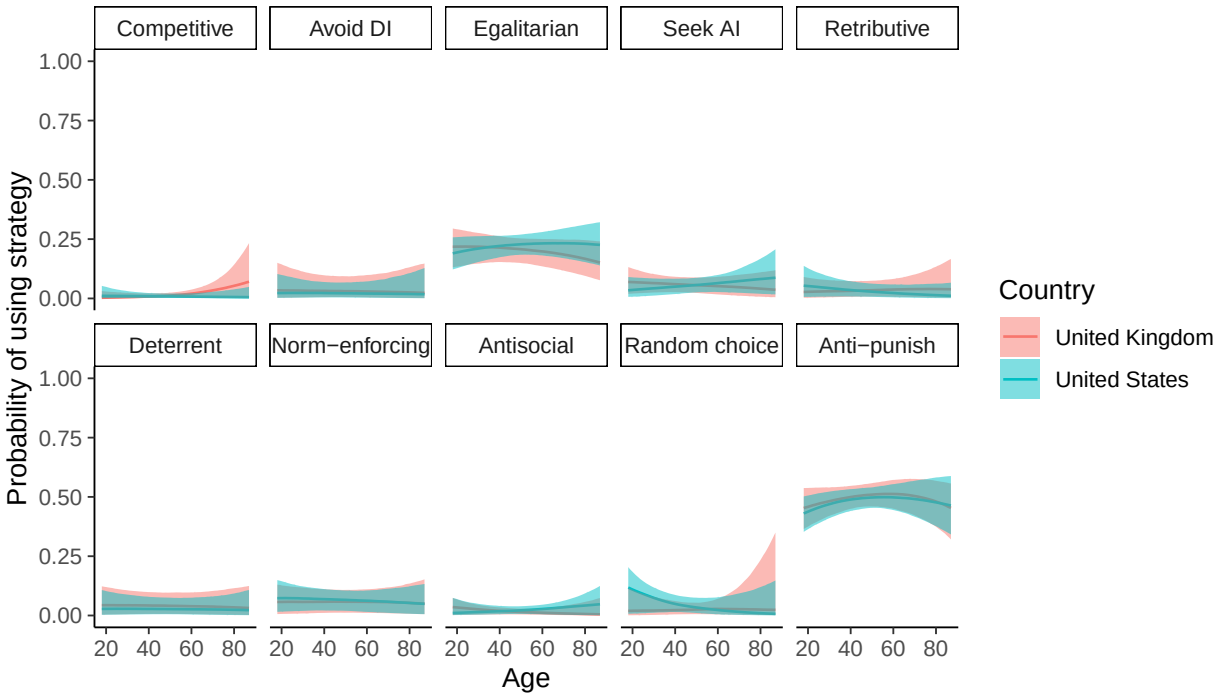
Why is egalitarian the winner in this model, considering that it wasn't an obvious winner in the raw counts? One explanation is that many participants punished only in the third-party punishment game (see table above). The model does not include this pattern of behaviour as an explicit strategy. This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

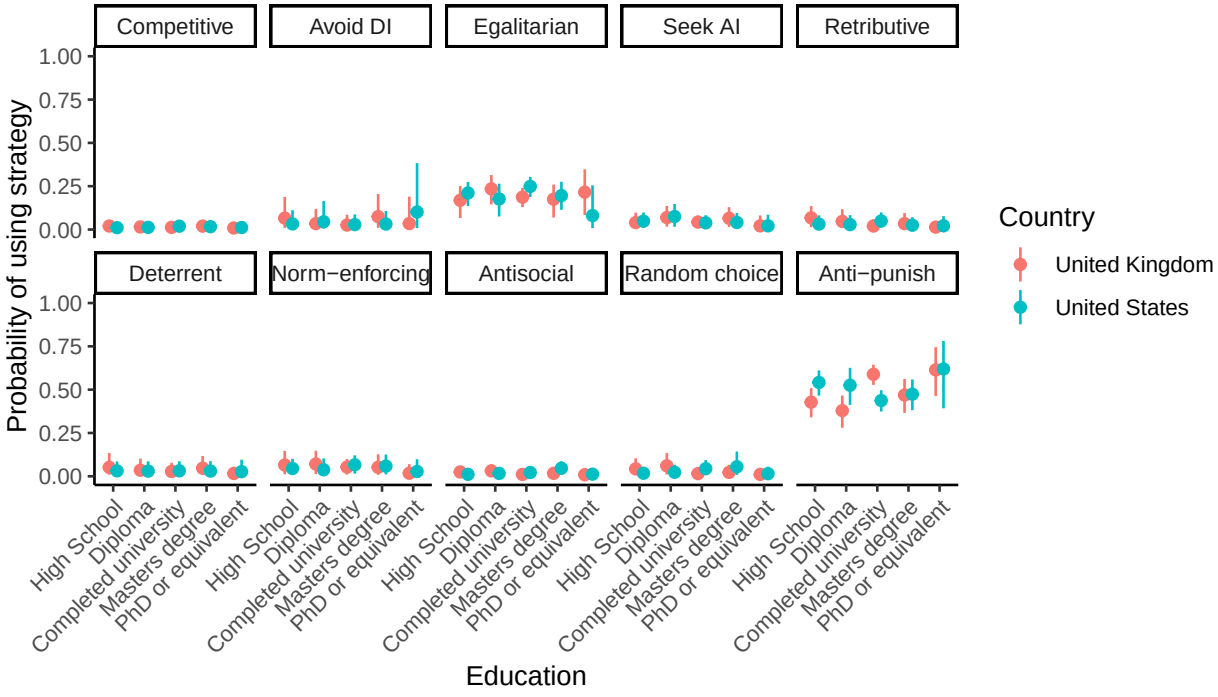
Predicting strategy usage

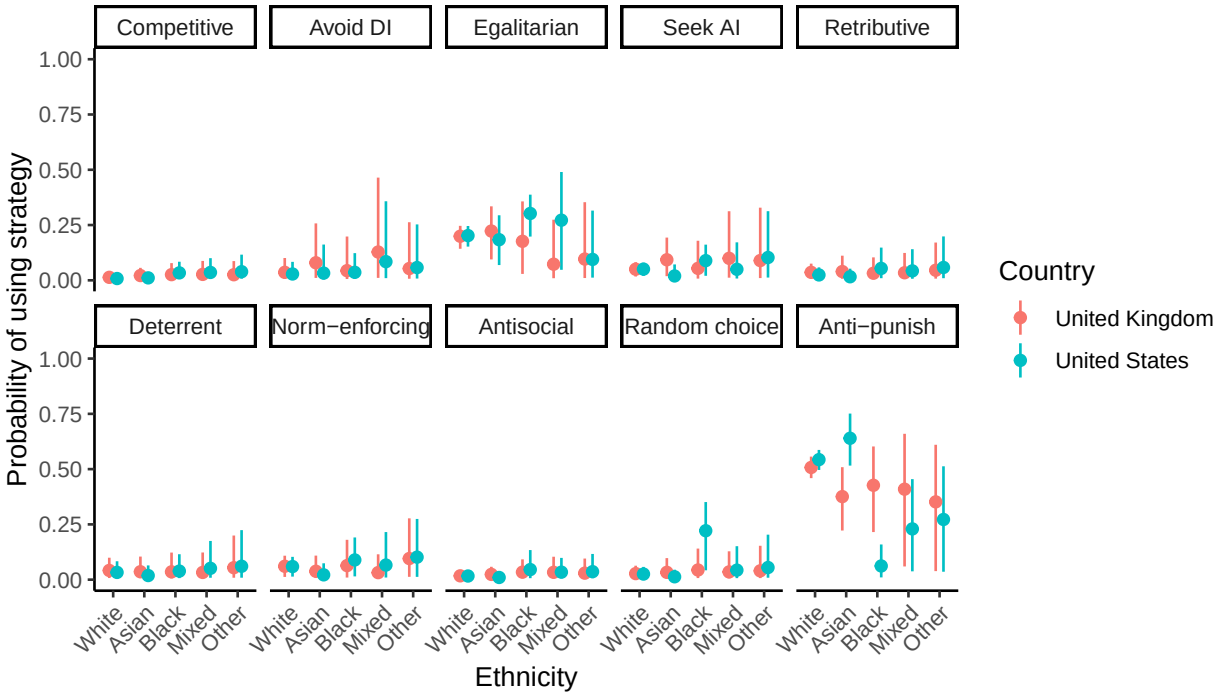
We then fit a series of models predicting the usage of different strategies from a variety of predictors. In each plot that follows, we show the posterior predictions from a model that includes a single variable as a predictor of all ten strategies and both countries simultaneously. The plots show the median regression lines with shaded 95% credible intervals.

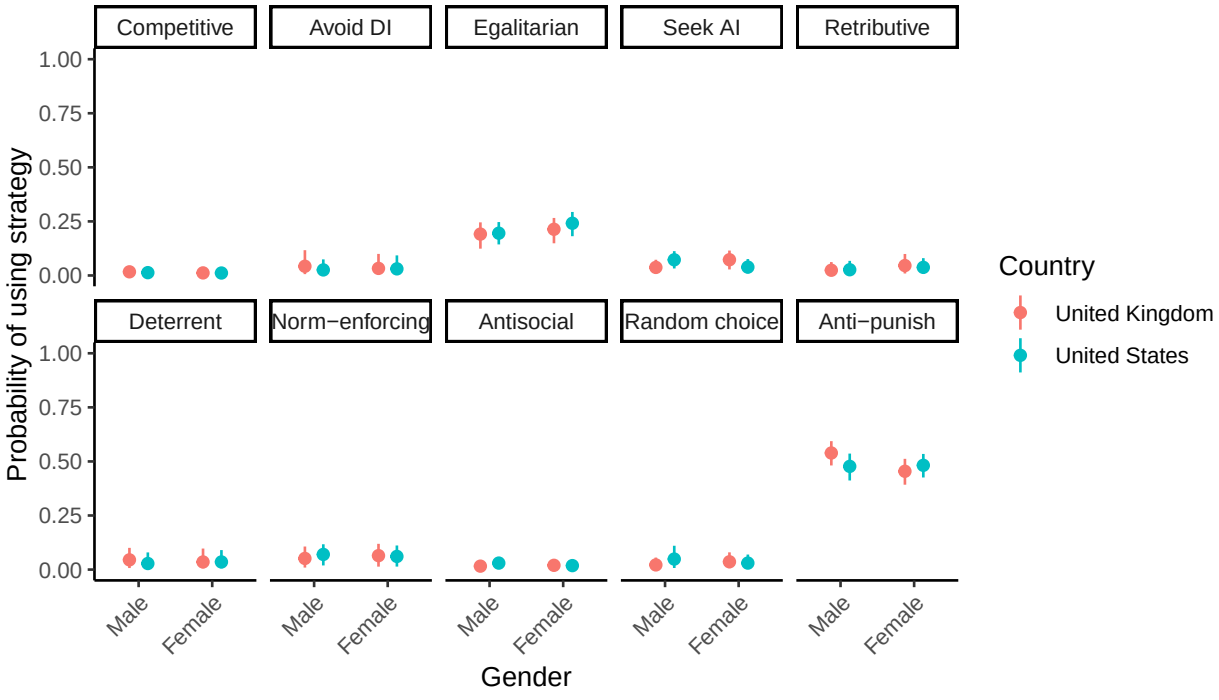
Demographics

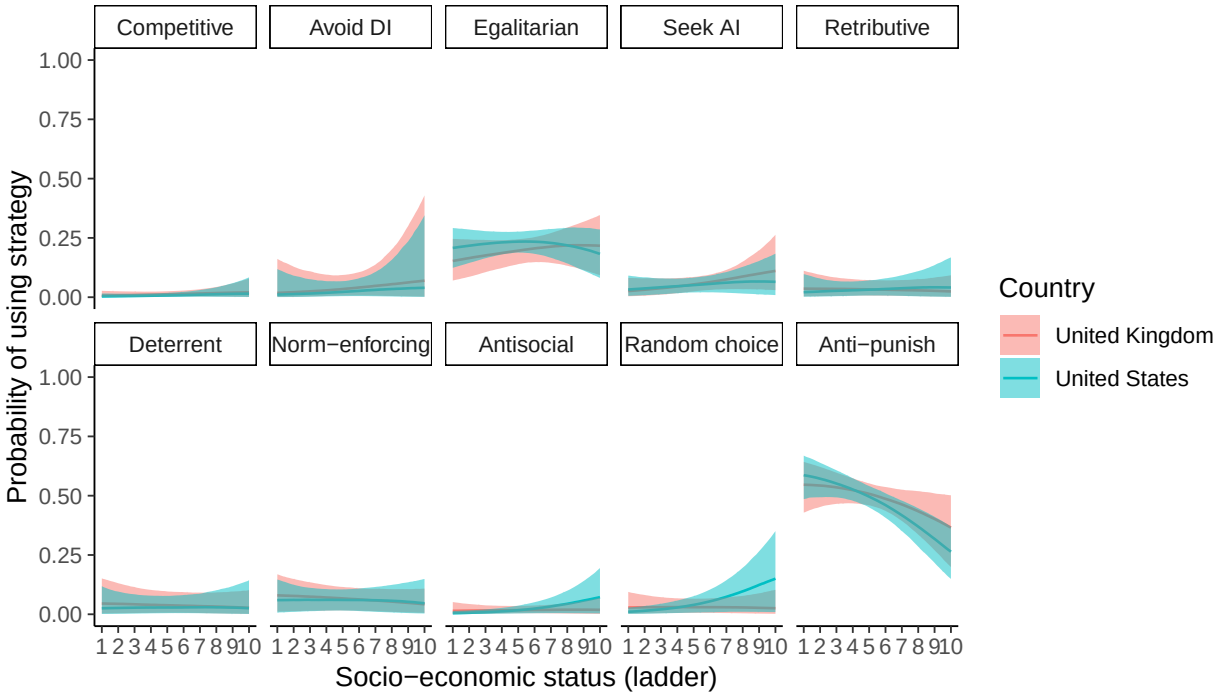
Age, gender, and socio-economic status were unrelated to strategy usage. Higher education tended to predict the anti-punish strategy. In the United States, non-students and white/asian people were more likely to follow the anti-punish strategy.

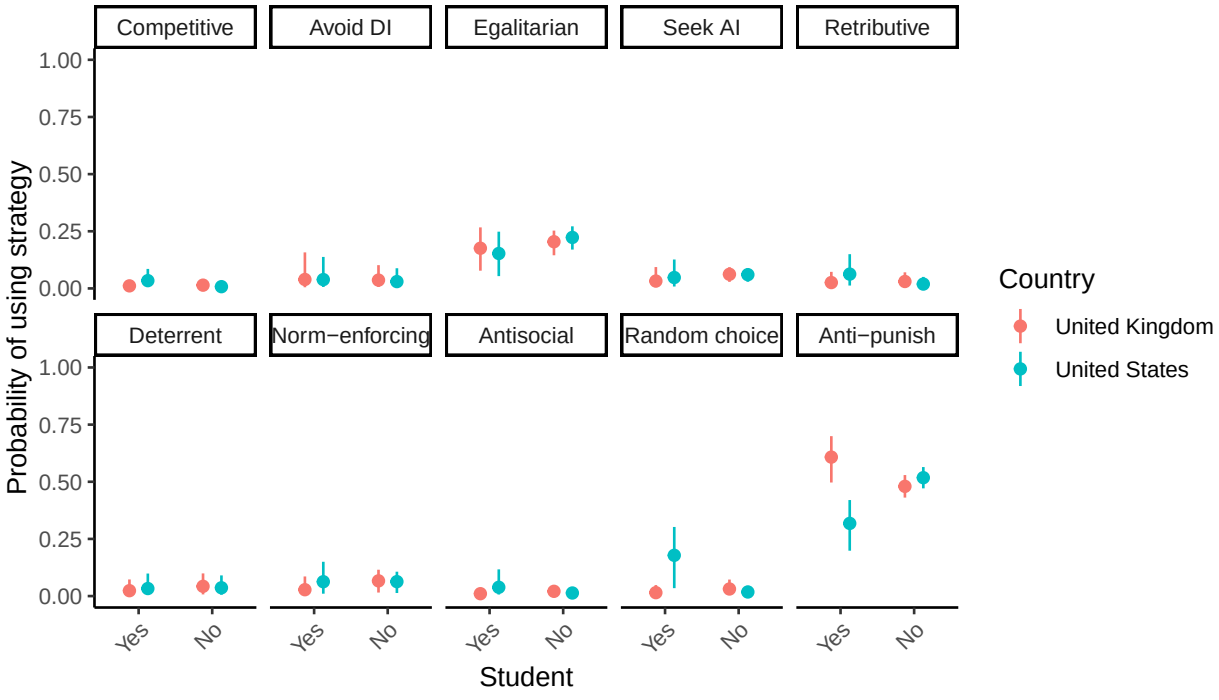

Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
20
40
60
80
Age
Country
United Kingdom
United States


Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
High School Diploma
Completed university
Masters degree
PhD or equivalent
Education
Country
United Kingdom
United States


Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
White
Asian
Black
Mixed
Other
Ethnicity
Country
United Kingdom
United States

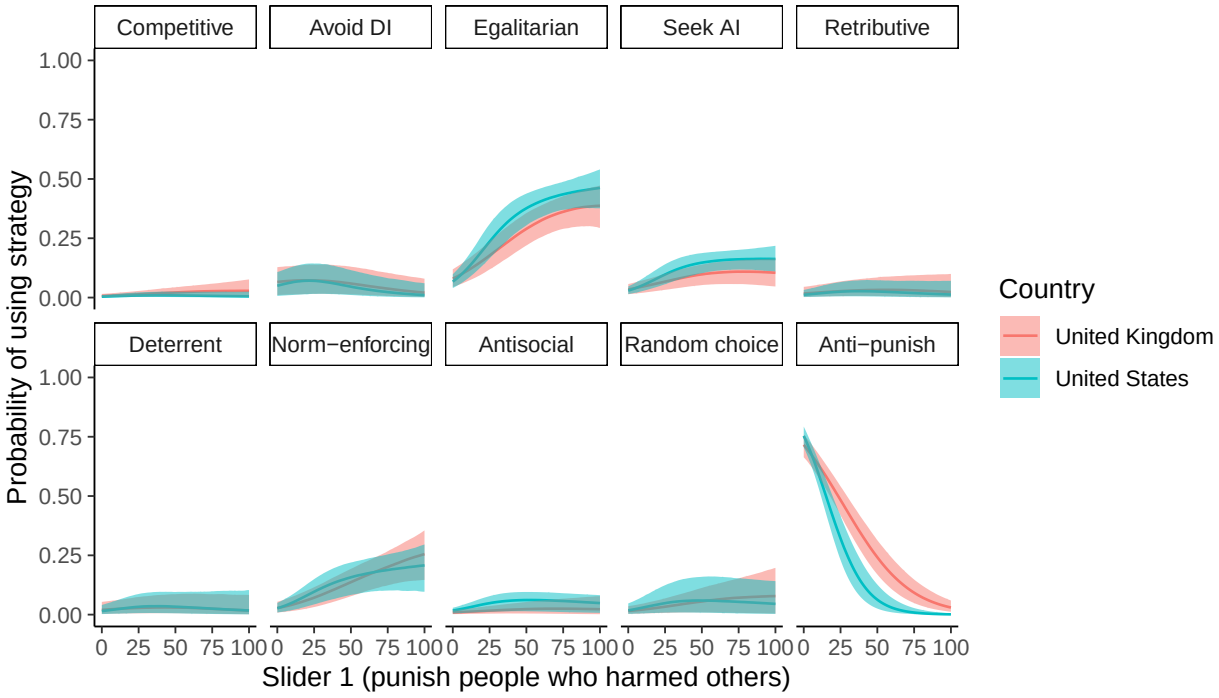

Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
Male
Female
Gender
Country
United Kingdom
United States


Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10
Socio−economic status (ladder)
Country
United Kingdom
United States

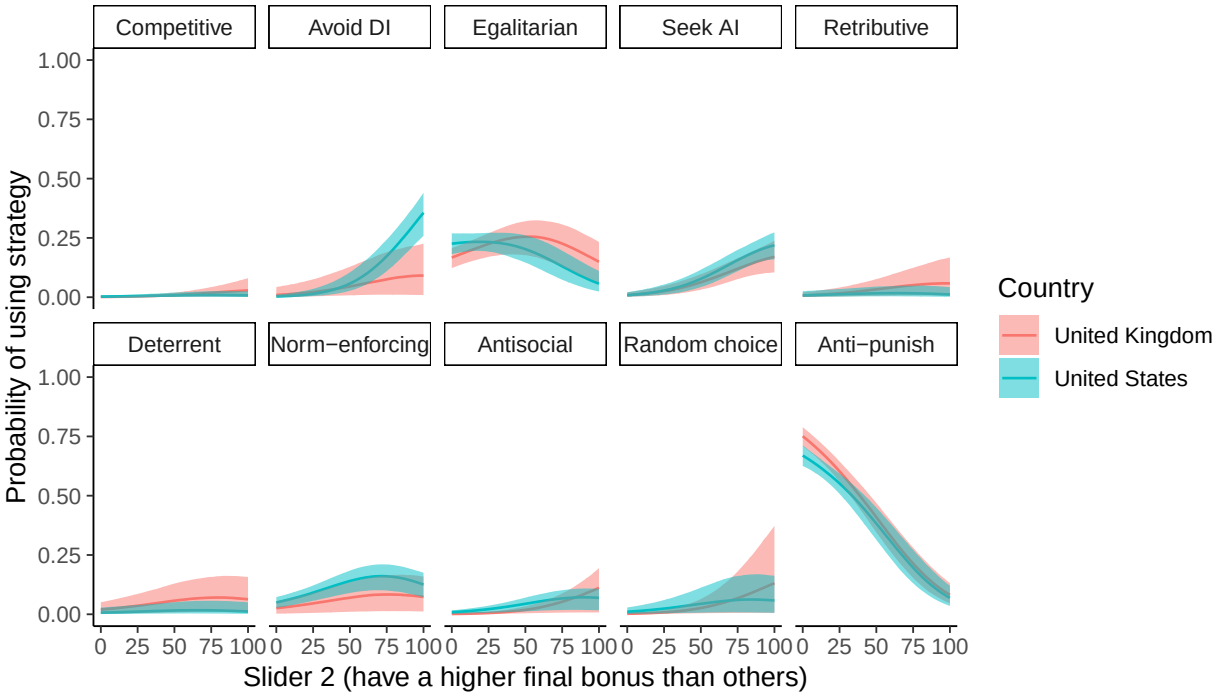

Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
Yes
No
Student
Country
United Kingdom
United States

Self-ratings

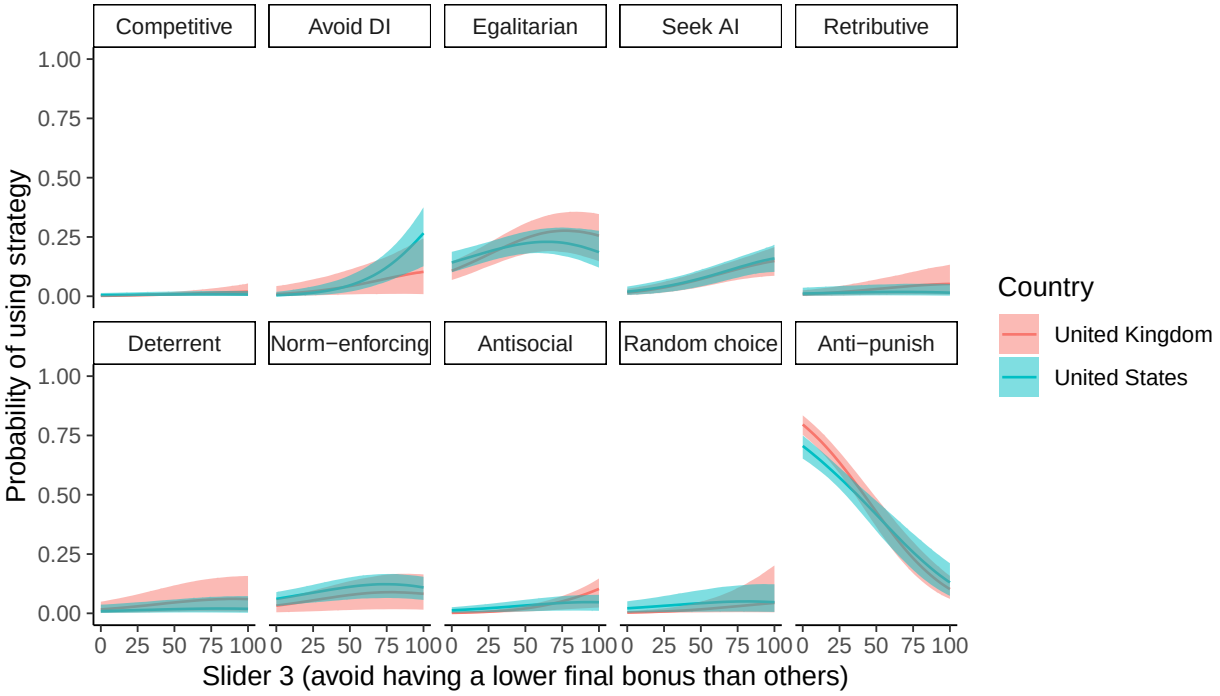
Participants in both countries who stated that they "punished those who harmed others" were more likely to follow egalitarian and norm-enforcing strategies and less likely to follow the anti-punish strategy. In the United States, this slider was additionally positively related to the seek advantageous inequity strategy.



Participants in both countries who stated that they wanted to "have a higher final bonus than others" were more likely to follow the seek advantageous inequity and antisocial strategies and less likely to follow the egalitarian and anti-punish strategies. In the United Kingdom, the slider was additionally related to the random choice strategy. In the United States, this slider was additionally related to the avoid disadvantageous inequity strategy.

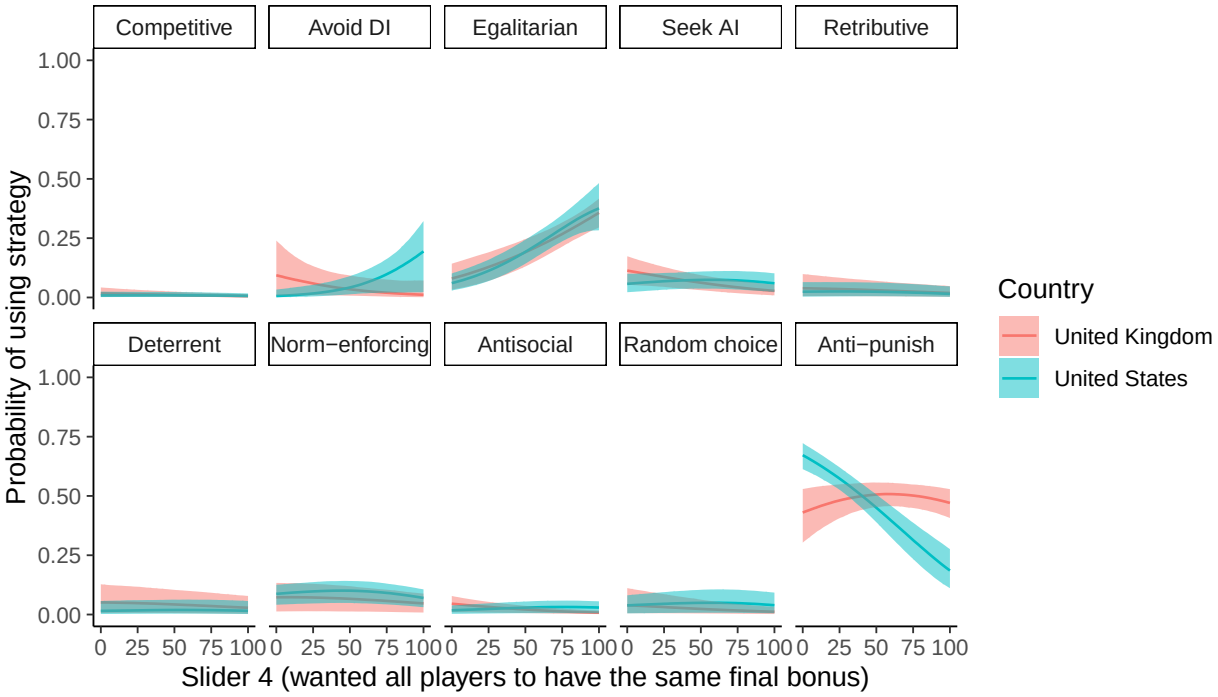


Participants in both countries who stated that they wanted to "avoid having a lower final bonus than others" were less likely to follow the egalitarian and anti-punish strategies. In the United Kingdom, the slider was additionally related to the antisocial strategy. In the United States, this slider was additionally related to the avoid DI and seek AI strategies.



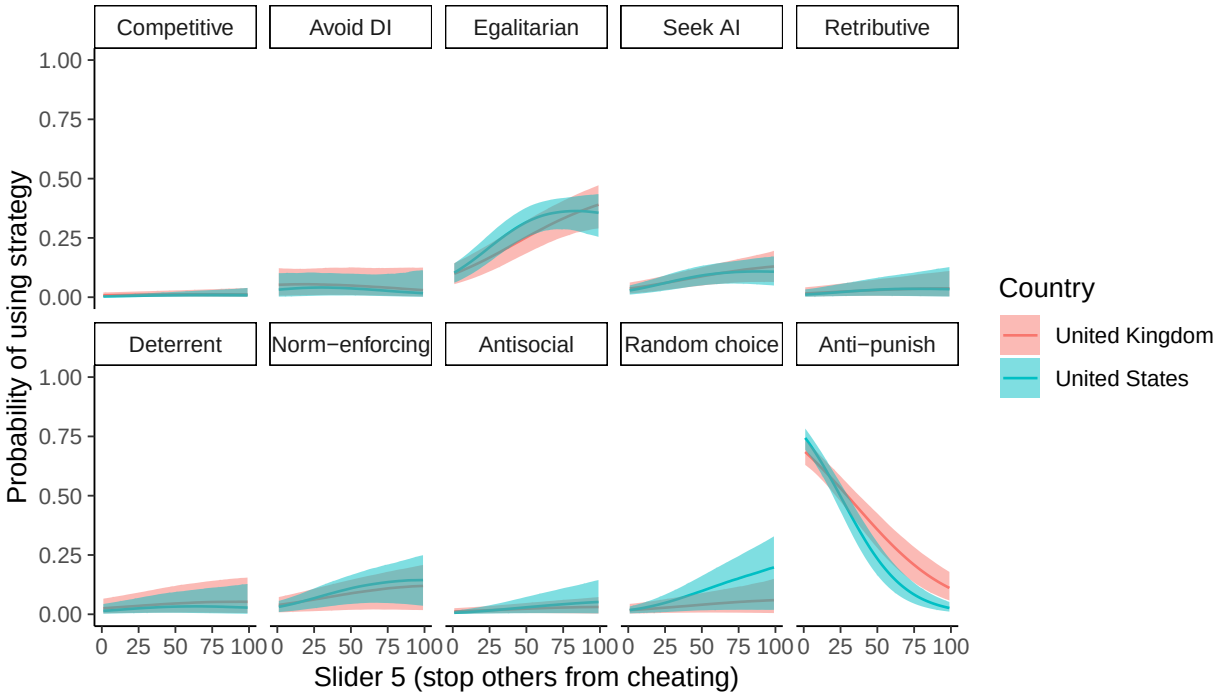
Participants in both countries who stated that they wanted "all players to have the same final bonus" were

more likely to follow the egalitarian strategy. In the United Kingdom, the slider was additionally negatively related to the antisocial strategy and positively related to the anti-punish strategy. In the United States, this slider was additionally positively related to the avoid DI strategy and negatively related to the anti-punish strategy.

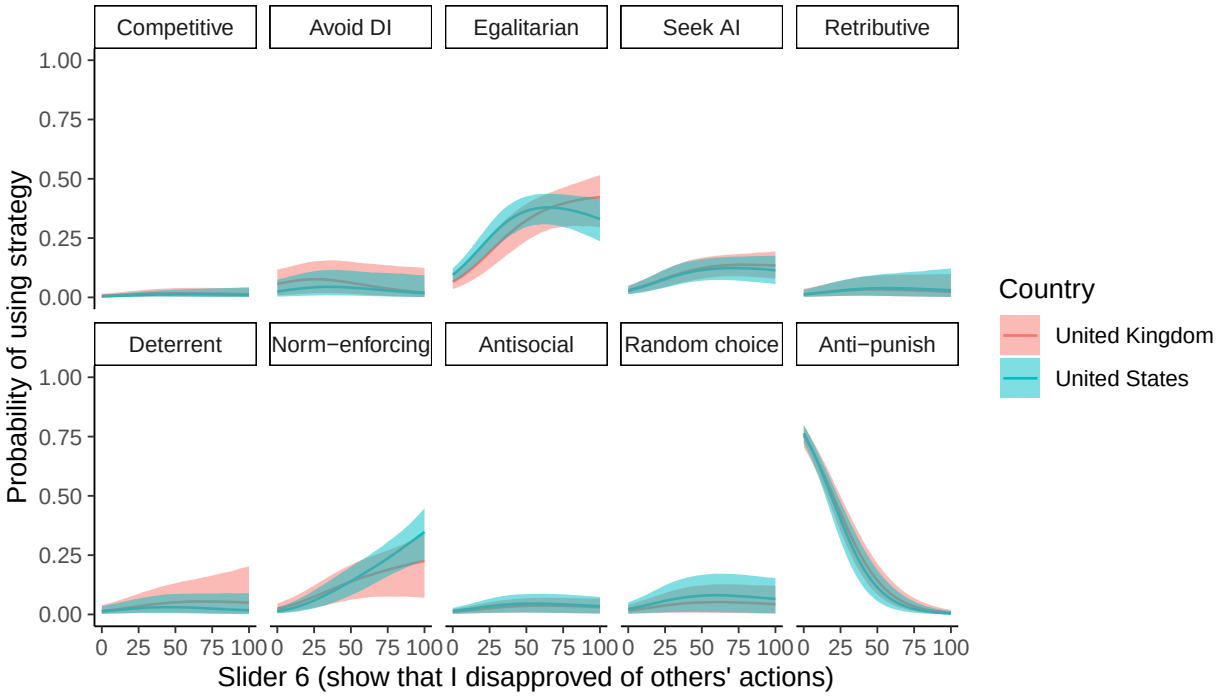


Participants in both countries who stated that they wanted to "stop others from cheating" were less likely to follow the anti-punish strategy. In the United Kingdom, the slider was additionally related to the egalitarian strategy. In the United States, this slider was additionally related to the random choice strategy.

```
## Warning: Removed 4 rows containing missing values (‘geom_line()‘).
```

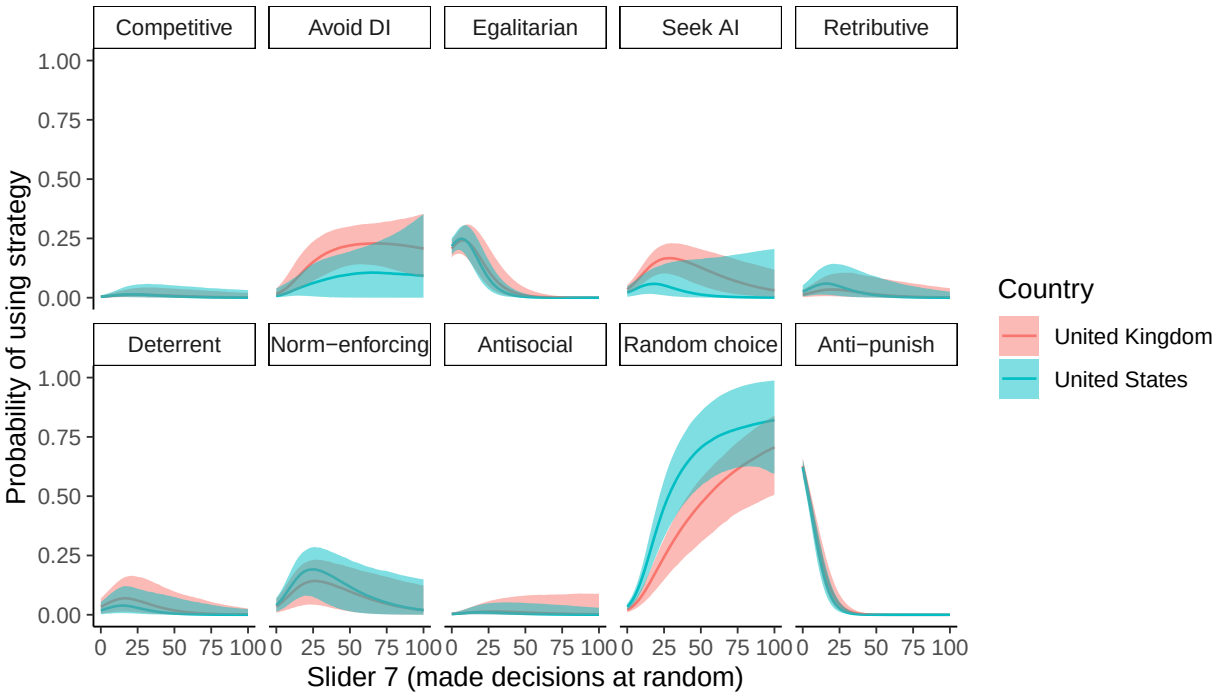


Participants in both countries who stated that they wanted to "show that they disapproved of others' actions" were more likely to follow the egalitarian, seek AI, and norm-enforcing strategies and less likely to follow the anti-punish strategy.

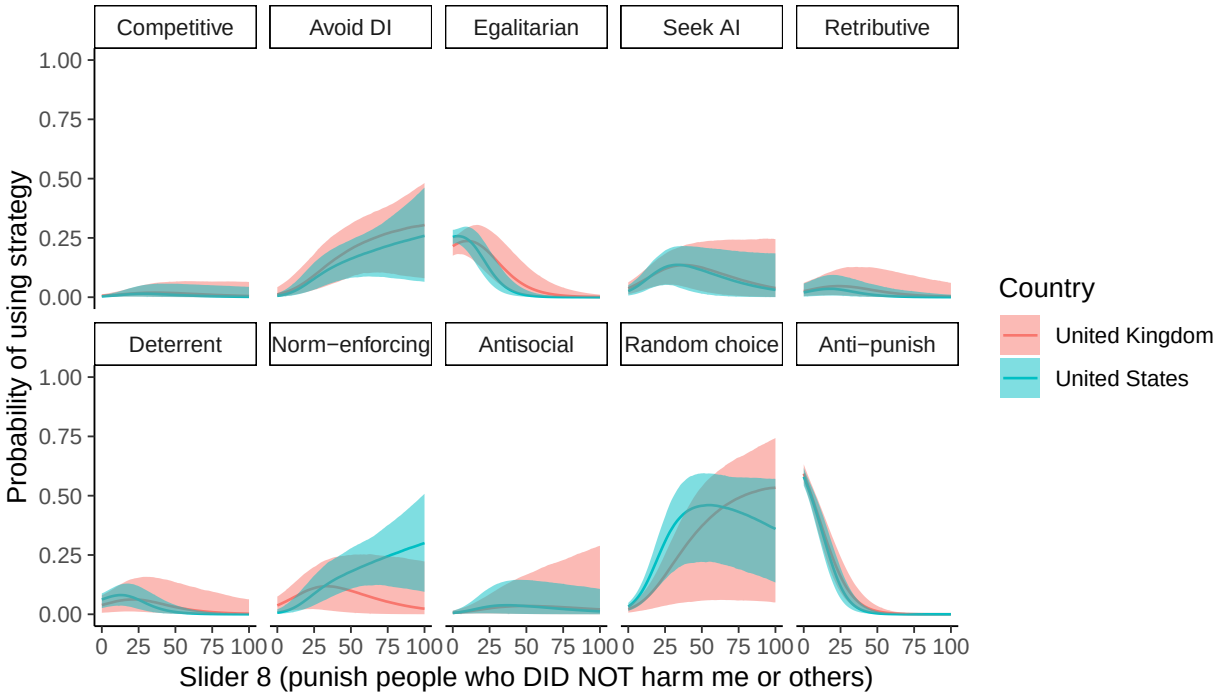


Participants in both countries who stated that they "made decisions at random" were more likely to follow the random choice strategy and less likely to follow the egalitarian and anti-punish strategies. In the United

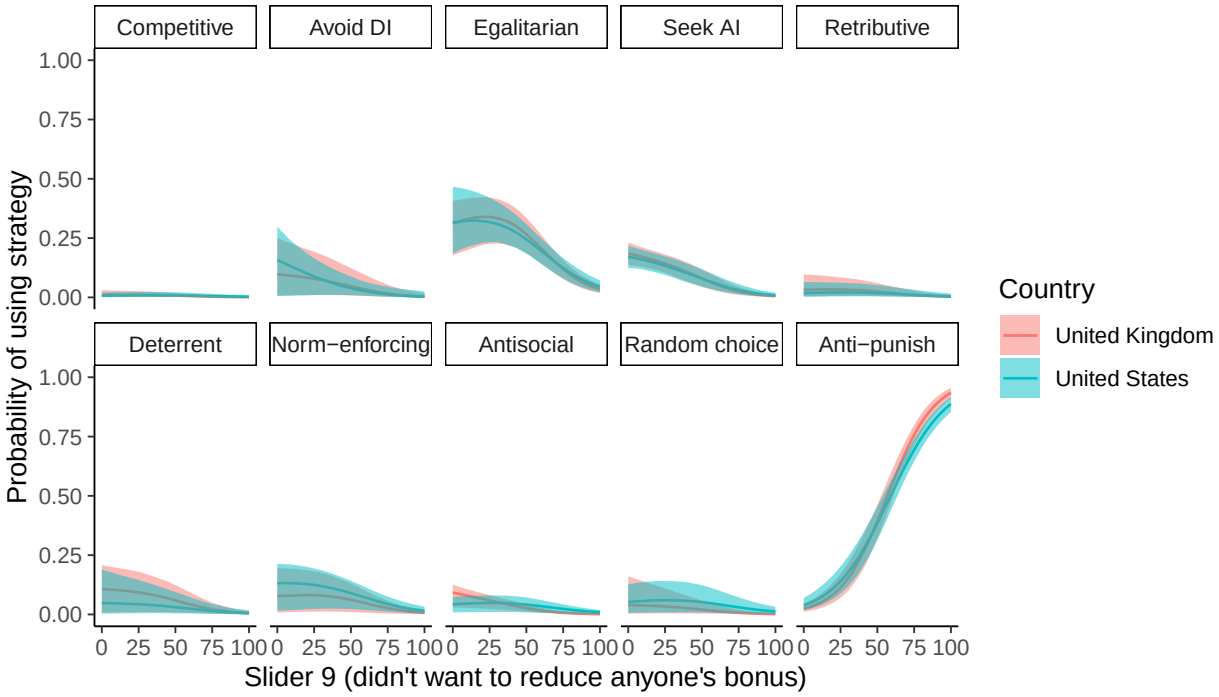
Kingdom, the slider was additionally related to the avoid DI and seek AI strategies. In the United States, this slider was additionally related to the norm-enforcing strategy.



Participants in both countries who stated that they wanted to "punish people who DID NOT harm them or others" were more likely to follow the avoid DI and random choice strategies and less likely to follow the egalitarian and anti-punish strategies. In the United States, this slider was additionally negatively related to the deterrent strategy and positively related to the norm-enforcing strategy.

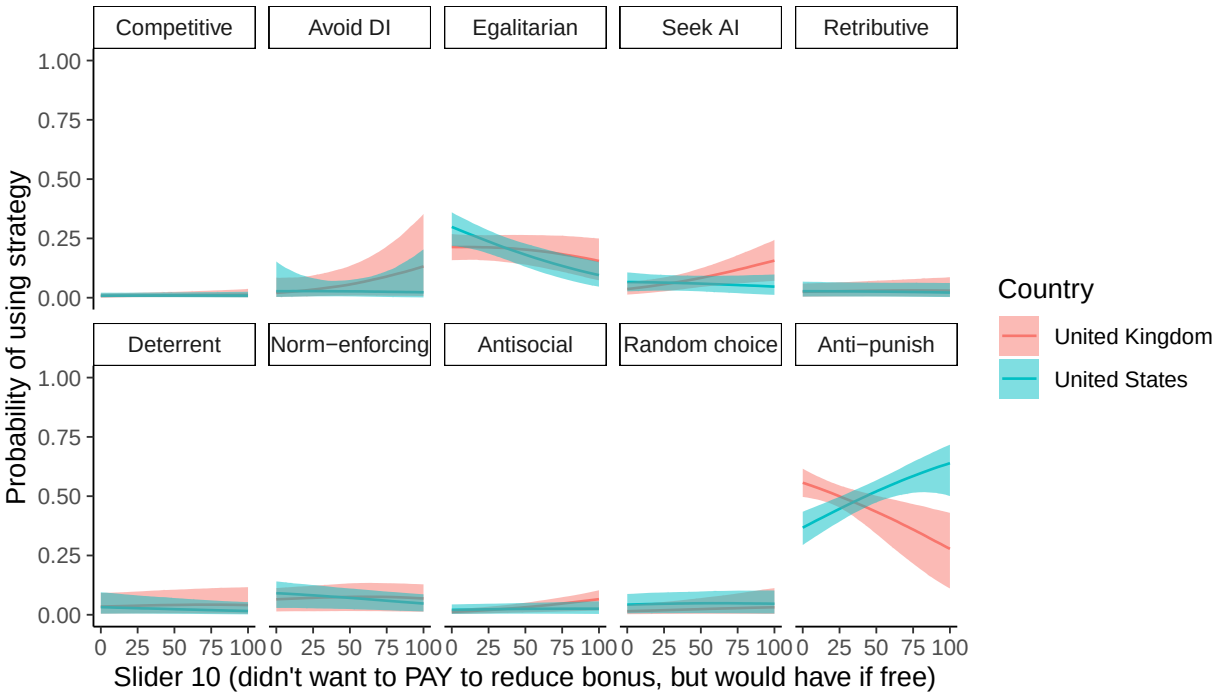


Participants in both countries who stated that they "didn't want to reduce anyone's bonus" were more likely to follow the anti-punish strategy and less likely to follow the seek AI strategy. In the United Kingdom, this slider was additionally negatively related to the antisocial strategy.

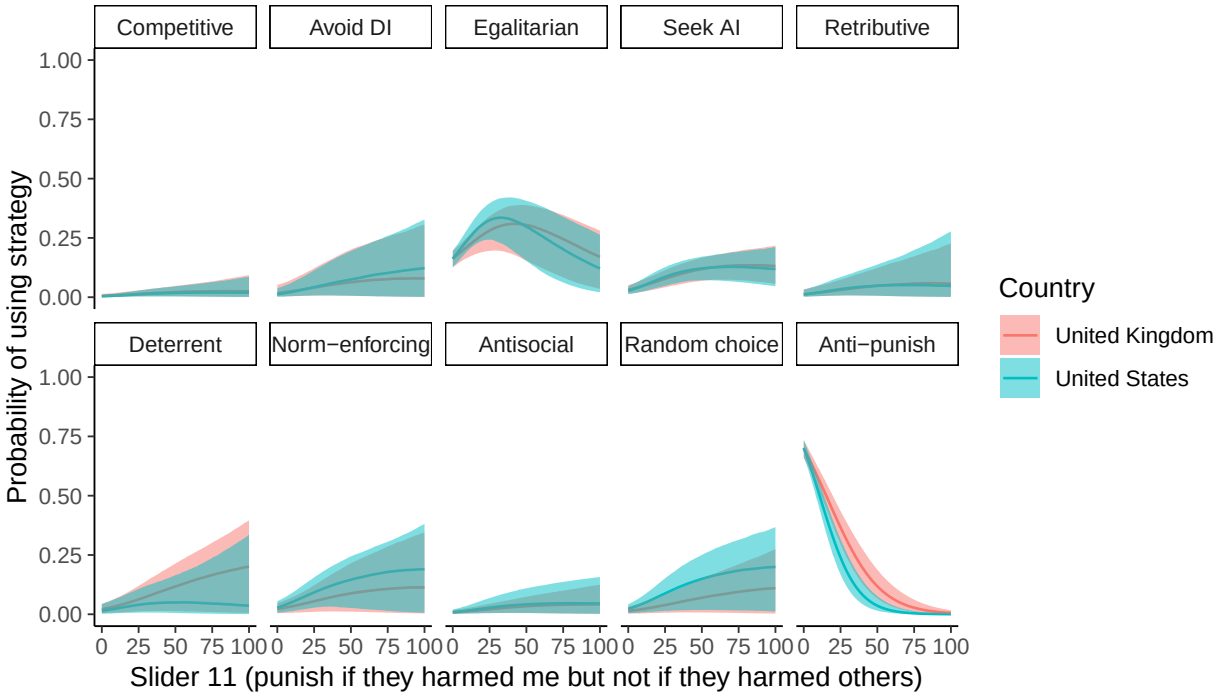


Participants in both countries who stated that they "didn't want to PAY to reduce bonus but would have if free" were less likely to follow the egalitarian strategy. In the United Kingdom, the slider was additionally

positively related to the seek AI and antisocial strategies and negatively related to the anti-punish strategy. In the United States, this slider was additionally positively related to the anti-punish strategy.

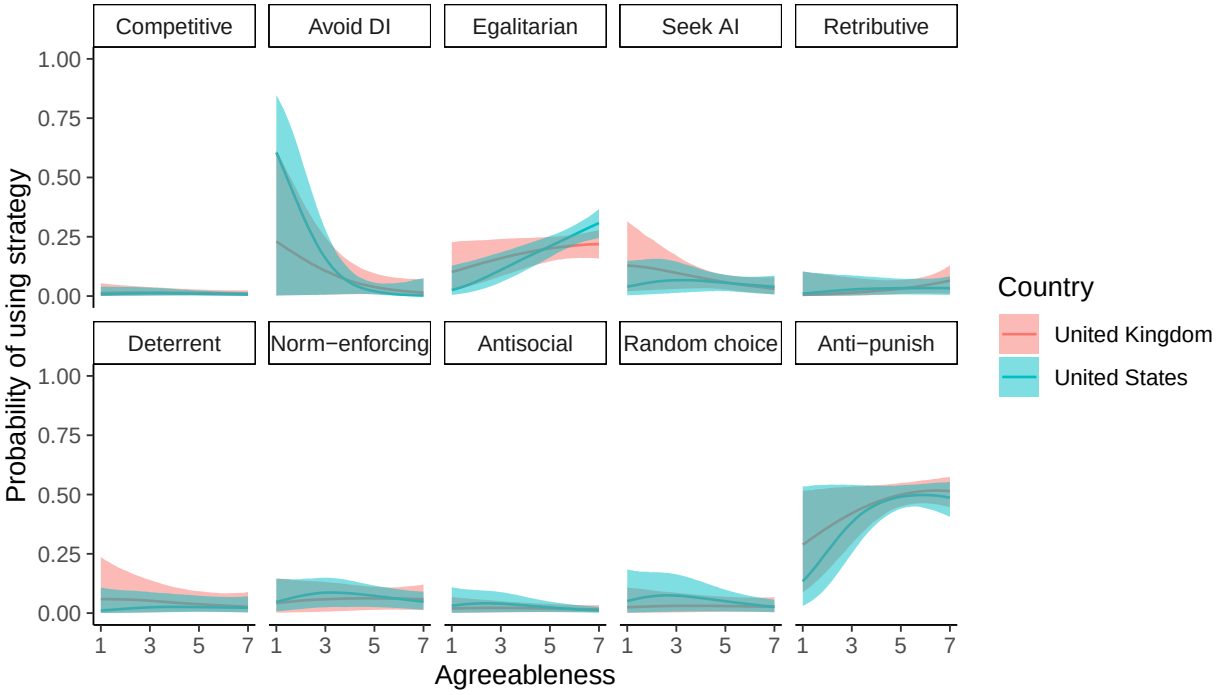


Participants in both countries who stated that they wanted to "punish if players harmed them but not if they harmed others" were less likely to follow the anti-punish strategy. In the United States, this slider was additionally related to the random choice strategy.

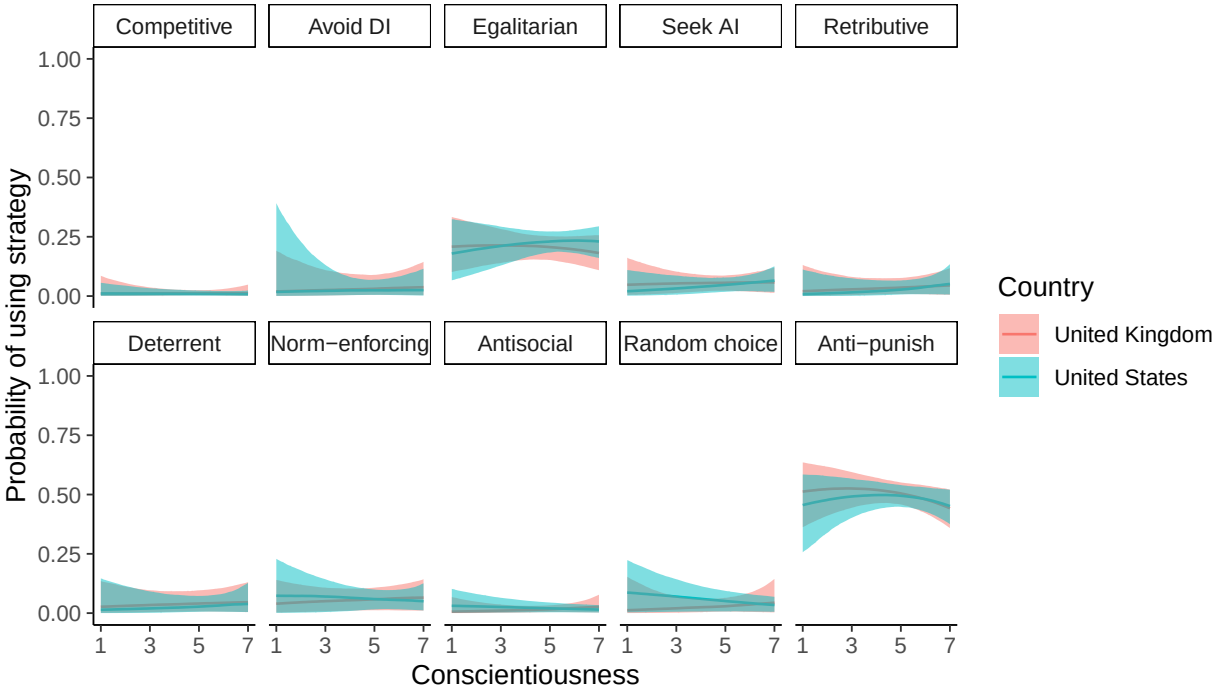

Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 11 (punish if they harmed me but not if they harmed others)
Country
United Kingdom
United States

Big-6 personality

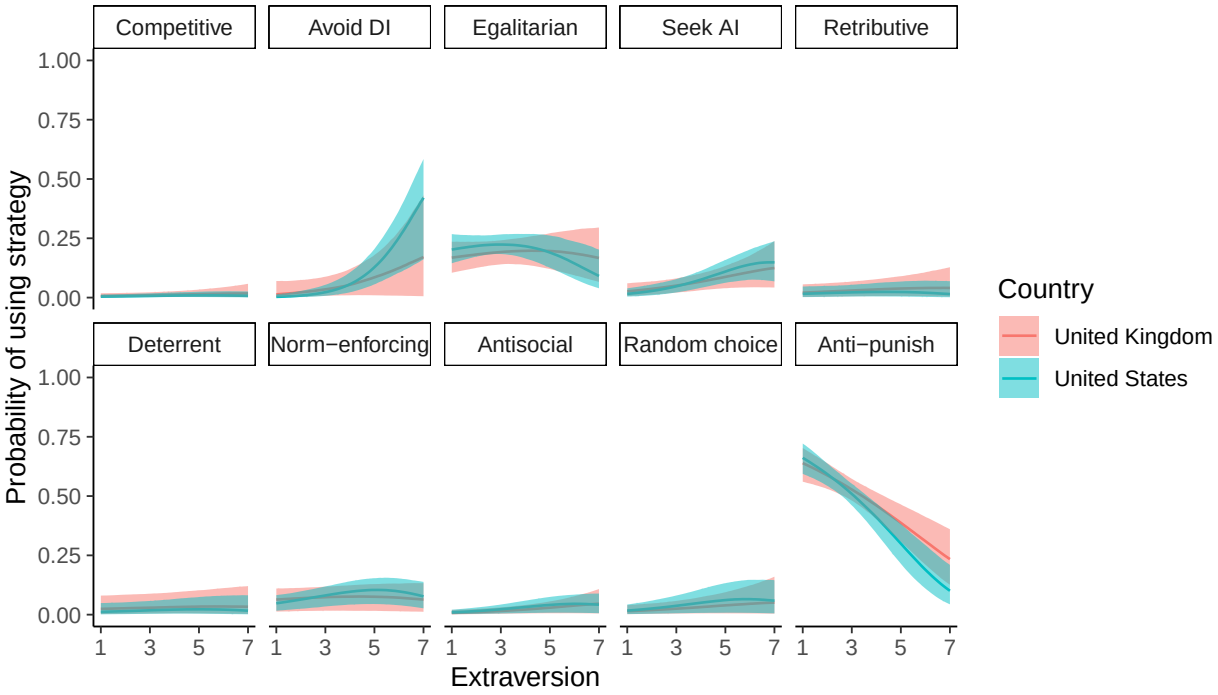
In the United States, agreeableness positively predicts usage of the egalitarian strategy.



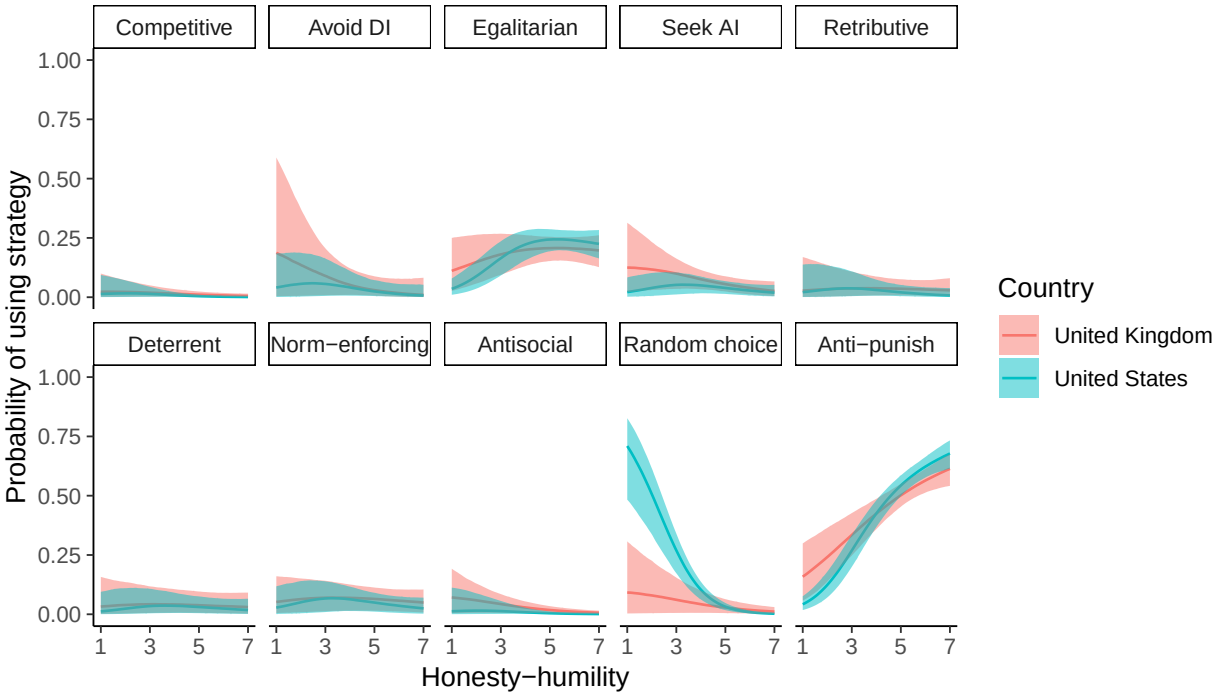
Conscientiousness was unrelated to strategy usage in both countries.



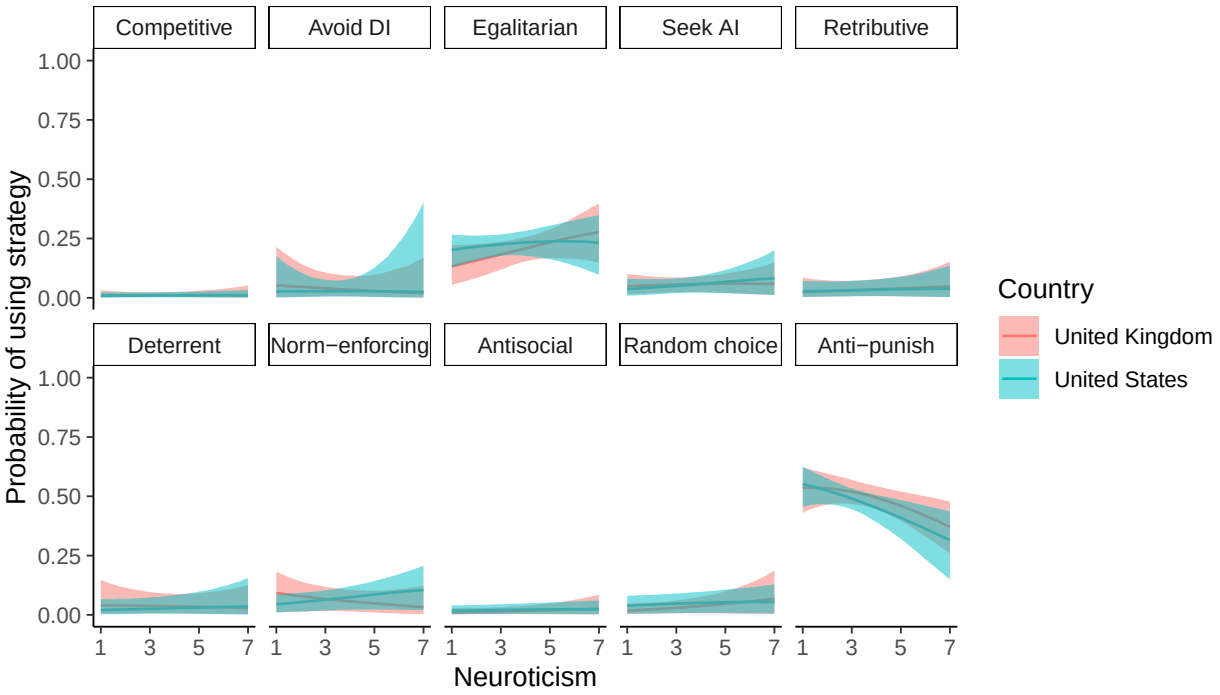
Extraversion negatively predicts usage of the anti-punish strategy in both countries. In the United States, extraversion positively predicts the avoid DI and seek AI strategies, and negatively predicts the egalitarian strategy.



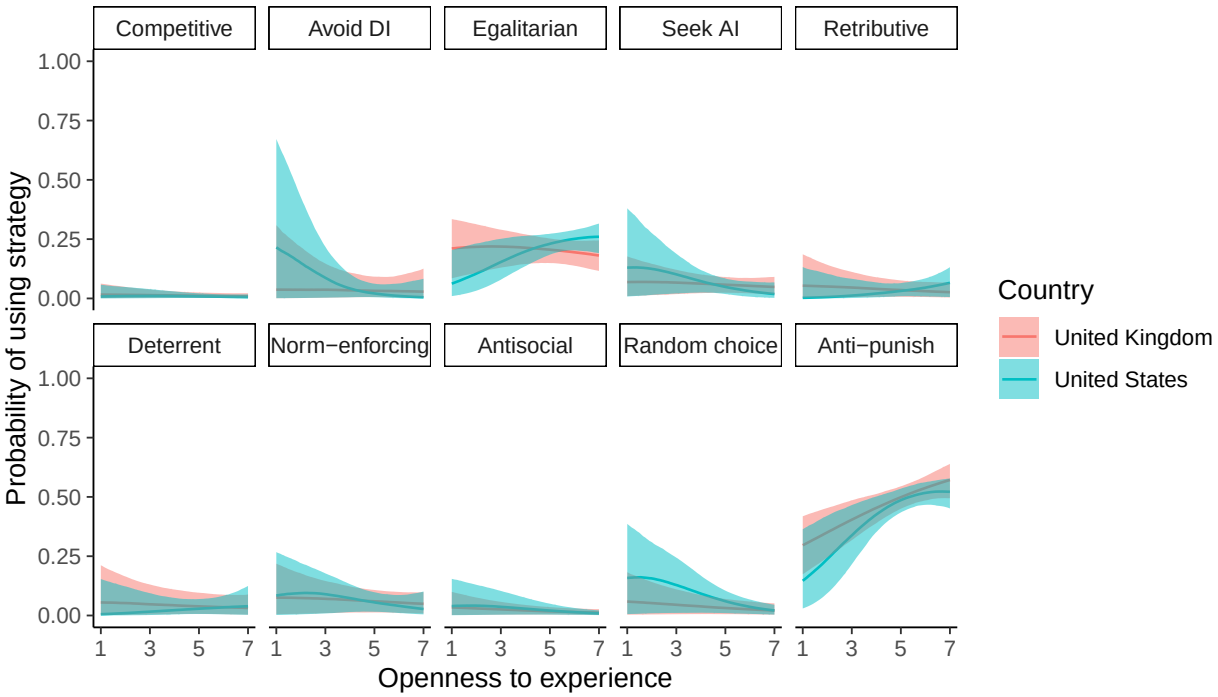
Honesty-humility positively predicts usage of the egalitarian and anti-punish strategies. In the United States, honesty-humility negatively predicts the random choice strategy.



Neuroticism was unrelated to strategy usage in both countries.

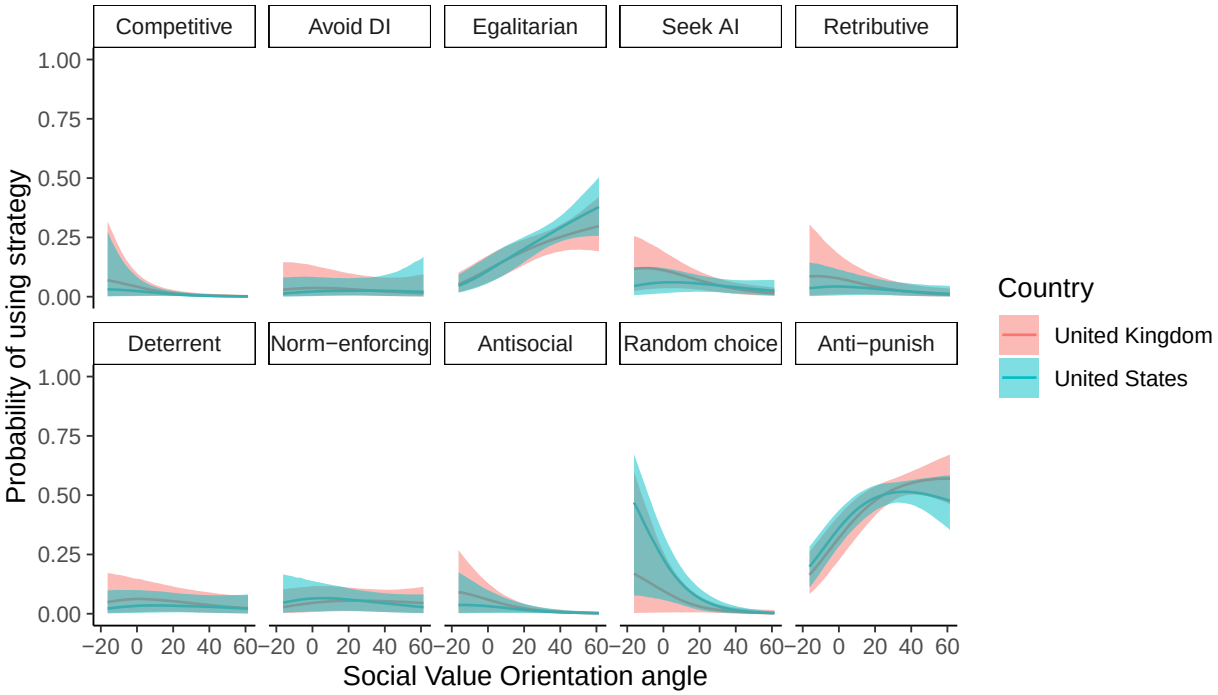


Openness to experience positively predicts usage of the anti-punish strategy in both countries. In the United States, openness positively predicts the egalitarian strategy.


Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
1
3
5
7
Openness to experience
Country
United Kingdom
United States

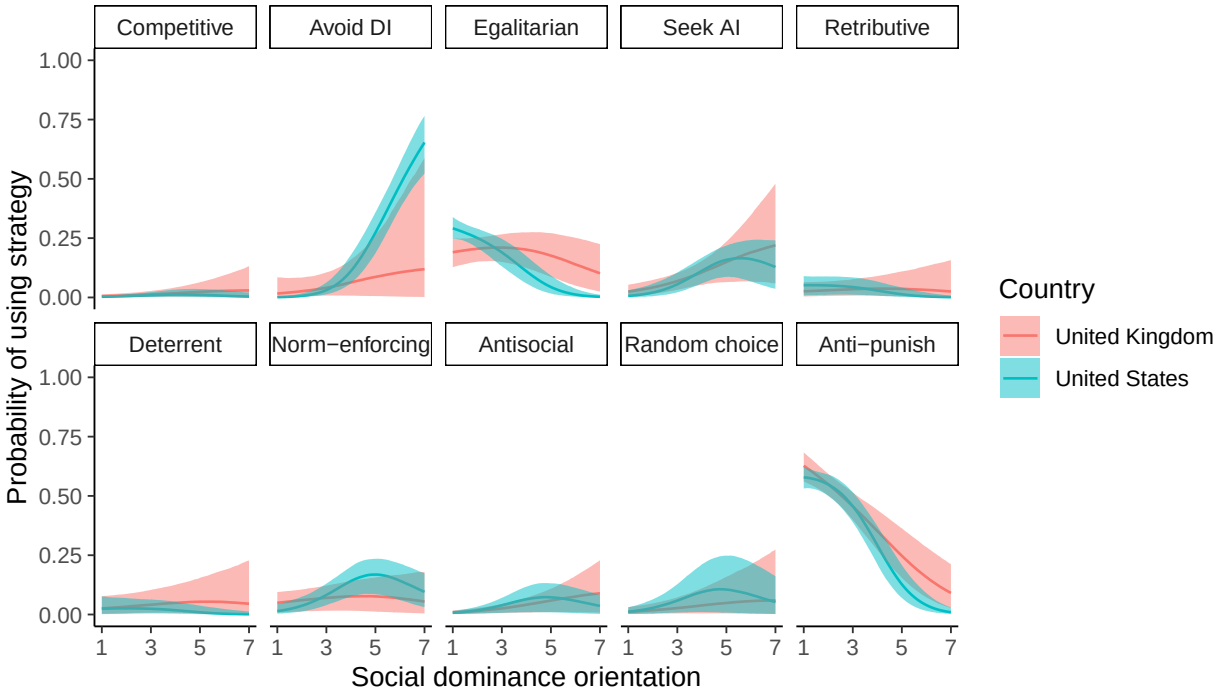
Social value orientation

In both countries, social value orientation angle positively predicted usage of the egalitarian and anti-punish strategies. In the United States, SVO angle also negatively predicted the random choice strategy.



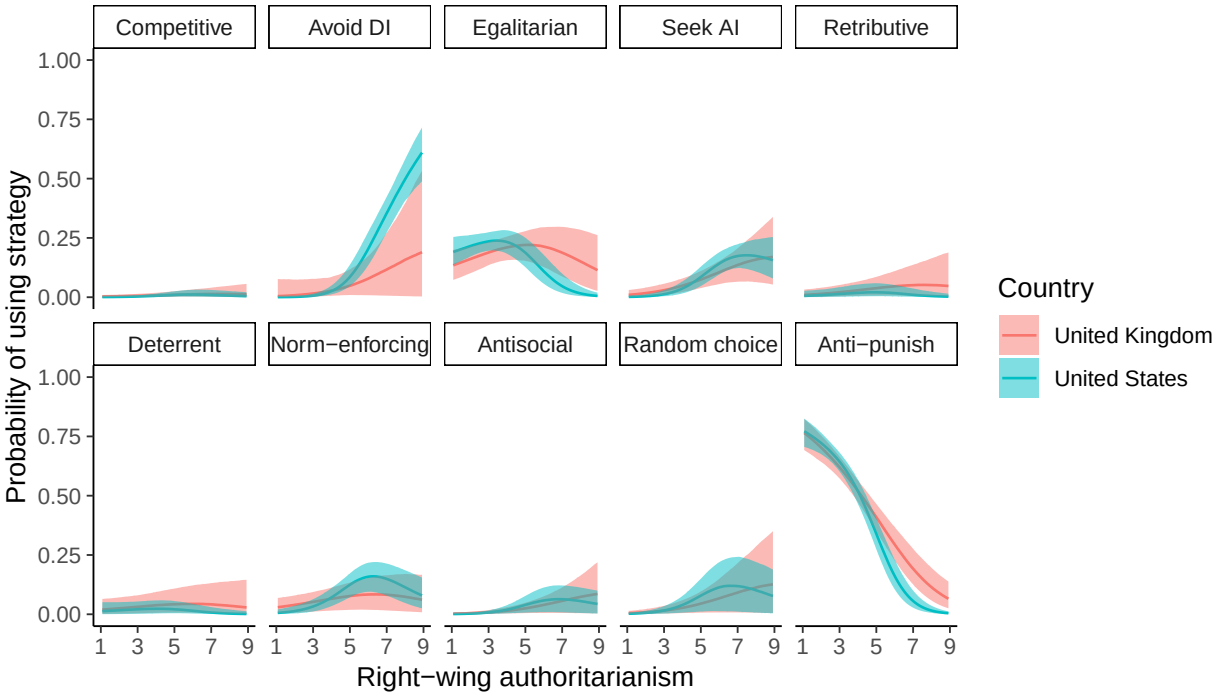
Political views

In both countries, Social Dominance Orientation was positively related to the seek AI and antisocial strategies and negatively related to the egalitarian and anti-punish strategies. In the United States, SDO was also positively related to the avoid DI and norm-enforcing strategies and negatively related to the retributive and deterrent strategies.

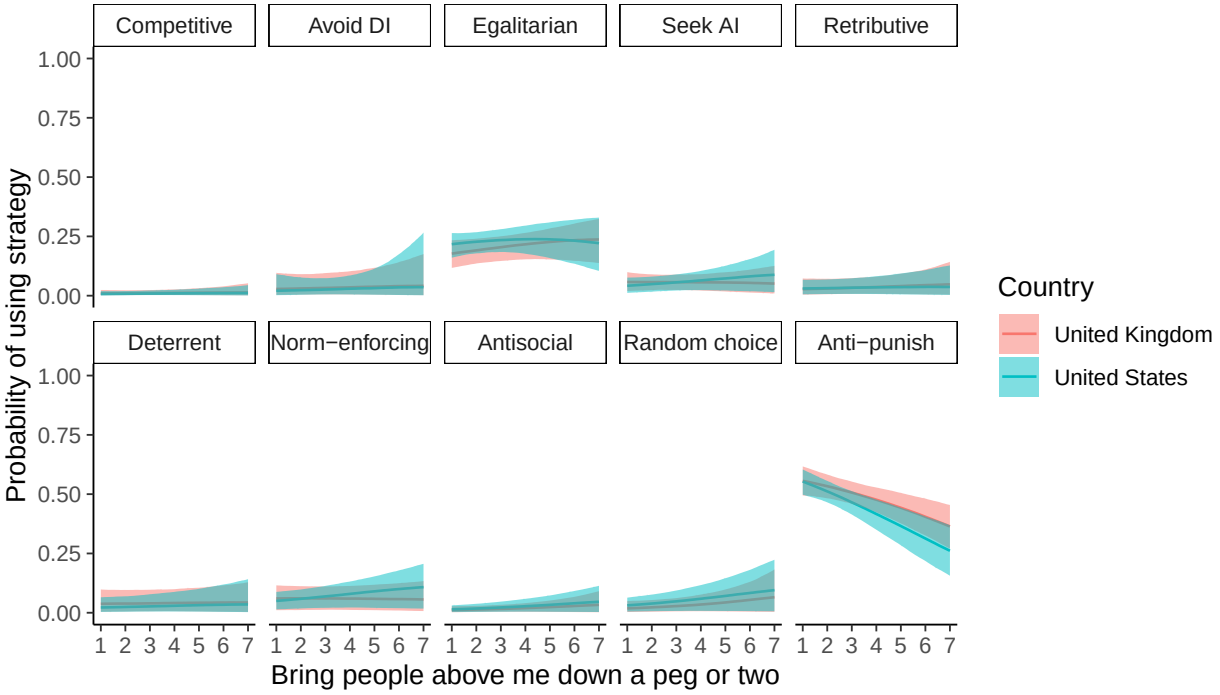


In both countries, Right Wing Authoritarianism was negatively related to the egalitarian and anti-punish strategies. In the United States, RWA was also positively related to the avoid DI, seek AI, antisocial, and random choice strategies.

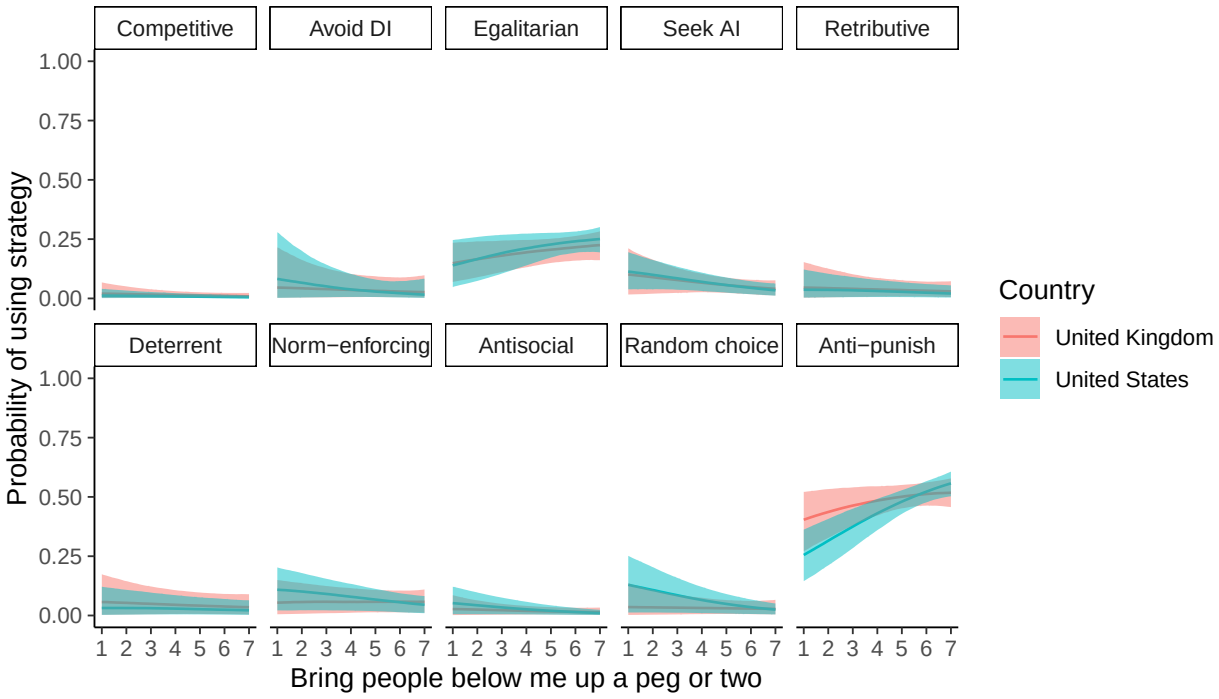
```
## Warning: Removed 4 rows containing missing values (‘geom_line()‘).
```



In the United States, agreement with the motivation to "bring people above me down a peg or two" was negatively related to the anti-punish strategy.



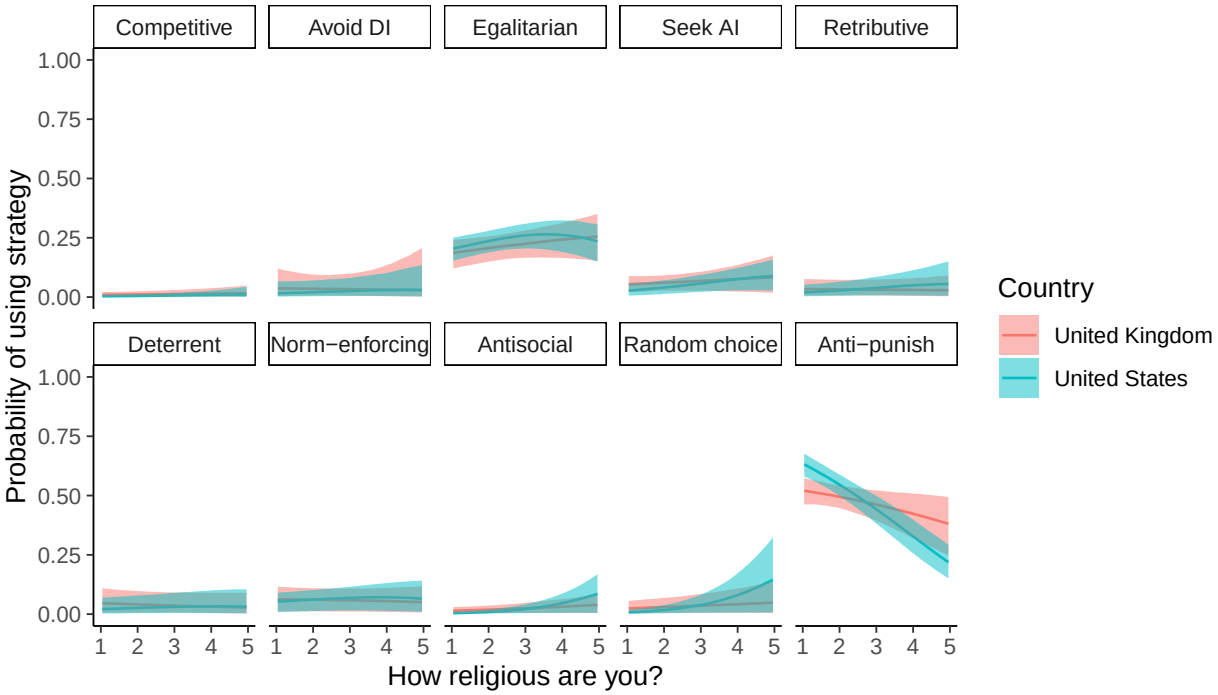
In the United States, agreement with the motivation to "bring people below me up a peg or two" was positively related to the egalitarian and anti-punish strategies.


Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7
Bring people below me up a peg or two
Country
United Kingdom
United States

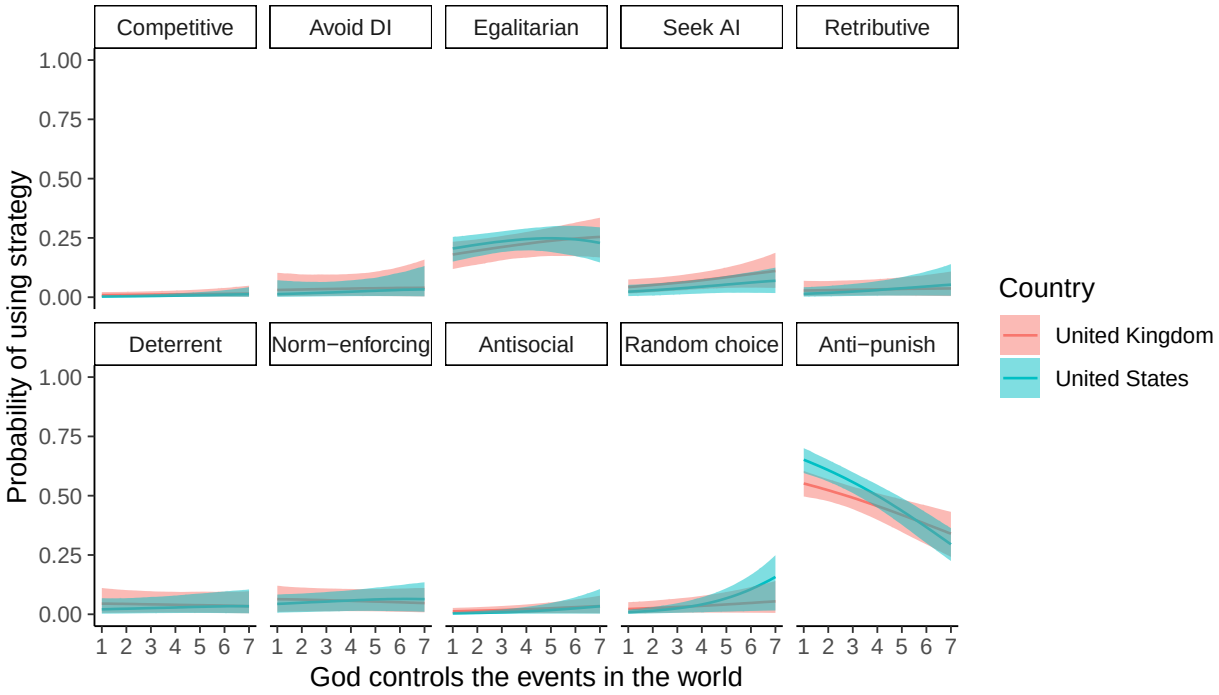
Religion

In the United States, the religiosity item positively predicted the antisocial and random choice strategies and negatively predicted the egalitarian and anti-punish strategies.

```
## Warning: Removed 4 rows containing missing values (`geom_line()`).
```



In both countries, agreeing with the statement "God controls the events in the world" negatively predicts the anti-punish strategy. In the United States, this item also positively predicts the random choice strategy and negatively predicts the egalitarian strategy.


Competitive
Avoid DI
Egalitarian
Seek AI
Retributive
Deterrent
Norm−enforcing
Antisocial
Random choice
Anti−punish
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7
God controls the events in the world
Country
United Kingdom
United States