Mapping out the punishment strategy space

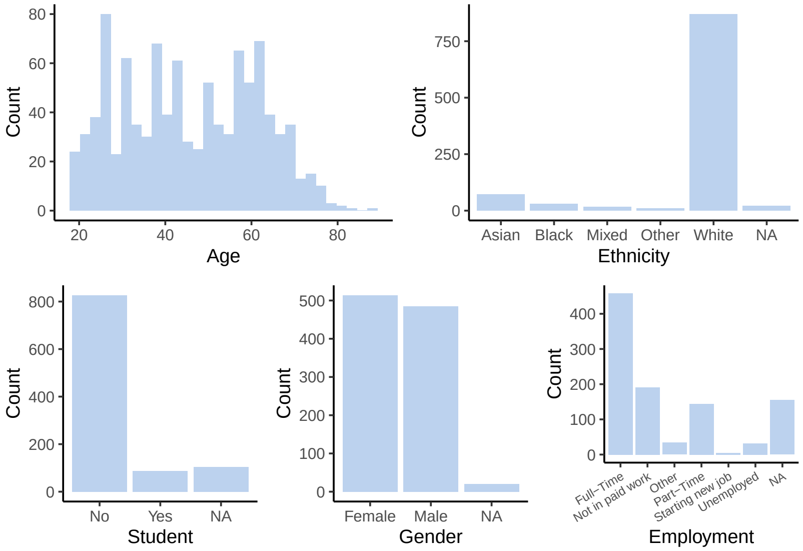
Scott Claessens

2022-12-12

This document outlines the initial data exploration and analyses for Study 1 of our project "Mapping out the punishment strategy space".

Sample

After cleaning the data, we have data for 1019 participants from Prolific. This sample is representative of the United Kingdom.



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.

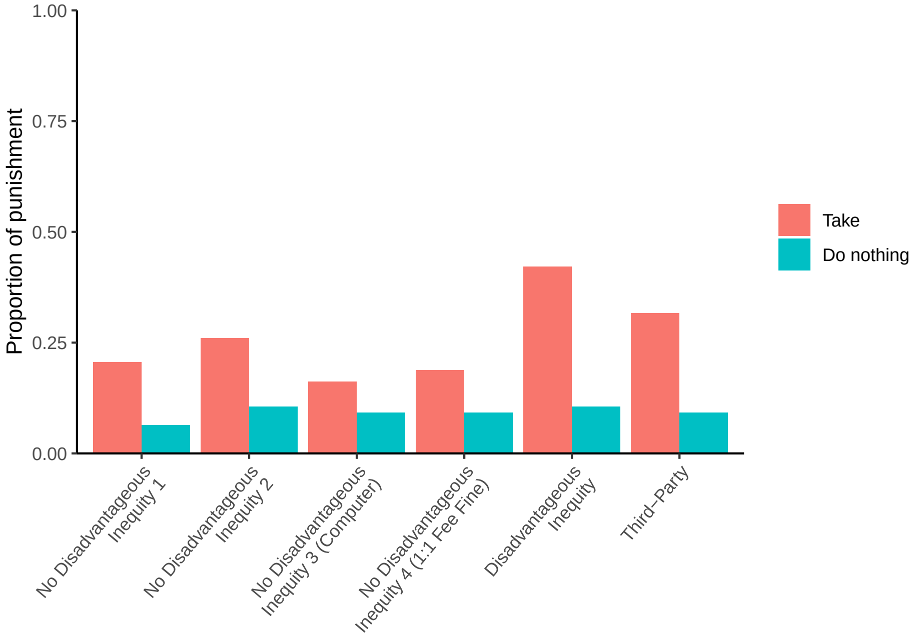
Comprehension

Answers to the comprehension questions revealed that participants were able to understand the payoff structure of all six punishment games.

Game	Comprehension Rate
No Disadvantageous Inequity 1	0.96
No Disadvantageous Inequity 2	0.95
No Disadvantageous Inequity 3 (Computer)	0.95
No Disadvantageous Inequity 4 (1:1 Fee Fine)	0.95
Disadvantageous Inequity	0.96
Third-Party	0.95

Punishment decisions

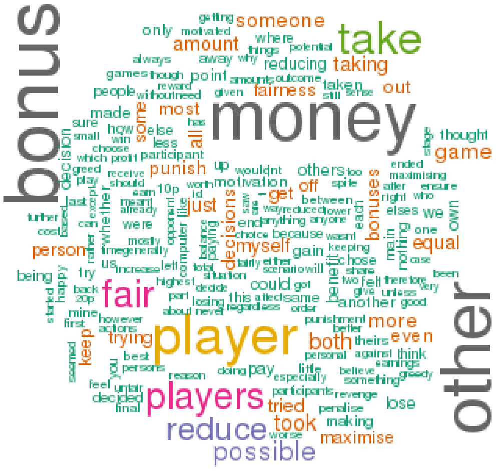
We can plot the proportion of participants who decided to punish in each game.



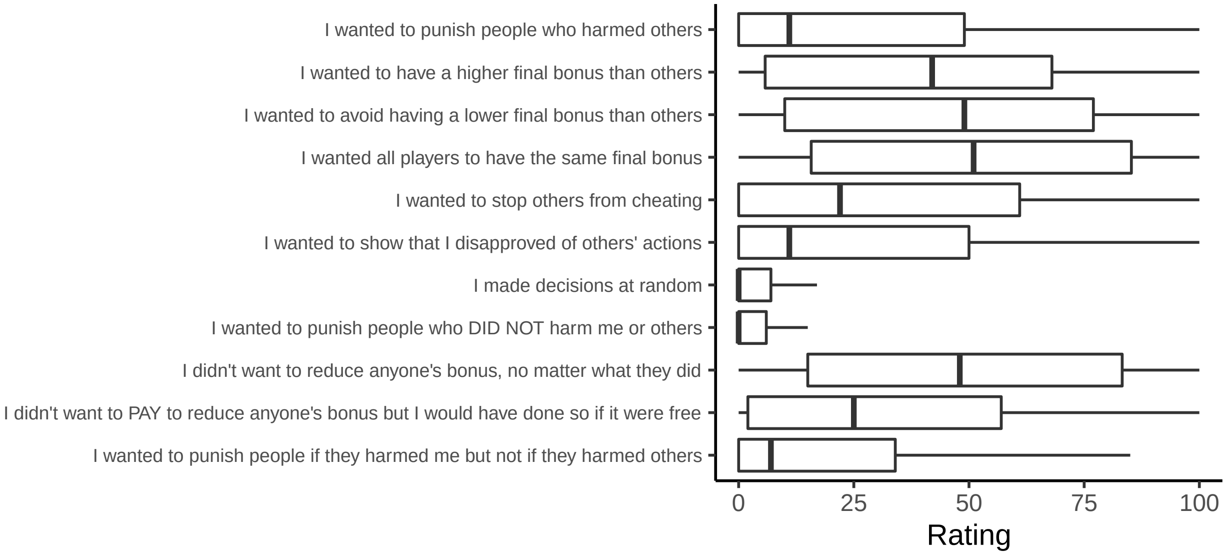
Participants appear more likely to punish if the other player took, compared to when they did nothing. Participants were most likely to punish when the other player took in the disadvantageous inequity game and in the third-party game.

Reasons given for punishing in the games

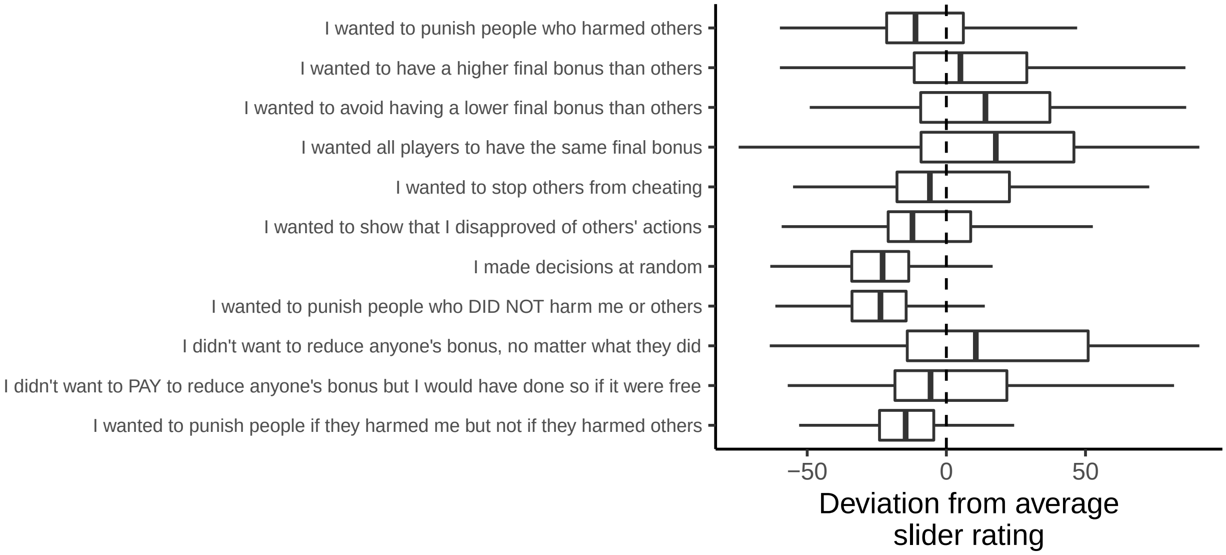
At the end of the survey, we asked participants why they decided to punish (if they ever did). First, we allowed them to provide an open-ended answer to this question. The following wordcloud summarises frequently used words in these open-ended answers.



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, 1000 = statement does apply).



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



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 from raw data

As a first step, we can look to see how many participants fitted these strategy predictions *exactly* across all games, by simply counting the raw data.

Strategy	N	Proportion
Competitive	3	0.003
Avoid disadvantageous inequity	67	0.066
Egalitarian	65	0.064
Seek advantageous inequity	2	0.002

Strategy	N	Proportion
Retributive	6	0.006
Deterrent	9	0.009
Norm-enforcing	8	0.008
Anti-punish	427	0.419
N/A	432	0.424

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.

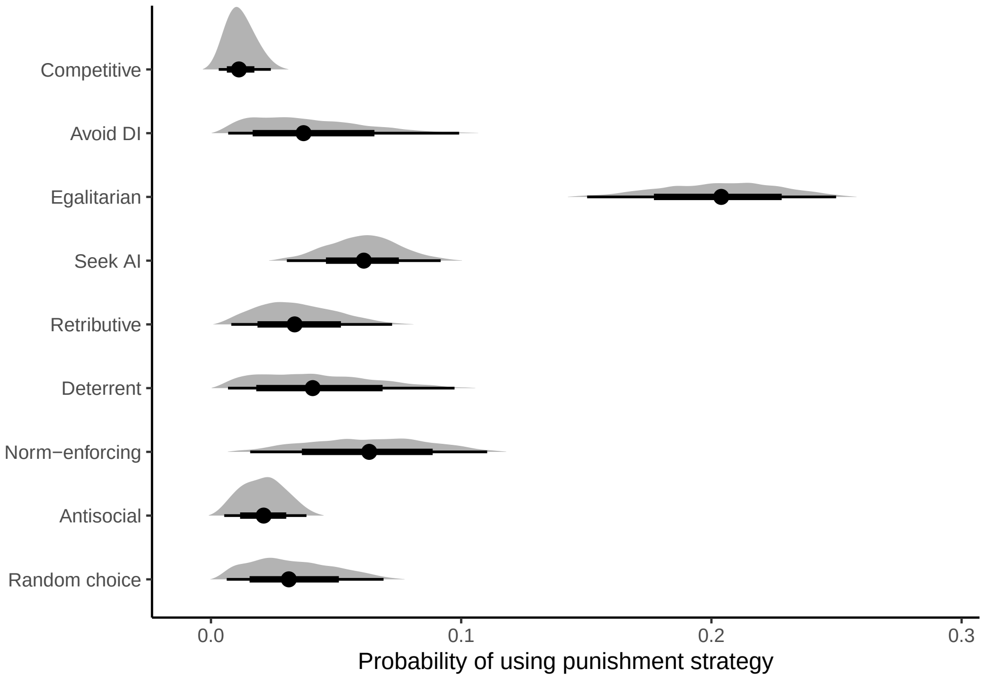
We can also look at the most common behavioural patterns across all six games (twelve punishment decisions in total). Below, we summarise the five most common patterns of behaviour as strings of twelve 0s and 1s, indicating whether participants did (1) or did not (0) punish in each game, with a brief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

Pattern	Explanation	N	Proportion
000000000000	Anti-punish strategy (exact)	427	0.42
000000001000	Avoid DI strategy (exact)	67	0.07
000000001010	Egalitarian strategy (exact)	65	0.06
000000000010	Only punish in third-party game	57	0.06
001000001000	Punish in No DI 2 and DI games	14	0.01

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 above. The Bayesian model is fitted in the probabilistic programming language Stan.

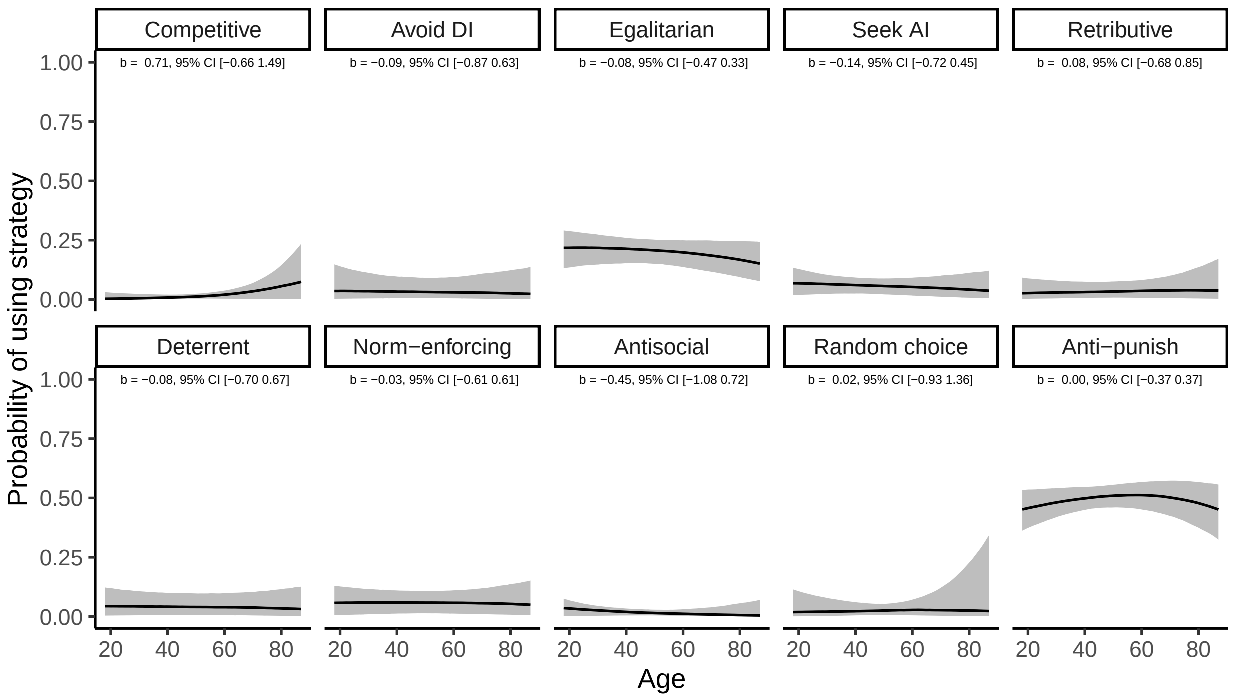
In line with our raw counts, the model predicts that participants follow the "anti-punish" strategy with a probability of 0.49, 95% credible interval [0.44 0.53]. We plot the posterior probabilities for the remaining punishment strategies below.

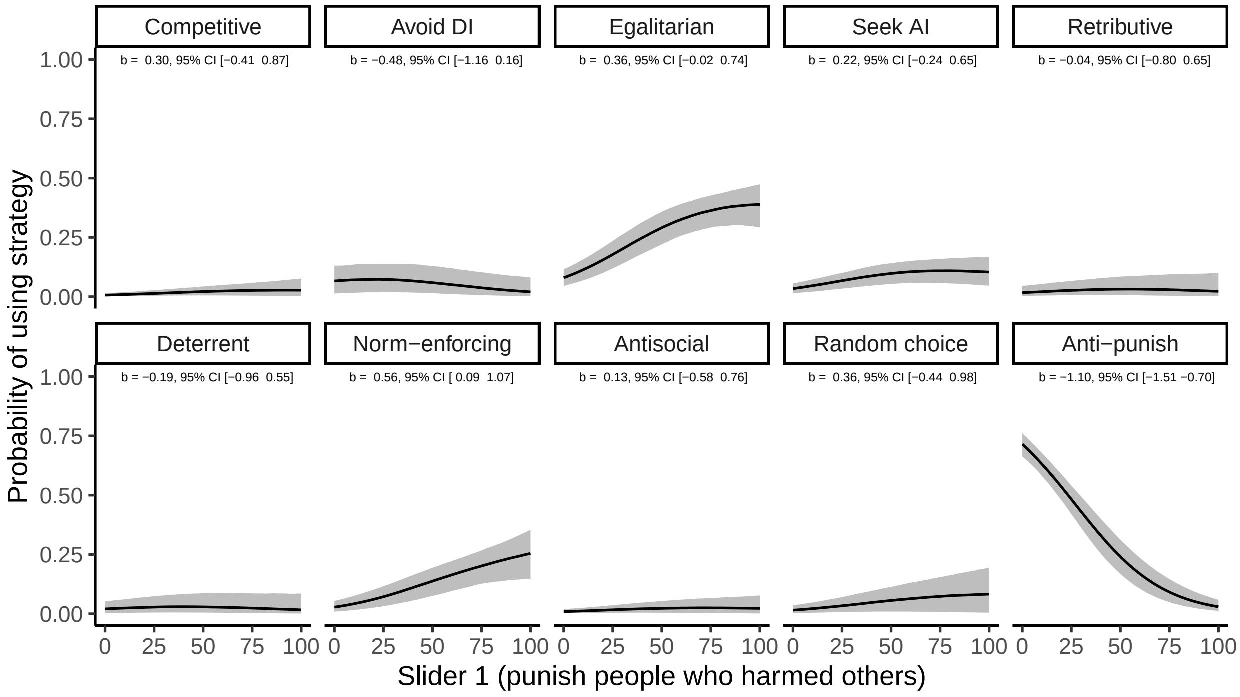


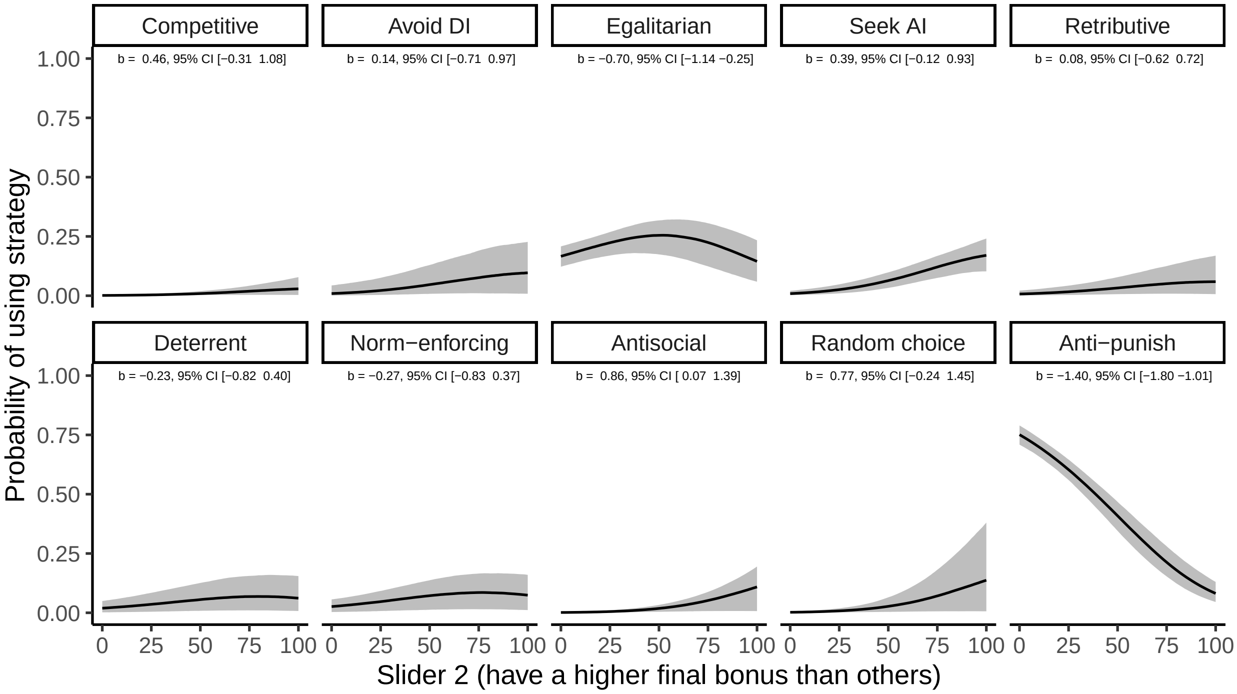
Taking advantage of all available data, the model suggests that "egalitarian" is the most common punishment strategy. All other strategies have median posterior probabilities less than 10%. In order of decreasing probability, the next most common punishment strategies are "norm-enforcing", "seek advantageous inequity", "deterrent", and "avoid disadvantageous inequity", although all are similarly likely. "Competitive" and "antisocial" punishment strategies are the most unlikely.

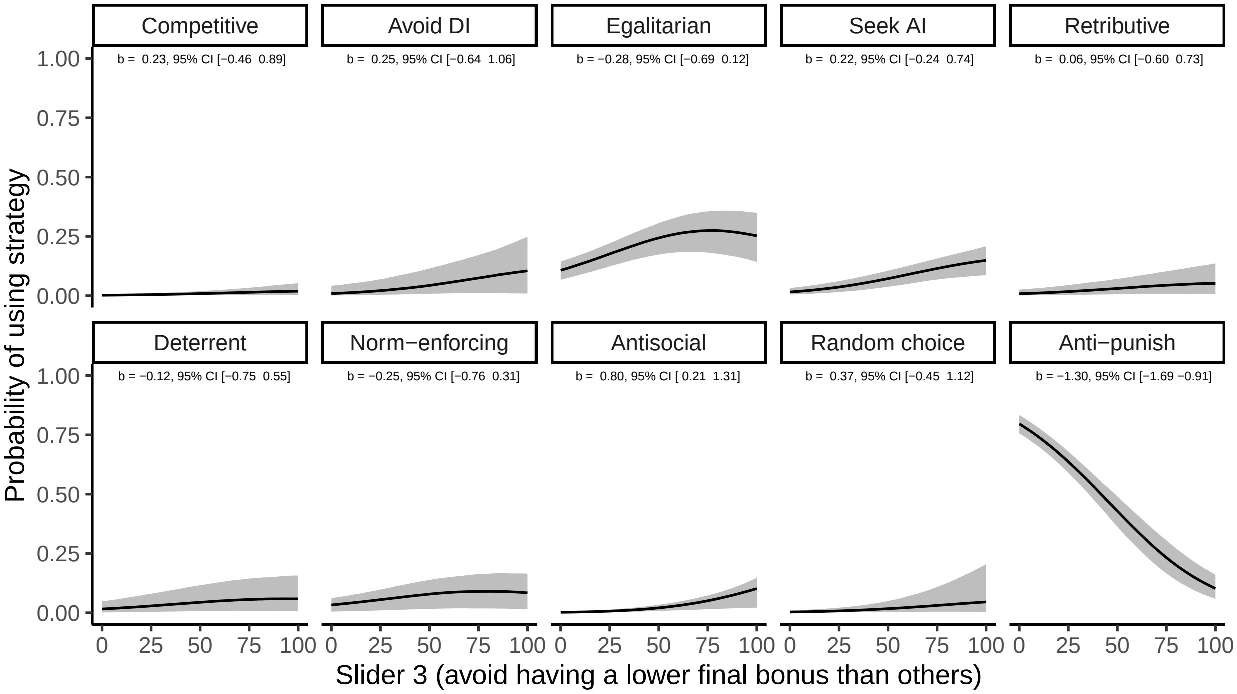
Predicting strategy usage

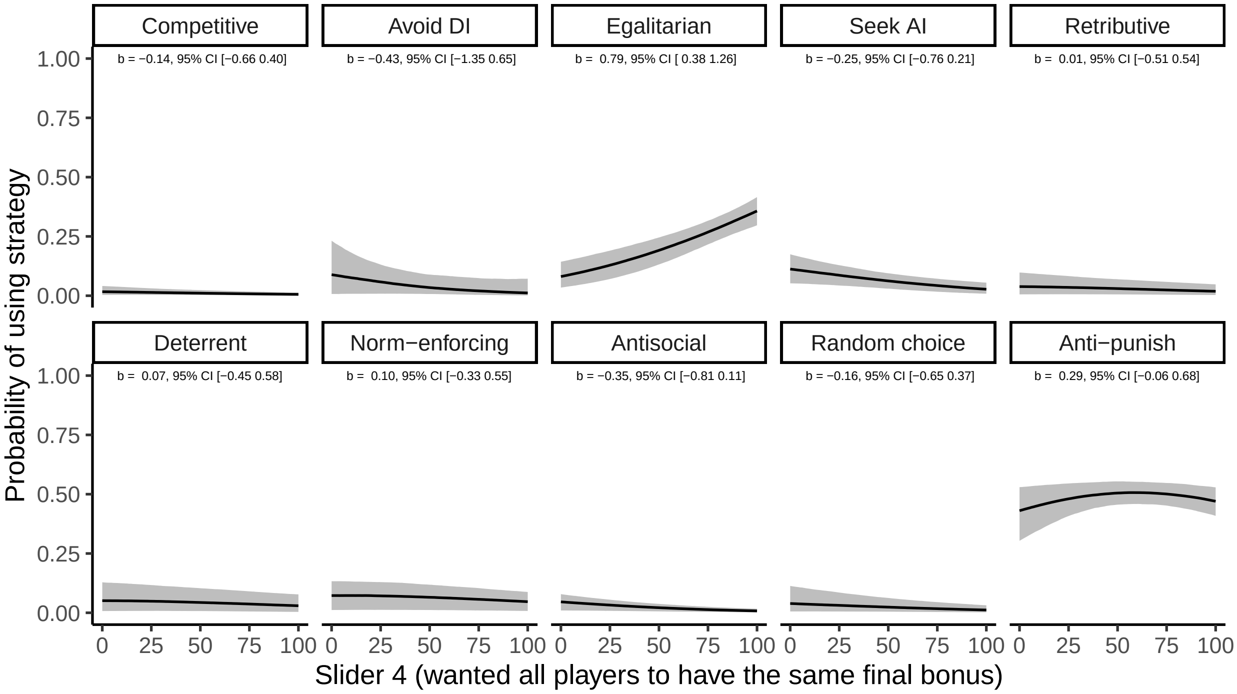
We then fit a series of models predicting the usage of different strategies from a variety of predictors. In each plot that follows, we show the posterior predictions from a model that includes a single variable as a predictor of all ten strategies simultaneously. The plots show the median regression lines with shaded 95% credible intervals. The model-estimated slopes for each strategy are included in each panel.

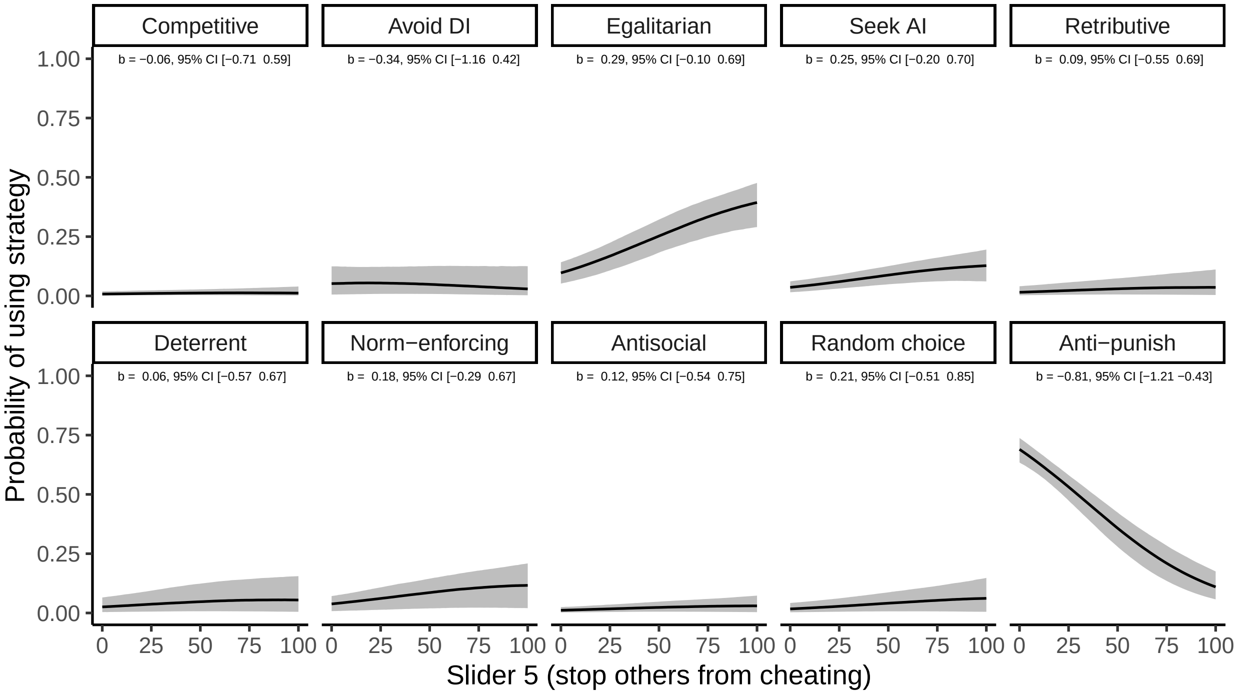

Competitive
b = 0.71, 95% CI [−0.66 1.49]
Avoid DI
b = −0.09, 95% CI [−0.87 0.63]
Egalitarian
b = −0.08, 95% CI [−0.47 0.33]
Seek AI
b = −0.14, 95% CI [−0.72 0.45]
Retributive
b = 0.08, 95% CI [−0.68 0.85]
Deterrent
b = −0.08, 95% CI [−0.70 0.67]
Norm−enforcing
b = −0.03, 95% CI [−0.61 0.61]
Antisocial
b = −0.45, 95% CI [−1.08 0.72]
Random choice
b = 0.02, 95% CI [−0.93 1.36]
Anti−punish
b = 0.00, 95% CI [−0.37 0.37]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
20
40
60
80
Age

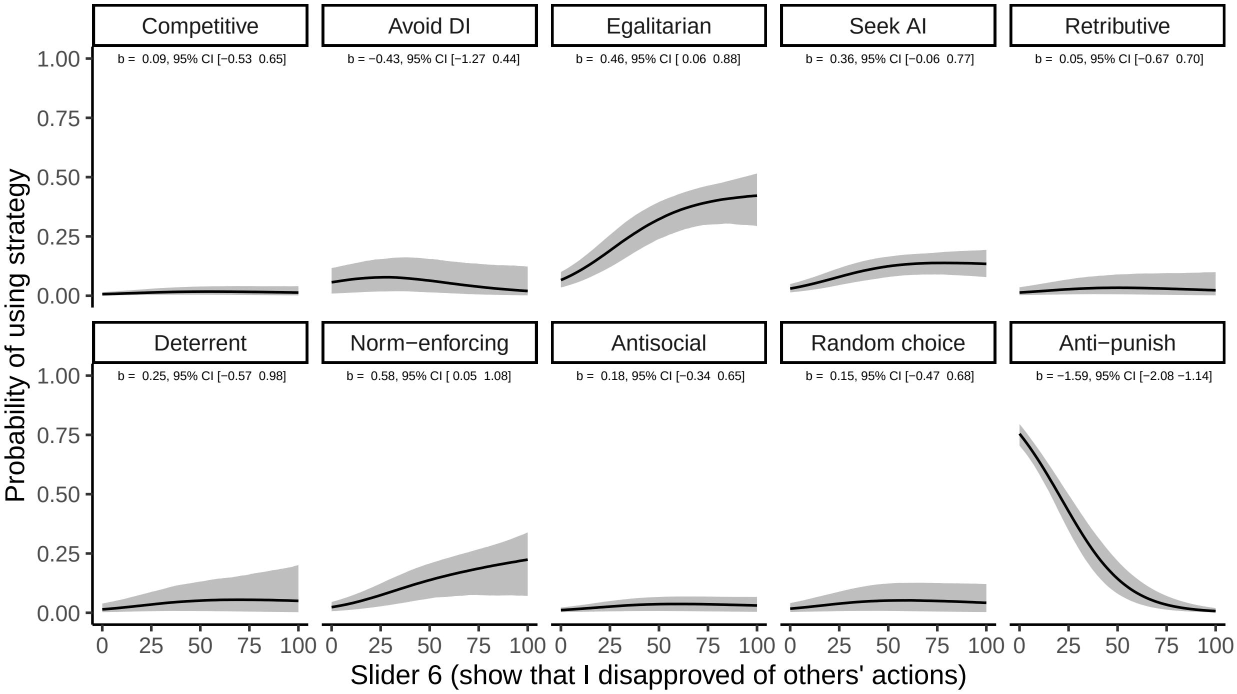

Competitive
b = 0.30, 95% CI [−0.41 0.87]
Avoid DI
b = −0.48, 95% CI [−1.16 0.16]
Egalitarian
b = 0.36, 95% CI [−0.02 0.74]
Seek AI
b = 0.22, 95% CI [−0.24 0.65]
Retributive
b = −0.04, 95% CI [−0.80 0.65]
Deterrent
b = −0.19, 95% CI [−0.96 0.55]
Norm−enforcing
b = 0.56, 95% CI [0.09 1.07]
Antisocial
b = 0.13, 95% CI [−0.58 0.76]
Random choice
b = 0.36, 95% CI [−0.44 0.98]
Anti−punish
b = −1.10, 95% CI [−1.51 −0.70]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 1 (punish people who harmed others)

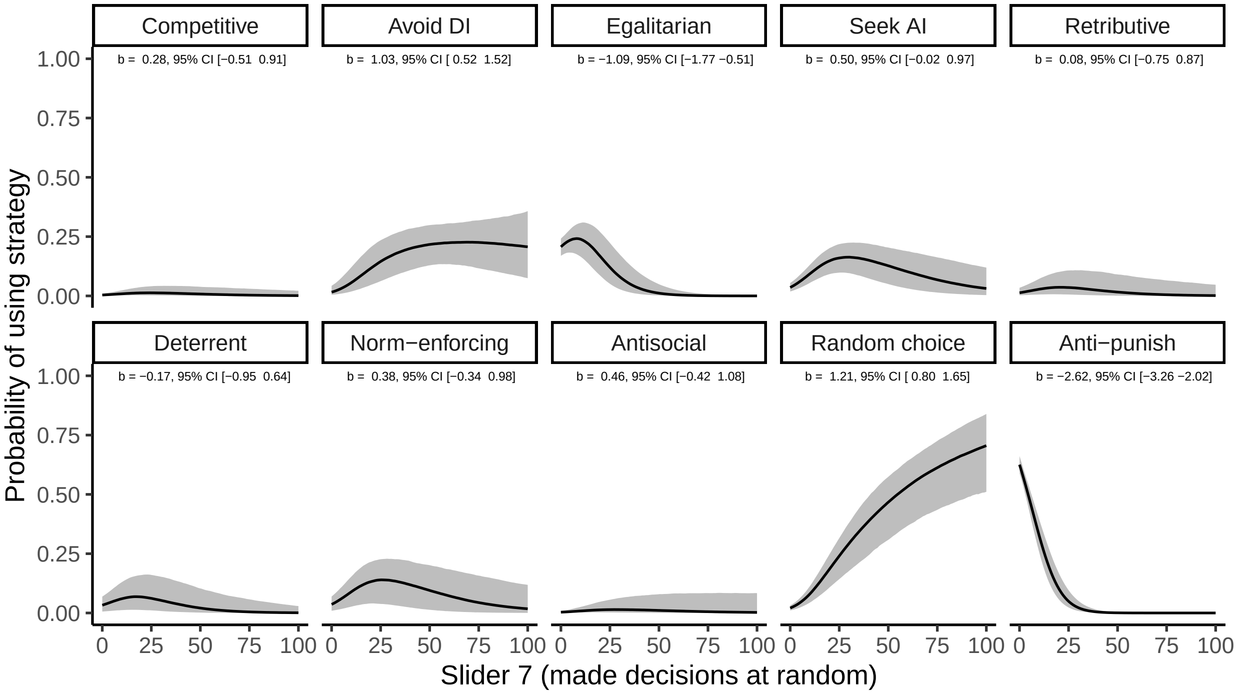

Competitive
b = 0.46, 95% CI [−0.31 1.08]
Avoid DI
b = 0.14, 95% CI [−0.71 0.97]
Egalitarian
b = −0.70, 95% CI [−1.14 −0.25]
Seek AI
b = 0.39, 95% CI [−0.12 0.93]
Retributive
b = 0.08, 95% CI [−0.62 0.72]
Deterrent
b = −0.23, 95% CI [−0.82 0.40]
Norm−enforcing
b = −0.27, 95% CI [−0.83 0.37]
Antisocial
b = 0.86, 95% CI [0.07 1.39]
Random choice
b = 0.77, 95% CI [−0.24 1.45]
Anti−punish
b = −1.40, 95% CI [−1.80 −1.01]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 2 (have a higher final bonus than others)

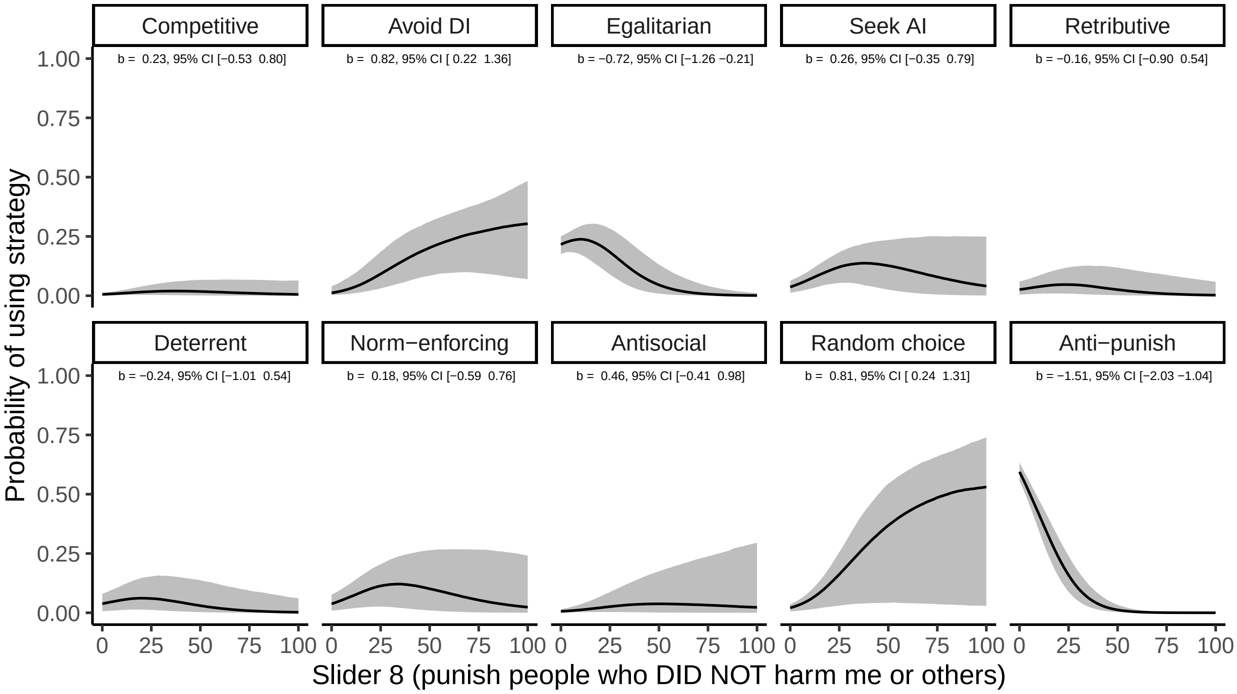

Competitive
b = 0.23, 95% CI [−0.46 0.89]
Avoid DI
b = 0.25, 95% CI [−0.64 1.06]
Egalitarian
b = −0.28, 95% CI [−0.69 0.12]
Seek AI
b = 0.22, 95% CI [−0.24 0.74]
Retributive
b = 0.06, 95% CI [−0.60 0.73]
Deterrent
b = −0.12, 95% CI [−0.75 0.55]
Norm−enforcing
b = −0.25, 95% CI [−0.76 0.31]
Antisocial
b = 0.80, 95% CI [0.21 1.31]
Random choice
b = 0.37, 95% CI [−0.45 1.12]
Anti−punish
b = −1.30, 95% CI [−1.69 −0.91]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 3 (avoid having a lower final bonus than others)

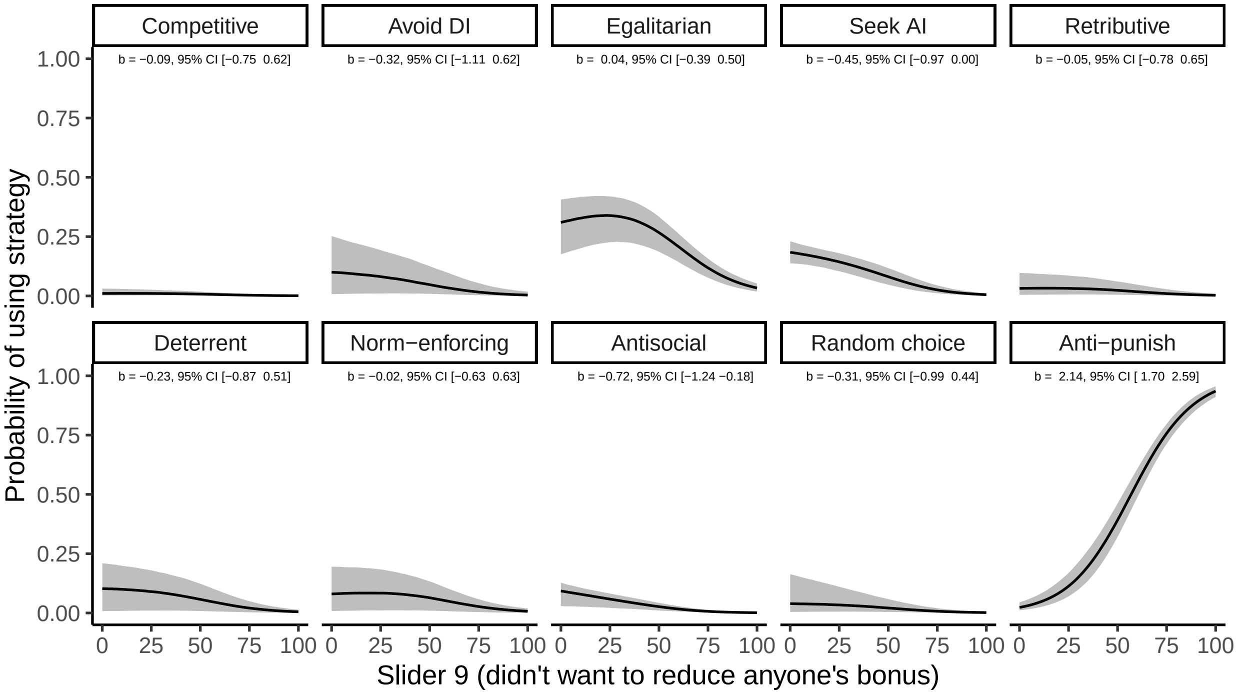

Competitive
b = −0.14, 95% CI [−0.66 0.40]
Avoid DI
b = −0.43, 95% CI [−1.35 0.65]
Egalitarian
b = 0.79, 95% CI [0.38 1.26]
Seek AI
b = −0.25, 95% CI [−0.76 0.21]
Retributive
b = 0.01, 95% CI [−0.51 0.54]
Deterrent
b = 0.07, 95% CI [−0.45 0.58]
Norm−enforcing
b = 0.10, 95% CI [−0.33 0.55]
Antisocial
b = −0.35, 95% CI [−0.81 0.11]
Random choice
b = −0.16, 95% CI [−0.65 0.37]
Anti−punish
b = 0.29, 95% CI [−0.06 0.68]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 4 (wanted all players to have the same final bonus)

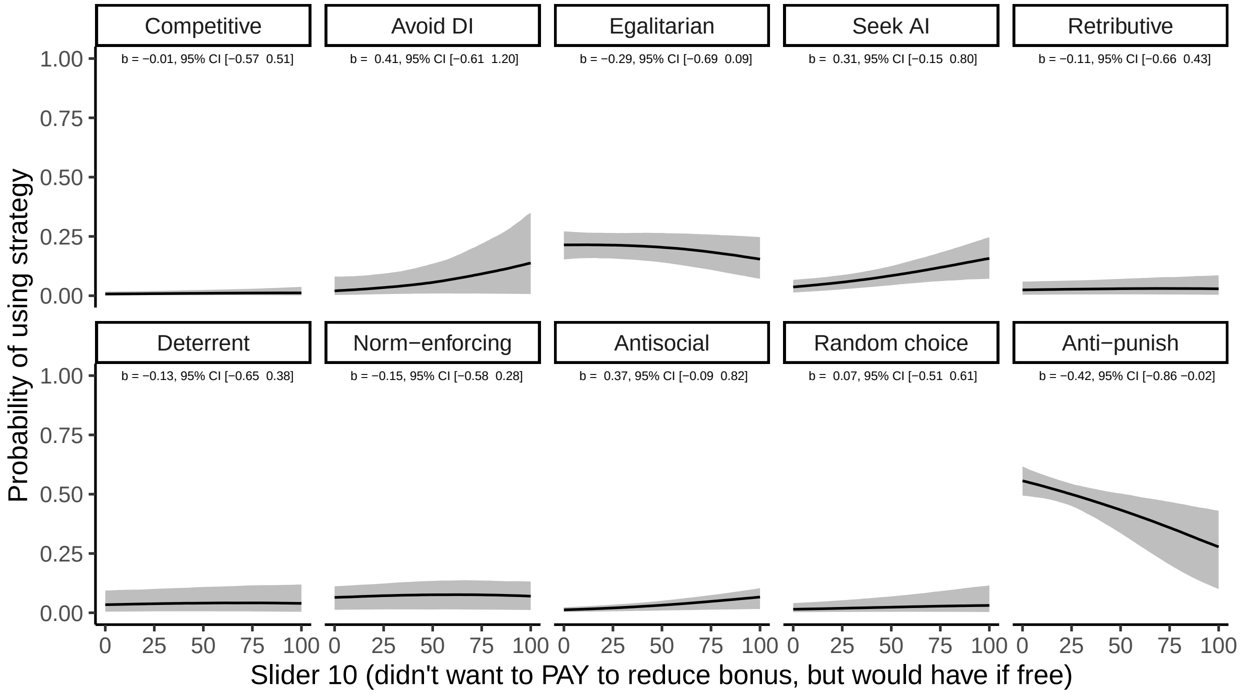

Competitive
b = −0.06, 95% CI [−0.71 0.59]
Avoid DI
b = −0.34, 95% CI [−1.16 0.42]
Egalitarian
b = 0.29, 95% CI [−0.10 0.69]
Seek AI
b = 0.25, 95% CI [−0.20 0.70]
Retributive
b = 0.09, 95% CI [−0.55 0.69]
Deterrent
b = 0.06, 95% CI [−0.57 0.67]
Norm−enforcing
b = 0.18, 95% CI [−0.29 0.67]
Antisocial
b = 0.12, 95% CI [−0.54 0.75]
Random choice
b = 0.21, 95% CI [−0.51 0.85]
Anti−punish
b = −0.81, 95% CI [−1.21 −0.43]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 5 (stop others from cheating)

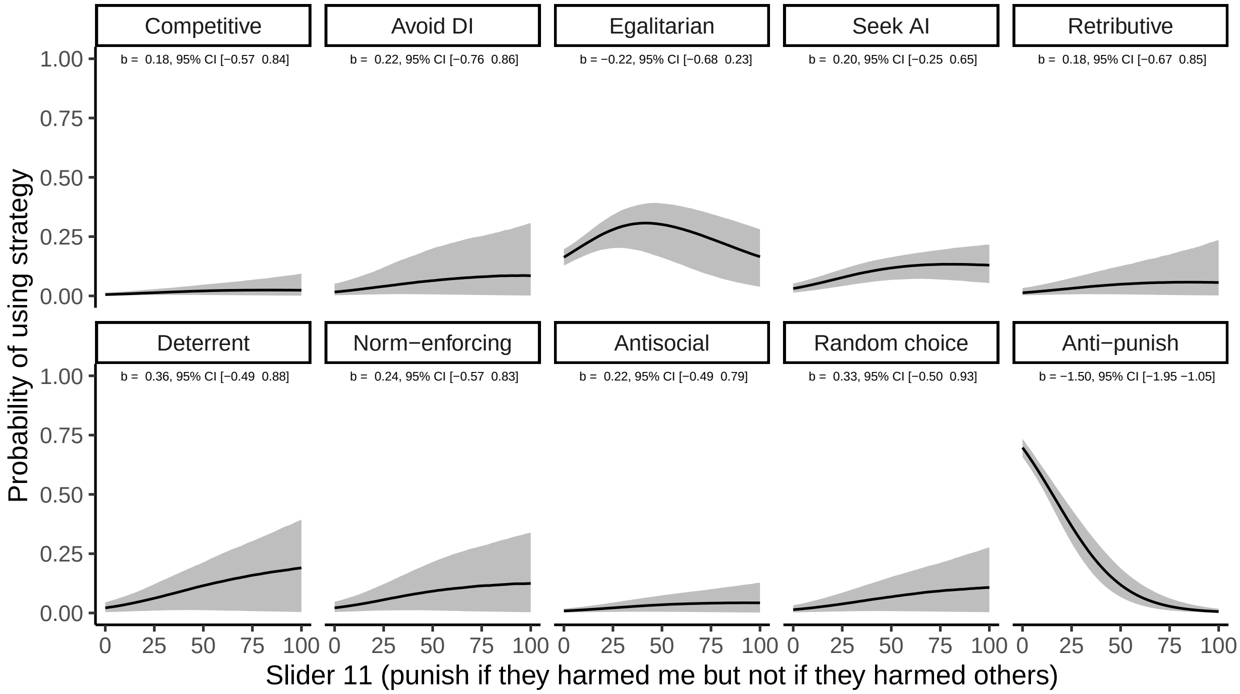

Competitive
b = 0.09, 95% CI [−0.53 0.65]
Avoid DI
b = −0.43, 95% CI [−1.27 0.44]
Egalitarian
b = 0.46, 95% CI [0.06 0.88]
Seek AI
b = 0.36, 95% CI [−0.06 0.77]
Retributive
b = 0.05, 95% CI [−0.67 0.70]
Deterrent
b = 0.25, 95% CI [−0.57 0.98]
Norm−enforcing
b = 0.58, 95% CI [0.05 1.08]
Antisocial
b = 0.18, 95% CI [−0.34 0.65]
Random choice
b = 0.15, 95% CI [−0.47 0.68]
Anti−punish
b = −1.59, 95% CI [−2.08 −1.14]
Probability of using strategy
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100
Slider 6 (show that I disapproved of others' actions)


Competitive
b = 0.28, 95% CI [−0.51 0.91]
Avoid DI
b = 1.03, 95% CI [0.52 1.52]
Egalitarian
b = −1.09, 95% CI [−1.77 −0.51]
Seek AI
b = 0.50, 95% CI [−0.02 0.97]
Retributive
b = 0.08, 95% CI [−0.75 0.87]
Deterrent
b = −0.17, 95% CI [−0.95 0.64]
Norm−enforcing
b = 0.38, 95% CI [−0.34 0.98]
Antisocial
b = 0.46, 95% CI [−0.42 1.08]
Random choice
b = 1.21, 95% CI [0.80 1.65]
Anti−punish
b = −2.62, 95% CI [−3.26 −2.02]
Probability of using strategy
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100
Slider 7 (made decisions at random)


Competitive
b = 0.23, 95% CI [−0.53 0.80]
Avoid DI
b = 0.82, 95% CI [0.22 1.36]
Egalitarian
b = −0.72, 95% CI [−1.26 −0.21]
Seek AI
b = 0.26, 95% CI [−0.35 0.79]
Retributive
b = −0.16, 95% CI [−0.90 0.54]
Deterrent
b = −0.24, 95% CI [−1.01 0.54]
Norm−enforcing
b = 0.18, 95% CI [−0.59 0.76]
Antisocial
b = 0.46, 95% CI [−0.41 0.98]
Random choice
b = 0.81, 95% CI [0.24 1.31]
Anti−punish
b = −1.51, 95% CI [−2.03 −1.04]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 8 (punish people who DID NOT harm me or others)


Competitive
b = −0.09, 95% CI [−0.75 0.62]
Avoid DI
b = −0.32, 95% CI [−1.11 0.62]
Egalitarian
b = 0.04, 95% CI [−0.39 0.50]
Seek AI
b = −0.45, 95% CI [−0.97 0.00]
Retributive
b = −0.05, 95% CI [−0.78 0.65]
Deterrent
b = −0.23, 95% CI [−0.87 0.51]
Norm−enforcing
b = −0.02, 95% CI [−0.63 0.63]
Antisocial
b = −0.72, 95% CI [−1.24 −0.18]
Random choice
b = −0.31, 95% CI [−0.99 0.44]
Anti−punish
b = 2.14, 95% CI [1.70 2.59]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 9 (didn't want to reduce anyone's bonus)


Competitive
b = −0.01, 95% CI [−0.57 0.51]
Avoid DI
b = 0.41, 95% CI [−0.61 1.20]
Egalitarian
b = −0.29, 95% CI [−0.69 0.09]
Seek AI
b = 0.31, 95% CI [−0.15 0.80]
Retributive
b = −0.11, 95% CI [−0.66 0.43]
Deterrent
b = −0.13, 95% CI [−0.65 0.38]
Norm−enforcing
b = −0.15, 95% CI [−0.58 0.28]
Antisocial
b = 0.37, 95% CI [−0.09 0.82]
Random choice
b = 0.07, 95% CI [−0.51 0.61]
Anti−punish
b = −0.42, 95% CI [−0.86 −0.02]
Probability of using strategy
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Slider 10 (didn't want to PAY to reduce bonus, but would have if free)


Competitive
b = 0.18, 95% CI [−0.57 0.84]
Avoid DI
b = 0.22, 95% CI [−0.76 0.86]
Egalitarian
b = −0.22, 95% CI [−0.68 0.23]
Seek AI
b = 0.20, 95% CI [−0.25 0.65]
Retributive
b = 0.18, 95% CI [−0.67 0.85]
Deterrent
b = 0.36, 95% CI [−0.49 0.88]
Norm−enforcing
b = 0.24, 95% CI [−0.57 0.83]
Antisocial
b = 0.22, 95% CI [−0.49 0.79]
Random choice
b = 0.33, 95% CI [−0.50 0.93]
Anti−punish
b = −1.50, 95% CI [−1.95 −1.05]
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)