Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent

Barnby, J.M.* [1,2,3], Raihani, N.[4], Dayan, P.[5,6]

*Corresponding Author

Author affiliations:

[1] Department of Psychology, Royal Holloway, University of London, London, UK

[2] Cultural and Social Neuroscience Group and [3] Neuropharmacology Group, Department of Neuroimaging, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

[4] Experimental Psychology, University College London, London, UK.

[5] Max Planck Institute for Biological Cybernetics, Tübingen, DE

[6] University of Tübingen, Tübingen, DE

Highlights

- Personal social values are integrated into prior beliefs about interaction partners
- Alignment of social values prior to interaction increased predictive accuracy
- Flexibility of prior beliefs increased predictive accuracy, regardless of alignment
- Misalignment of social values resulted in larger attributions of harmful intent
- Paranoia predicts less flexible beliefs about a partner's relative payoff preferences

Abstract

To benefit from social interactions, people need to predict how their social partners will behave. Such predictions arise through integrating prior expectations with evidence from observations, but where the priors come from and whether they influence the integration into beliefs about a social partner is not clear. Furthermore, this process can be affected by factors such as paranoia, in which the tendency to form biased impressions of others is common. Using a modified social value orientation (SVO) task in a large online sample (n=697), we showed that participants used a Bayesian inference process to learn about partners, with priors that were based on their own preferences. Paranoia was associated with preferences for earning more than a partner and less flexible beliefs regarding a partner's social preferences. Alignment between the preferences of participants and their partners was associated with better predictions and with reduced attributions of harmful intent to partners. Together, our data and model expand upon theories of interpersonal relationships by demonstrating how dyadic similarity mechanistically influences social interaction by generating more accurate predictions and less threatening impressions.

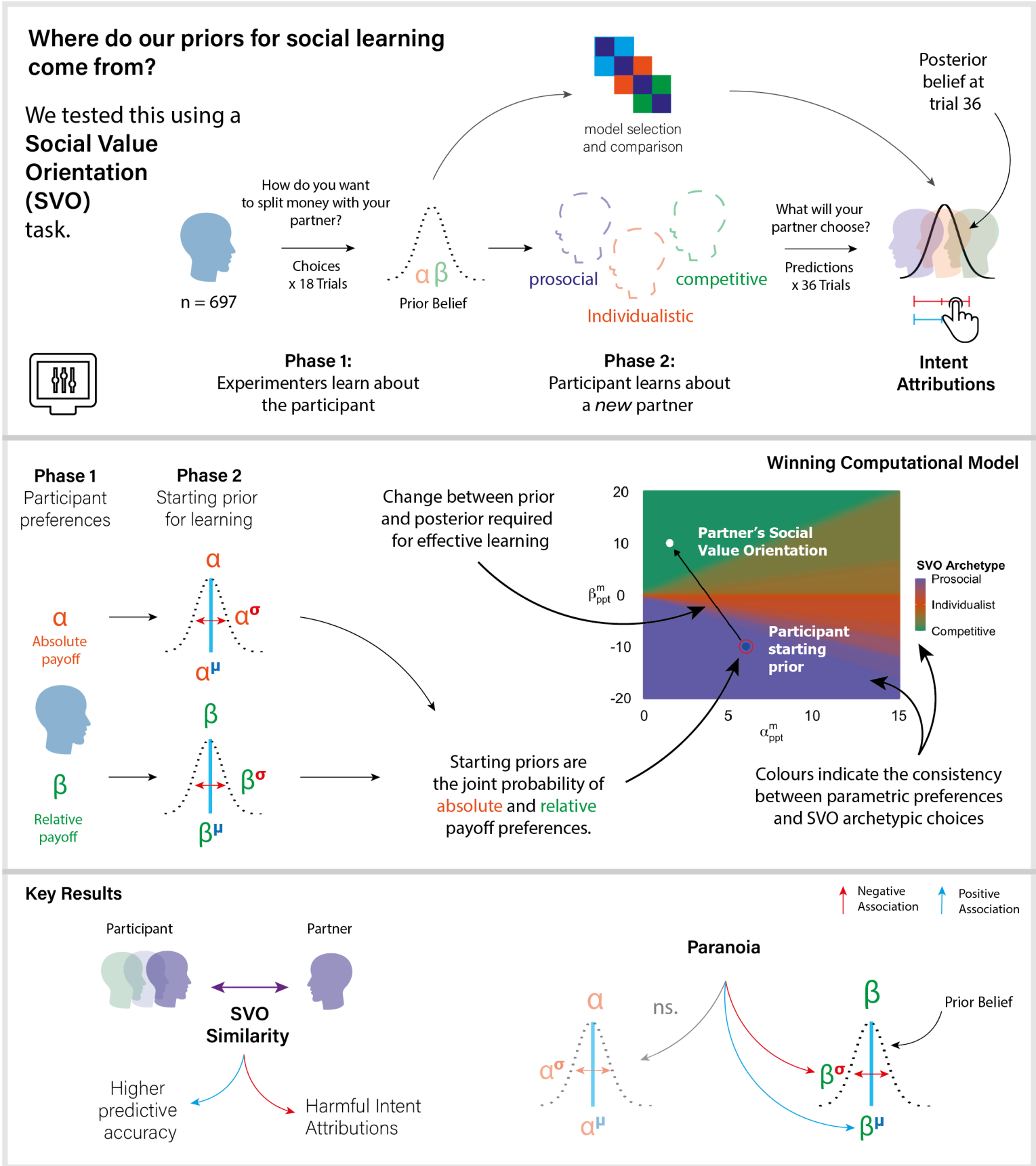
Keywords

Social-Value Orientation; Bayesian Belief; Paranoia; Interpersonal Alignment; Social Learning; Belief Integration

Abbreviations

AICc: Akaike Information Criterion (corrected); BIC: Bayesian Information Criterion; CBM: Concurrent Bayesian Modelling; ICAR: International Cognitive Ability Resource, Progressive Matrices; R-GPTS-B: Subscale B of the Revised Green Paranoid Thoughts Scale; SVO: Social-Value Orientation

Graphical Abstract



1. Introduction

How do people learn about the properties or preferences of other people in the world? Generically, they start with prior expectations and update them in the light of observations (Chater et al., 2006; Plitt & Giocomo et al., 2021; Vilares & Kording, 2011). When lacking specific information about others, a natural and readily accessible source of prior expectations for people is their own beliefs or preferences (Krueger & Clement, 1994). This will generate self-related biases in the inferences made about others (Andersen & Chen, 2002; Andersen & Glassman, 1996; Buckner & Carroll et al., 2007; Robbins & Krueger, 2005; Suzuki et al., 2016). These biases will be particularly important when data are scarce – a common regime in social contexts.

People's priors can also exert an enduring influence over the whole course of learning about others, as if their interpretation of observations is coloured by what they themselves think or would do in the same circumstances. This can affect predictions of the beliefs or actions of others. For instance, peoples' predictions about the choices partners would make between alternative snack foods remain partly biased by their own preferences, even after substantial observations of their partners' picks (Tarantola et al., 2017).

Priors can also influence learning in social interactions, for example when cooperating or trusting others. This is particularly apparent in psychiatric disorders, many of which involve distressing changes in inferences about the social orientations of others. For example, persecutory beliefs are associated with biases in social impression formation (Barnby et al., 2020a; Diaconescu et al., 2020; Lincoln et al., 2010; Raihani & Bell 2017; Saalfeld et al. 2018; Wellstein et al., 2020), and a defining feature of paranoia is an exaggerated belief that harm will occur, and that other people intend for it to happen (Freeman & Garety, 2000). In experimental settings, more paranoid people attribute more harmful intentions to others, including in scenarios where the partner's true intentions are ambiguous (Barnby et al. 2020b; Greenburgh et al. 2019; Raihani & Bell 2017; Saalfeld et al. 2018). Paranoia is also associated with variation in social preferences: more paranoid individuals are less trusting and less cooperative in experimental economic games (Fett et al., 2012; Hula et al., 2015; 2018; King-Casas et al., 2008; Raihani & Bell 2018; Raihani et al. 2021; Xiang et al., 2012) and other work has shown that paranoia positively predicts the enjoyment of negative social interactions (Raihani et al. 2021) and the willingness to inflict financial harm on a partner ('punishment', Raihani & Bell 2018; Raihani et al. 2021). Given the striking disruption of paranoia on interpersonal dynamics, both in health and illness, it is critical to understand the role of priors on the process of belief formation when information is scarce.

Here, we asked how participants' own social preferences influenced learning about the social preferences of others when an interaction partner's social preferences had financial consequences for the participant themselves. We also examined how variation in paranoia correlated with participants' own social preferences and the way that they formed and updated beliefs about their partners. To do this, we used a modified social-value orientation approach (SVO, Murphy et al. 2011; Murphy & Ackermann, 2014). SVO describes a participant's social preferences and can be measured using a task where decision outcomes impact both the participant and a notional recipient. Specifically, participants can be classified as being prosocial if they typically prefer equal outcomes, individualistic if they prefer to maximise absolute earnings for themselves, and competitive if they prefer to maximise the relative payoff difference between themselves and a partner. In phase 1 of our modified SVO task, participants acted in the SVO decider role for 18 trials; in phase 2, they acted as the recipient for 36 trials when choices were made by a (new) partner. In phase 2, participants predicted which option the partner would choose in each trial, allowing us to measure initial priors about the partner and subsequent learning (Figure 1).

We formalised the influence of a participant's SVO on their learning about their partner by building and comparing ten Bayesian belief and five heuristic models (Table A.1). We used phase 1 behaviour to estimate participants' own SVO, operationalised in terms of parameters of their subjective utility for earnings for themselves and their partner. In phase 2, models either a) integrated a participant's SVO into their prior beliefs about their partner in a Bayesian updating process, b) used separate prior values for the inferred SVO of a partner (but still used a Bayesian updating process), or c), adopted a heuristic reinforcement learning approach.

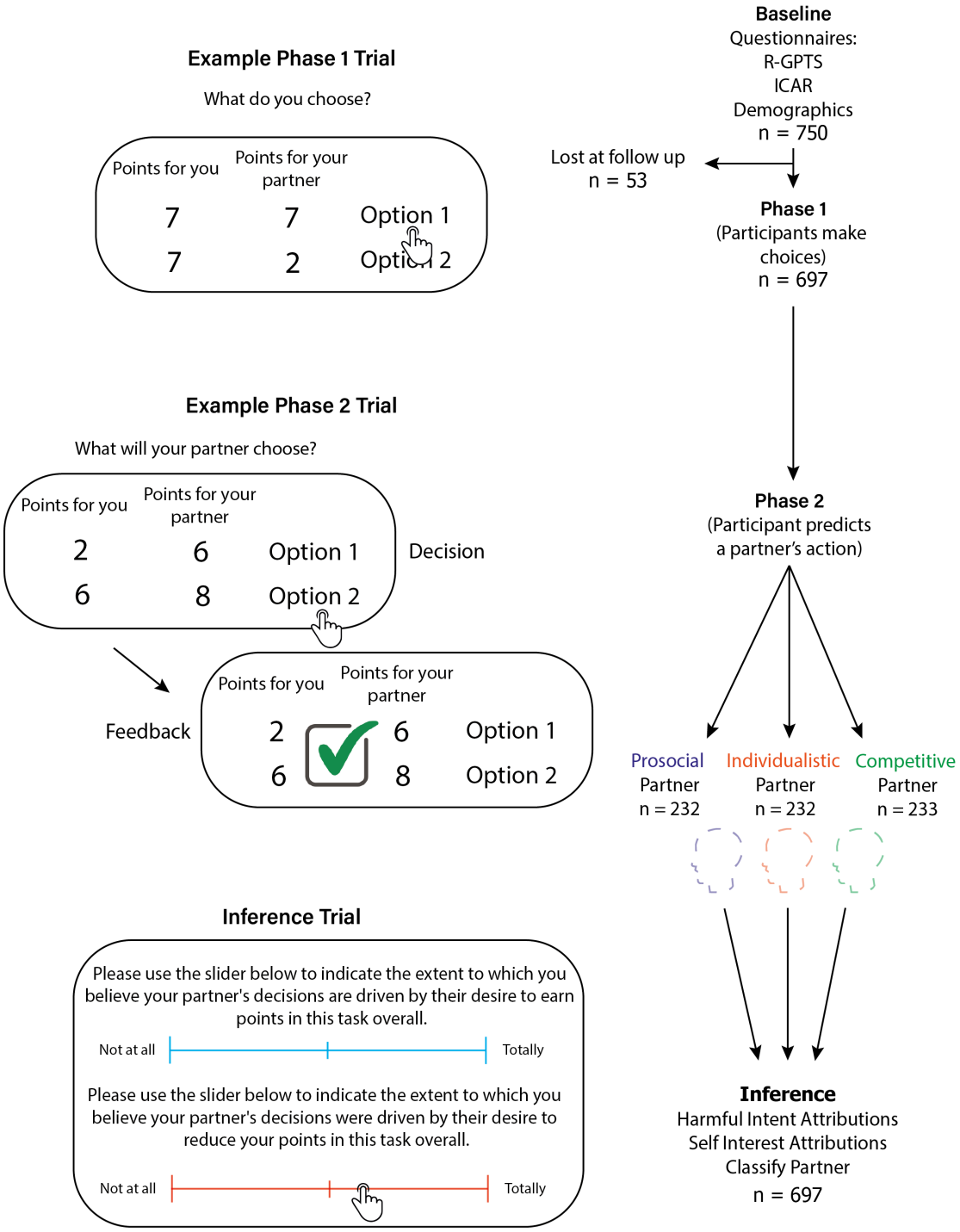


Figure 1: Study design

In phase 1, participants chose between two options for 18 trials. In phase 2, participants were randomly matched with a prosocial, competitive, or individualistic partner and predicted which options their partner would choose on 36 trials. Participants received feedback after each trial about what their partner actually chose. After phase 2, participants were asked to infer the extent to which they believed their partner was motivated by self-interest and harmful intent, respectively. They were also asked to classify their partner into one of three categories depicting the partner's social preferences. These categories were described as whether they thought their partner was primarily aiming to (i) equalise payoffs, (ii) earn as much money as possible, or (iii) prevent the participant from earning money.

2. **Materials and Methods**

2.1. Participants

We recruited 750 participants for the initial baseline assessment via Prolific Academic in March 2020. Participants first completed the R-GPTS (Freeman et al., 2020), the ICAR matrices (Condon & Revelle, 2014), and provided demographic information (age, sex, education). After a minimum interval of seven days, we were successful in recalling 697 participants to take part in the experimental paradigms. All participants were between the ages of 16-65, were UK residents, fluent in English, had at least a 90% approval rating on Prolific Academic, and had no prior or current psychiatric diagnosis.

Overall, our sample passed quality control checks, with only 7.1% failing both control questions, 14.9% getting one control question wrong, and 78% getting both control questions correct. Task comprehension was included as an explanatory variable in all regression models. In addition, on a scale of 0-100 (0 = I did not believe that my partner was a real person, 100 = I believed that my partner was a real person), participants were more inclined to believe their partners in the game were real (mean = 59.61, sd = 30.37, median = 66, min = 0, max = 100, skew = -0.43).

2.2. *The modified SVO task*

We built a modified SVO task based on existing paradigms (Murphy et al., 2011). Participants were asked to play in two phases of the task. Participants were informed that the points they earned in the task would contribute to an overall point total which was pooled over a series of tasks. Other tasks in the series that contributed to the points total are reported in a different paper (Barnby et al., *In Prep*). Participants were informed they would be matched with two anonymous partners (one for Phase 1 and another for Phase 2) online.

In Phase 1, participants played the role of the decider in a two-player SVO task (Murphy et al., 2011) over 18 trials. In each trial, participants chose between two options determining financial rewards (framed as points) for themselves and an anonymous partner (Table B.1). Participants made 6 choices between prosocial and competitive options, 6 choices between individualistic and competitive options and 6 choices between prosocial and individualistic/competitive options. Options in this latter category were classified as being individualistic/competitive because they were consistent both with a participant's preference to maximise own payoffs and with a preference to establish a payoff advantage over the partner. We did not use these categorical labels when presenting options to participants and instead used neutral labels, Option 1, and Option 2, in each trial.

In Phase 2, participants played in the recipient role with a new partner with whom they were randomly matched. For a breakdown of the distribution of participants preferences within each partner type see Table F.1 and Figure G.1. Over 36 trials, participants predicted which of two options the partner would choose. Participants were incentivised to predict accurately because accurate predictions contributed to their total point score, which determined entry into a financial lottery. Partner decisions were decided manually a priori, and without noise: prosocial partners always chose the option that maximized equality and in cases of competitive/individualist option pairs the prosocial partner never decided upon the competitive option; competitive partners always tried to reduce the participant's bonus as much as possible (and chose the individualist option otherwise); and individualist partners always chose the highest payoff for themselves (and chose the prosocial option otherwise). After predicting which option their partner would choose, participants were provided with feedback about whether their answer was correct or not (Figure 1). Finally, participants were asked to what extent they thought their partner was motivated by harmful intent and self-

interest (using two separate slider scales from 0-100, with the slider invisible until the participant had made the first click). They then answered (using a 3-option forced-choice question) whether they thought their partner was (1) aiming to share the money equally, (2) trying to earn as much money as possible, or (3) to prevent the participant from earning money.

2.3. Regression Modelling

All linear models were constructed using the 'LME4' package (Bates et al., 2015; v1.1-23) and averaged using the 'MuMIn' package (Bartoń, 2020; v1.43.17) with data wrangling using 'dplyr' (Wickham et al., 2021; v1.0.7) and plotting using 'ggplot2' (Wickham, 2016; v3.3.3) in R (R Core Team, 2020; Version 4.0.0, 2020/04/24) on a Mac OS (Big Sur v11.1). All continuous variables, including model derived parameters, were centred, and scaled using the 'scale' base function in r which normalises the values of the distribution but does not change the shape of the distribution.

Categorical pairwise choice analysis reported used the cumulative sum for each type of decision a participant made within each choice pair. For example, within prosocial-competitive choice pairs, prosocial choices were dummy coded as 1, and all instances of a participant choosing the prosocial option when this choice pair was available were summed (for a maximum of 6 per participant, per choice pair); see Table B.1). We constructed 3 separate models, one for each choice pair (prosocial-competitive; prosocial-individualist/competitive; competitive-individualist).

We derived regression estimates (which we refer to as 'estimates' in the text) associated with parameters in our statistical models using model averaging (Burnham & Anderson, 1998; 2002; Grueber et al., 2011), which accounts for the uncertainty in producing parameter estimates when more than one model is consistent with the observed data. Briefly, we defined a global model, containing all explanatory terms and interactions of interest and then derived a top model set, using the dredge function in MuMIn (Bartoń, 2020) which compares the global model and all possible sub models. We defined the top model set as being the model with the lowest AICc value and all models within 2 AICc units of that top model. Parameter estimates were then obtained by averaging across this top model set, which accounts for the fact that some parameters do not appear in all the top models. We report full rather than conditional model-averaged estimates, as the former are more conservative. All associations derived from model averaging control for paranoia, general cognitive ability, task comprehension, age, and sex, unless stated otherwise.

All averaged regression models are given an identifier (e.g., Model 1a) to signify which estimates came from the same model and correspond to the model signifier in the RMarkdown workbook available on GitHub (see below for link).

In addition to the model-averaging approach described above, we also conducted stepwise linear regression models to explore the variance explained by each parameter that appeared in the top model set; we report r^2 values associated with these models, generated via a standard stepwise approach (i.e., by sequentially adding terms of interest to our baseline model). We used the 'lm' function in R and we report the overall model fit in addition to regression coefficient strength (referred to as 'non-averaged estimates').

All data, model code, and analysis scripts/workbooks are available on GitHub: https://github.com/josephmbarnby/Barnby_etal_2021_SVO.

2.4. Computational modelling

We implemented a suite of models that belong to two broad 'classes' of computational theory in social learning: reinforcement learning models (k = 5) and more structured inferential Bayesian models (k = 10; Table A.1). While reinforcement learning models allow the tracking of reward-predictive signals from social others, Bayesian models instantiate explicit adherence to the potential structured nature of inference about others (Vélez & Gweon, 2021). Both are important to test which method may be most suited and explanatory to the way our participants integrate their own preferences into the beliefs about their partner. Rather than using the three categorical SVO definitions as with the heuristic models (individualist, competitive, prosocial), the Bayesian models decompose the preferences of participants and partners according to a reduced form of Fehr-Schmidt inequality aversion model (Fehr & Schmidt, 1999) which parameterises the subjective utility U of a choice between reward R_{self} for the chooser and R_{other} for the partner as follows:

(1)

$$U_{\alpha,\beta}(\mathbf{R}) = \alpha * R_{\text{self}} + \beta * \max(R_{\text{self}} - R_{\text{other}}, 0)$$

where $\mathbf{R}=\{R_{\text{self}}, R_{\text{other}}\}$. Given a choice between two such option pairs, $\mathbf{R} = \{\mathbf{R^1}; \mathbf{R^2}\}$, the probability of choosing the first option is taken to be

$$P(c = 1|\alpha, \beta; \mathbf{R}) = \sigma(U_{\alpha,\beta}(\mathbf{R^1}) - U_{\alpha,\beta}(\mathbf{R^2})),$$

where $\sigma(\cdot)$ is the logistic sigmoid.

Here, α describes the weight a participant places on their own payoff (in one reduced Bayesian model we set $\alpha = 0$), and β, the weight a participant places on their payoff relative to the payoff of their partner. Large positive or negative values of β indicate respectively that participants like or dislike earning more than their partner. We can therefore describe these terms α and β as reflecting preferences for absolute and relative payoffs, respectively. For the option set we used, $R_{\text{self}} > R_{\text{other}}$ so one can also write $U_{\alpha,\beta}(\mathbf{R}) = (\alpha + \beta) * R_{\text{self}} - \beta * R_{\text{other}}$.

Following usual SVO practice, the partners in our study acted according to the choices reported in the Appendix (Table B.1). Broadly, individualist partners had high α and minimal β; prosocial partners had substantially negative values of β, so that their subjective utility was reduced if they earned a lot more than their partners; and competitive partners tended to have more positive values of β, so their subjective utility was increased when they earned more than their partners. Competitive and prosocial partners still tended to have positive values of α, because given choices with equal values of $R_{\text{self}} - R_{\text{other}}$, they tend to favour the option with a larger R_{self}.

All models were built, fitted, and compared using *Matlab* (Mathworks, 2020) using the CBM toolbox (Piray et al., 2019). Model comparison metrics (e.g., iBIC; Huys et al., 2011) estimated from Laplace approximation are useful for individually fitted models but treat each model during comparison as a fixed effect which leaves parameter estimation in hierarchical, mixed effect, models susceptible to outliers (Stephan et al., 2009). To overcome the issue of fixed effects in model comparison we used concurrent Bayesian model fitting that hierarchically estimated participant's parameters for each model and simultaneously compared all models using the CBM toolbox (Piray et al., 2019) using a stepwise method. We used broad priors to fit each model and individual hierarchically (mean = 0, variance = 7.5).

Model comparison consisted of four phases. We first fitted all model parameters to individuals using Laplace approximation with the 'CBM lap' function in the CBM toolbox. We then initially compared 11 models. In the second phase, we retained the viable models from step 1 and compared them with three additional models. In the fourth phase, we compared the winning models with a responsibility greater than or equal to 1%. We performed recovery analysis to confirm each model could simulate and recover data as expected before model comparison. All models that were weighted with <1% contribution to the overall hierarchical fit over all models were excluded at each step. We concurrently compared models 1-15 together using CBM in a subsample of 100 random participants to ensure our outcome from the stepped approach was reproducible across model space. Here, we note the formalisation of the winning model (Model 2). See the Appendix for the formalisation of the other models (Text A.1; B.1).

In the winning model (Model 2; Table A.1), the actual value of a participant's own preferences in Phase 1 influences the inferences they make about their partner in Phase 2. Therefore, both phases are important to clarify the preferences of the partner, and we engaged in simultaneous estimation of parameters from Phase 1 and Phase 2 rather than separating each segment of the task with separate models.

2.4.1. Phase 1 – estimating participants' social preferences

We modelled participants' social preferences) as ranging along two dimensions: absolute payoffs (α_{ppt}), indicating the weight participants place on their own payoffs; and relative payoffs (β_{ppt}), indicating the weight participants place on payoffs relative to the partner. The Fehr-Schmidt model also includes a term for quantifying how much a participant dislikes relative discrepancies in payoffs with their partner that results in them earning less ('disadvantageous inequality') which can arise when $R_{\text{self}} < R_{\text{other}}$, although this circumstance does not arise in our options. For one (ultimately losing) model, we adopted the restriction $\alpha = 0$ (meaning that only β determined the SVO of a participant and their belief about a partner).

Over 18 trials, participants made binary choices c^t, $t = \{1 \ldots T\}$ about whether option 1 or option 2 should be chosen given the returns $\mathbf{R^t}=\{\mathbf{R^{t;1}}; \mathbf{R^{t;2}}\} = \{R^{t;1}_{\text{self}}, R^{t;1}_{\text{other}};\ R^{t;2}_{\text{self}}, R^{t;2}_{\text{other}}\}$ for self and other for both offers, such that the log likelihood of the participant choosing option $c^t = 1$ is:

(2)

$$U_{\alpha_{ppt},\beta_{ppt}}(\mathbf{R^{t;1}}) = \alpha_{ppt} * R^{t;1}_{\text{self}} + \beta_{ppt} * \max(R^{t;1}_{\text{self}} - R^{t;1}_{\text{other}}, 0)$$

$$p(c^t = 1|\alpha_{ppt}, \beta_{ppt};\ \mathbf{R^t}) = \sigma(\Delta U_{\alpha_{ppt},\beta_{ppt}}(\mathbf{R^t})) \text{ or}$$

$$p(c^t|\alpha_{ppt}, \beta_{ppt};\ \mathbf{R^t}) = \sigma((2c^t - 1)\, \Delta U_{\alpha_{ppt},\beta_{ppt}}(\mathbf{R^t}))$$

$$\Delta U_{\alpha_{ppt},\beta_{ppt}}(\mathbf{R^t}) = U_{\alpha_{ppt},\beta_{ppt}}(\mathbf{R^{t;1}}) - U_{\alpha_{ppt},\beta_{ppt}}(\mathbf{R^{t;2}})$$

$$LL = \log\left(\, p(c^t|\alpha_{ppt}, \beta_{ppt};\ \mathbf{R^t})\, \right)$$

2.4.2. Phase 2 – how participants learned about their partners' social preferences

We then modelled participants' beliefs about their partner's SVO as ranging along two dimensions, α_{par} & β_{par}.

Over 36 trials, participants made binary predictions $\hat{d}^t$, t = {1…T} about whether option 1 or option 2 would be chosen by their partner given the returns $\mathbf{R}^t$=$\{\mathbf{R}^{t;1}; \mathbf{R}^{t;2}\}$ = $\{R^{t;1}_{\text{self}}, R^{t;1}_{\text{other}};\ R^{t;2}_{\text{self}}, R^{t;2}_{\text{other}}\}$ for each pair of options. They then discovered what the partner chose, which we write as d^t.

The participant assumed that the partner chooses the same way that they do themselves, but with SVO parameters $\alpha_{par}, \beta_{par}$, which they needed to infer from observation. That is, the likelihood that the partner chose d^t is $LL = \log\left(\, p(d^t|\alpha_{par}\beta_{par};\ \mathbf{R}^t)\,\right)$ using the same formula as in equation 1.

The partner's decisions $D^t = \{d^1, d^2, \ldots, d^t\}$ were used to update a participant's beliefs about a partner's $\alpha_{par}, \beta_{par}$, written as $p(\alpha_{par}, \beta_{par}|D^t)$. The starting point for these beliefs (written as $p\left(\alpha_{par}, \beta_{par} \mid D^0\right)$ was the participant's prior. For this model, we assumed that this was a factorised distribution with each parameter centred on the participant's own preferences α^m_{ppt}, β^m_{ppt} but with standard deviation parameters α^σ and β^σ that characterised the extent to which the participant thought their partner might differ from themselves (belief flexibility). Therefore, we have

(3)

$$p\left(\alpha_{par}|D^0\right) \sim N\left(\alpha_{par}; \alpha^m_{ppt}, \alpha^\sigma\right)$$

$$p(\beta_{par}|D^0) \sim N\left(\beta_{par}; \beta^m_{ppt}, \beta^\sigma\right)$$

$$p\left(\alpha_{par}, \beta_{par} \mid D^0\right) = p\left(\alpha_{par}|D^0\right)\, p\left(\beta_{par}|D^0\right)$$

Where equation 3 signifies that the independent probability of α_{par} and β_{par} are predicated on a normal density distribution over all possible values of α_{par} and β_{par}, determined by the participant priors ($\alpha^m_{ppt}, \alpha^\sigma$ with regard to α_{par}, and $\beta^m_{ppt}, \beta^\sigma$ with regard to β_{par}). We then assumed that a participant's posterior beliefs about their partner from trials t = 1…36 given a partner's decisions followed Bayes rule:

(4)

$$p\left(\alpha_{par}, \beta_{par} \,\middle|\, D^t\right) = \frac{p(d^t|\alpha_{par}, \beta_{par};\ \mathbf{R}^t)\, p\left(\alpha_{par}, \beta_{par}|D^{t-1}\right)}{p(d^t|D^{t-1})}$$

For efficiency, we conveniently represented $p\left(\alpha_{par}, \beta_{par} \,\middle|\, D^t\right)$ as a matrix over a fixed grid of α and β values, $\theta^t_{\alpha_{par},\beta_{par}}$. We could then calculate the participant's beliefs about their partner's SVO preferences for each trial:

(5)

$$\theta^t_{\alpha_{par},\beta_{par}} = \frac{p(d^t|\alpha_{par}, \beta_{par};\ \mathbf{R}^t)\,\theta^{t-1}_{\alpha_{par},\beta_{par}}}{\sum_{\alpha'_{par},\beta'_{par}} p(d^t|\alpha_{par}, \beta_{par};\ \mathbf{R}^t)\,\theta^{t-1}_{\alpha'_{par},\beta'_{par}}}$$

We could then marginalise along $\theta^t_{\alpha_{par},\beta_{par}}$ to calculate the belief a participant had over their partner's SVO:

(6)

$$p(\alpha_{par} \mid D^t) = \sum_{\beta_{par}} \theta^t_{\alpha_{par},\beta_{par}}$$

$$p(\beta_{par} \mid D^t) = \sum_{\alpha_{par}} \theta^t_{\alpha_{par},\beta_{par}}$$

The model then stated that the participant predicts the partner's decision in the next trial by calculating the probability determined by the utility differences $\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^{t+1}})$ as in equation (1), summed over the joint distribution $\theta^t_{\alpha_{par},\beta_{par}}$ over the partner's parameters

(7)

$$p(d^{t+1} = 1 \mid D^t; \mathbf{R^{t+1}}) = \sum_{\alpha_{par},\beta_{par}} \theta^t_{\alpha_{par},\beta_{par}} \cdot \sigma(\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^{t+1}}))$$

$$p(d^{t+1} = 2 \mid D^t; \mathbf{R^{t+1}}) = 1 - p(d^{t+1} = 1 \mid D^t; \mathbf{R^{t+1}})$$

and then performed probability matching, so that

$p(\hat{d}^{t+1} = 1 \mid D^t; \mathbf{R^{t+1}}) = p(d^{t+1} = 1 \mid D^t; \mathbf{R^{t+1}})$.

2.5. Updates in our inferences about the beliefs of participants about their partners

Updates between the prior and posterior distribution of our inferences about what a participant believes about their partner were calculated as the absolute difference between the mean of their prior at the start (trial 0) of phase 2 (which, according to the winning model, came from their own value) and the mean posterior approximation of the participant's belief about a partner along each dimension at trial 36 of phase 2, weighted by the baseline similarity of the participant and their partner, i.e. the number of the same decisions a participant and their partner would have made over Phase 2 choices had no learning occurred:

$$\Delta(\beta^m_{par}) = \mid \hat{\beta}^m_{par} - \beta^m_{ppt} \mid \cdot \frac{1}{Baseline\ Similarity + 1}$$

$$\Delta(\alpha^m_{par}) = \mid \hat{\alpha}^m_{par} - \alpha^m_{ppt} \mid \cdot \frac{1}{Baseline\ Similarity + 1}$$

3. Results

3.1. *Participants*

We recruited 697 participants via Prolific (66% identified as female). Paranoia (measured by scale B of the Revised Green Paranoid Thoughts Scale, R-GPTS-B; Freeman et al., 2020) was highly skewed to the left and low, although it covered the entire spectrum (mean [sd] = 3.83 [6.07], median = 1, skew = 2.23, range = 0-33). General cognitive ability was approximated using the International Cognitive Ability Resource progressive matrices scores (ICAR; ICAR Team, 2014) and was approximately normally distributed (mean [sd] = 4.95 [2.38], median = 5, skew = 0.09, range = 0-10).

3.2. Model-agnostic Analysis

3.2.1. *Phase 1*

When given prosocial-competitive choices, participants mostly made prosocial choices (mean[sd] = 84.2%, [29.9]). When given competitive-individualistic choices, participants mostly made individualistic choices (mean[sd] = 90.3% [22.1]). When given prosocial-individualistic/competitive choices, participants mostly made prosocial choices (mean[sd] = 55.3% [39.7]).

More paranoid people made fewer prosocial choices, both in the prosocial-competitive choice pairs (non-averaged estimate = -0.08, 95%CI: -0.10, -0.06; only one model was supplied in averaging and so a non-averaged linear model was used; Model 1a), and in the prosocial-individualistic/competitive choice pairs (estimate = -0.07, 95%CI: -0.09, -0.05; Model 1b). Paranoia was not associated with preferences during individualistic-competitive choice pairs. When faced with prosocial-competitive choice pairs, people with higher general cognitive ability made more prosocial choices (estimate = 0.10, 95%CI: 0.08, 0.11; Model 1a). People with higher general cognitive ability made more individualistic/competitive choices when traded off against prosocial decisions (estimate = -0.11, 95%CI: -0.13, -0.09; Model 1b), and more individualistic choices when traded off against competitive decisions (estimate = 0.10, 95%CI: 0.08, 0.11; Model 1c).

3.2.2. Phase 2

3.2.2.1. Predictive Accuracy

A participant's accuracy was quantified as the total number of correct predictions (out of 36) they made about their partner. Participants correctly predicted on average 30.2 [sd: 4.98, range = 1-36] trials, indicating high accuracy. Accuracy was highest for participants assigned to prosocial partners (mean [sd]: 33.4 [3.56], range = [12, 36]); then competitive partners (30.6 [4.58], range = [1, 36]); and lowest for individualistic partners (26.6 [4.11], range = [11, 32]). Accuracy was greater for prosocial than individualistic (estimate: -1.39, 95%CI: -1.53, -1.24; Model 2a) or competitive (estimate: -0.56, 95%CI: -0.72, -0.42; Figure 2A; Model 2a) partners; and higher for competitive compared to individualistic partners (estimate: 0.82, 95%CI: 0.67, 0.97; Model 2b).

There was no association between paranoia and accuracy (estimate: -0.04, 95%CI: -0.13, 0.02; Figure 2B; Model 3a). General cognitive ability was associated with increased accuracy (estimate: 0.07, 95%CI: 0.00, 0.15; Model 3a), and specifically associated with increased accuracy for predicting individualistic partner behaviour (estimate: 0.19, 95%CI: 0.06, 0.30; Model 3b) and prosocial partner behaviour (estimate: 0.19, 95%CI: 0.06, 0.32; Model 3c), but not the behaviour of competitive partners (Figure 2B; Model 3d).

When asked to explicitly classify their partners according to prosocial, individualistic, or competitive preferences following phase 2 predictions, participants were generally accurate: 71 % of people correctly identified competitive partners as trying to stop them earning money; 84 % correctly identified individualistic partners as trying to earn as much money as possible; and 93 % correctly identified prosocial partners as trying to share payoffs as equally as possible (Figure C.1). Kruskal-Wallis rank sum test found that those who were more paranoid significantly (although marginally) differed in their classification of prosocial partners (Kruskal-Wallis $\chi^2(2) = 6.00$; $p = 0.0497$). To explore this further, pairwise comparisons using Dunn test with Benjamini-Hochberg correction found a significant difference between competitive and prosocial classifications (Dunn estimate = -2.32, $p = 0.03$), such that more paranoid individuals were more likely to classify prosocial partners as trying to stop them earning money rather classifying them as trying to share the money equally.

3.2.2.2. Intention Attributions

Participants attributed more harmful intent to competitive (estimate: 1.87, 95%CI: 1.75, 1.99; Model 4a) and individualistic (estimate: 1.03, 95%CI: 0.91, 1.14; Model 4a) partners than to prosocial partners; and harmful intent attributions were stronger for competitive than for individualistic partners (estimate: 0.84, 95%CI: 0.73, 0.96; see Figure 2C; Model 4a), as

expected. Self-interest attributions also varied with partner SVO. Participants attributed higher self-interest to individualistic (estimate: 1.38, 95%CI: 1.18, 1.57; Model 4b) and competitive (estimate: 0.57, 95%CI: 0.40, 0.73; Model 4b) partners than to prosocial partners; and individualistic partners were perceived as being more self-interested than competitive partners (estimate: -0.67, 95%CI: -0.83, -0.51; see Figure 2C; Model 4b). As expected, intention attribution also varied with paranoia: more paranoid participants attributed more harmful intent to partners (estimate: 0.07, 95%CI: 0.02, 0.11; Model 4a), but did not make stronger self-interest attributions (Model 4b).

When paired with prosocial partners, people who made stronger harmful intent attributions were less accurate at predicting the partner's behaviour (estimate: -0.31, 95%CI: -0.43, -0.19; Model 4c). Conversely, when paired with competitive partners, people who made stronger harmful intent attributions were more accurate at predicting the partner's behaviour (estimate: 0.19, 95%CI: 0.06, 0.31; Model 4e). When paired with individualistic partners, people who made stronger self-interest attributions were more accurate, as expected (estimate: 0.58, 95%CI: 0.47, 0.68; Figure 2D; Model 4g).

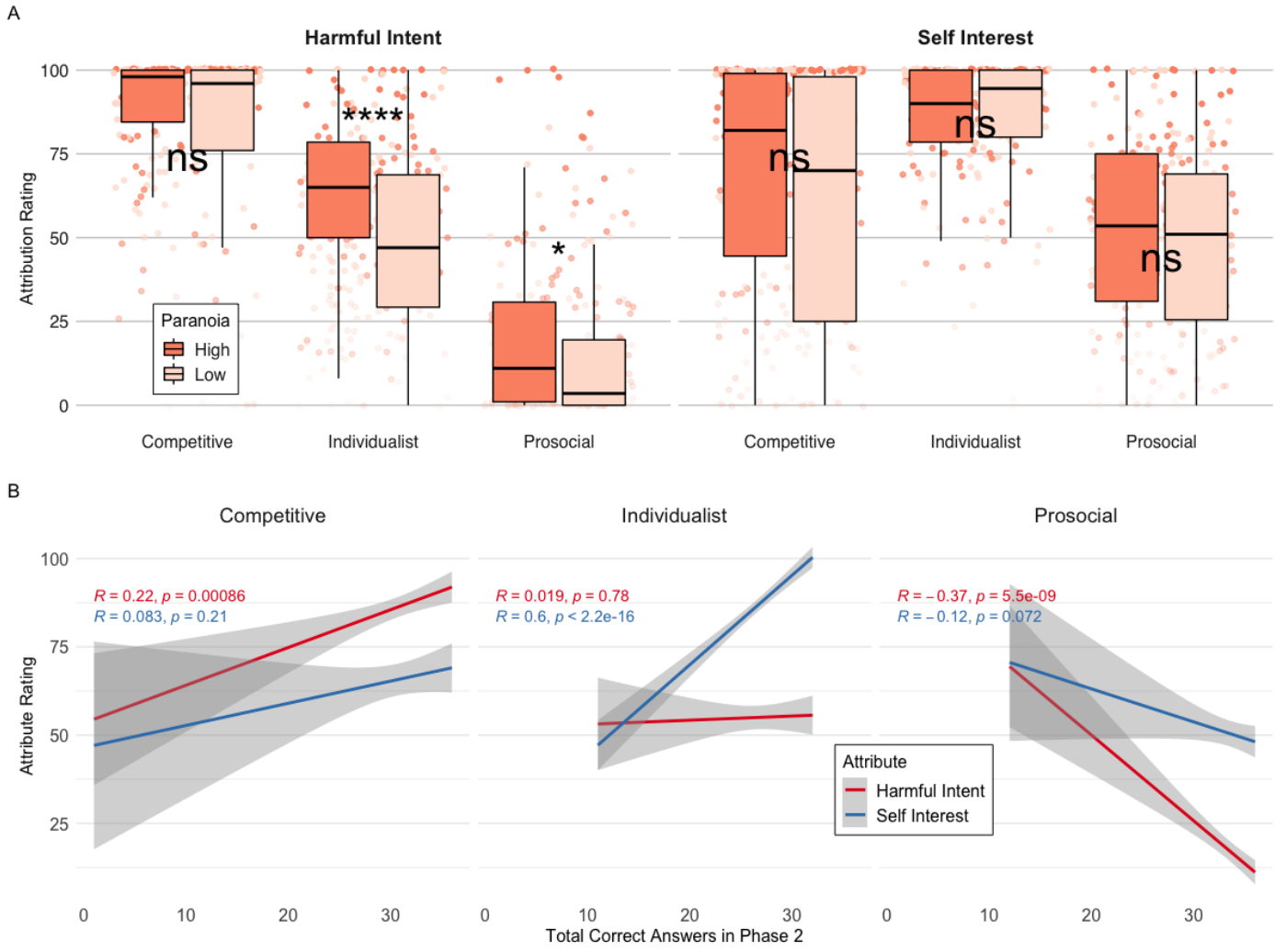


Figure 2. Model agnostic analysis of phase 2.

(A) Harmful intent and self-interest attributions made at the end of the task for each level of paranoia – determined via split mean (median = 1) – for each partner policy. Linear modelling was used for the main analysis and confirmed the split-mean comparison differences. Split mean differences calculated for this visualisation using non-parametric ANOVA. ns = not significant, * = p < 0.05, ** = p < 0.01 (B) Pearson correlations between total correct answers and attributions made about the partner at the end of phase 2. These associations were confirmed with more complex models in the text controlling for age, sex, task comprehension, general cognitive ability, and paranoia.

3.3. Computational Modelling

We found that a four-parameter Bayesian updating model fitted the data best (Figure 3A). The model estimates a participant's own $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$ from the preferences they express in trials 1-18 and characterises the participant as performing Bayesian inference about the α_{par} and β_{par} of the partner they observe over the following 36 trials. In the model, a participant's own SVO (characterised by their $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$) forms the central tendency of their prior beliefs about a partner's SVO, and the participant's willingness to update these beliefs in the light of the partner's choices is governed by standard deviation terms α^{σ} and β^{σ} which parameterize the width of these priors (Figure 3B & 3C). Concretely, the participant's starting priors for their partner are Gaussian with means $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$, and standard deviations α^{σ} and β^{σ}. The model well predicts the trial-by-trial responses of the participants and overall, for each participant (Figure 3D & E), was able to reproduce behaviour (Figure 3G) and had parameters that could be appropriately recovered (Figure 3H). The model fits $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$ to data from phase 1 and phase 2; see Table A.1 for the relationship to the fit of parameters estimated from just phase 1 and phase 2 separately.

All alternate models are defined in the Appendix (Table A.1; Formalisms: Text S1; Text S2). The best-performing model fit better than similar models (inspired by Tarantola, 2017) in which the participants own preferences exerted a persistent bias over choices as well as affecting the initial condition. See Appendix (Figure A.1) for the generative performance of the model across different participant and partner SVO types.

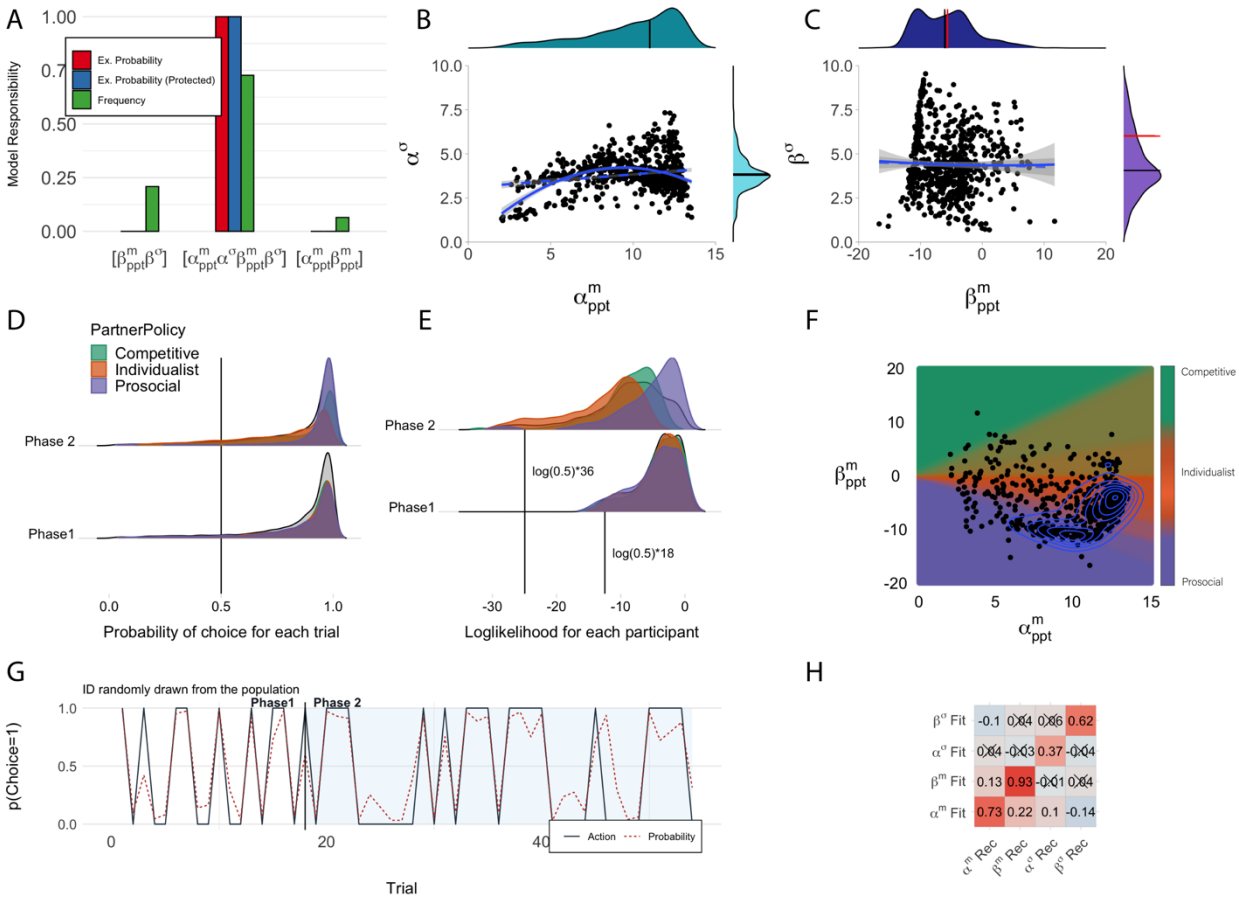


Figure 3: Winning model summary statistics.

Model comparison across competing models with responsibility greater than 1% for the population. The best-fitting model integrated a participant's preferences into their beliefs about a partner and therefore generated congruency between the two. Models that only parameterised relative payoff (β; for both participant and partner) or only considered the participant's α^{m}_{ppt} and β^{m}_{ppt} accounted for 22% and 7% of the population, respectively. (B) Relationship between α^{m}_{ppt} and α^{σ} with their marginal distributions. Black lines within marginals denote the hierarchical group mean. Blue dashed line is the linear relationship between central tendencies and standard deviation parameters. Blue solid line is the quadratic relationship between central tendencies and standard deviation parameters. (C) Relationship between β^{m}_{ppt} and β^{σ} with their marginal distributions. Black lines within marginals denote the hierarchical group mean. Red lines denote the hierarchical group means of the beta-only model; β^{m}_{ppt}; β^{σ} mean estimates [±sd] were -5.63[7.80] and 6.01[2.64]. Blue dashed line is the linear relationship between central tendencies and standard deviation parameters. Blue solid line is the quadratic relationship between central tendencies and standard deviation parameters. (D) Trial-wise simulated probabilities across the population for each partner policy. (E) Phase-wise sum loglikelihood. Lines delineate the cut off where the model is predicting a participant at chance level (50%). (F) Relationship between α^{m}_{ppt} and β^{m}_{ppt}. Background colours denote the summed number of simulated correct model predictions of being competitive, individualist, or prosocial given the full spectrum of possible α^{m}_{ppt} and β^{m}_{ppt} used in our study. Black dots show estimated participant parameters of our sample. Contours represent the 2D density distribution of parameter estimates across our population. (G) Posterior predictive check on a random participant drawn from the population for choices across both phase 1 and phase 2. (H) Recovery analysis for the winning model. X = non-significant correlation.

3.4. Model-based Analysis

3.4.1. Phase 1

We used a hierarchical Bayesian fit for the four parameters of the model, with one level for the population and a second for the participants. At the population level, participants preferred higher absolute payoffs ($\alpha^{\mathrm{m}}_{\mathrm{ppt}}$ = 11.13[0.05]; for mean [±standard error]) and more equal relative payoffs ($\beta^{\mathrm{m}}_{\mathrm{ppt}}$ = -6.08[0.20]; Figure 3F). That $\alpha^{\mathrm{m}}_{\mathrm{ppt}}$ is larger than $-\beta^{\mathrm{m}}_{\mathrm{ppt}}$ (t test, t(1157.7) = 77.6, p < 0.001) implies that participants on average valued returns to themselves more than they valued equality. General cognition was positively associated with participants' preferences for higher absolute payoffs ($\alpha^{\mathrm{m}}_{\mathrm{ppt}}$; estimate = 0.22, 95%CI: 0.15, 0.29; Model 5a), and older people exhibited a reduced preference for absolute payoffs (estimate = -0.09, 95 % CI: -0.16, -0.02; Table C.1; Model 5a). However, there was no link between paranoia and $\alpha^{\mathrm{m}}_{\mathrm{ppt}}$ (Table C.1; Model 5a). Paranoia was positively associated with $\beta^{\mathrm{m}}_{\mathrm{ppt}}$ (estimate = 0.09, 95%CI: 0.01, 0.16; Model 5b), indicating participants' preference for options in which they earned more than their partners. Participants who identified as female were also found to be less competitive compared to those identifying as male (estimate = -0.16, 95%CI: -0.32, -0.01; Model 5b). There was no effect of general cognition or age on social preferences for relative payoffs (Table C.1; Model 5b).

3.4.2. Phase 2

3.4.2.1. Predictive Accuracy

We asked what affected a participant's ability to predict the social preferences of their partner. Of course, if a participant's own values of $\alpha^{\mathrm{m}}_{\mathrm{ppt}}$ and $\beta^{\mathrm{m}}_{\mathrm{ppt}}$ were such that they would make the same choices as their partner, prediction would be particularly straightforward. Thus, we calculated how many answers a participant *would* have gotten correct based on their own parameter values, and included it as a regressor in our models. This approximated baseline similarity between participant and partner in the absence of learning (see Figure D.1 for the distrubution of baseline similarity). When $\alpha^{\mathrm{m}}_{\mathrm{ppt}}$ and $\beta^{\mathrm{m}}_{\mathrm{ppt}}$ specify preferences that differ from those of the partner, accurate predictions depend on the participant being willing to entertain the possibility that such a difference might exist (represented as higher α^{σ} and β^{σ}). We therefore also include these terms in the model exploring predictive accuracy. Belief flexibility terms, α^{σ} and β^{σ}, were positively associated with each other (estimate = 0.35, 95%CI: 0.28, 0.42; Model 6).

Predictive accuracy was positively associated with baseline similarity and with belief flexibility parameters, α^{σ} and β^{σ} (Table D.1; Model 7). Neither paranoia nor general cognitive ability in this model were associated with predictive accuracy (Model 7).

To unpack these relationships: initially regressing predictive accuracy against general cognition, age, sex, and task comprehension generated an r^2 of 0.03 (p<0.001). Including paranoia did not improve the model (r^2 = 0.03, p < 0.001; F = 2.22, p = 0.14). Including baseline similarity significantly improved the model (r^2 = 0.07; F = 27.8, p < 0.001), indicating that similarity between participants and their partners was associated with increased accuracy (non-averaged estimate: 0.19, 95%CI: 0.12, 0.27). Including α^{σ} and β^{σ} significantly improved the model (r^2 = 0.52, p < 0.001; F = 319.8, p < 0.001), with both being positively associated with predictive accuracy (α^{σ} non-averaged estimate: 0.39, 95%CI: 0.34, 0.45; β^{σ} non-averaged estimate: 0.56, 95%CI: 0.50, 0.62). We also allowed for α^{σ} , β^{σ} and baseline similarity to interact in a final model; this significantly improved the model (r^2 = 0.61, p < 0.001; F = 37.7, p < 0.001). In this final model there was an interaction between baseline similarity and β^{σ} (non-averaged estimate: -0.31, 95%CI: -0.37, -0.25), as well as baseline similarity and α^{σ} (0.12, 95%CI: 0.04, 0.19).

When predicting general cognitive ability and paranoia in two separate models, general cognitive ability was not associated with either α^{σ} or β^{σ} (Model 8a), although paranoia was negatively associated with β^{σ} (estimate: -0.06, 95%CI: -0.15, -0.00; Model 8b), after controlling for participant-partner baseline similarity, age, sex, and task comprehension. There was no interaction between α^{σ}, β^{σ} and baseline similarity in either model. This suggests that paranoia was specifically associated with increased belief rigidity concerning the value a partner placed on relative (rather than absolute) payoffs.

3.4.2.2. Inferential Updating

We explored the change in the statistical inferences made about the beliefs participants held about their partners by testing the difference in the mean values of inferred distributions before and after a participant learnt about their partner normalised by their similarity prior to learning [$\Delta(\alpha_{par}^{m})$; $\Delta(\beta_{par}^{m})$], and the impact of paranoia on this process (see Table F.1 for summary statistics).

We observed significantly larger $\Delta\left(\alpha_{par}^{m}\right)$ and $\Delta(\beta_{par}^{m})$ in competitive partner conditions compared to individualist and prosocial partner conditions (see Table F.1). Pearson

correlations identified that similarity between a participant and partner at baseline was negatively associated with $\Delta(\alpha_{par}^{m})$ (r = -0.56, 95%CI: -0.61, -0.50) and $\Delta(\beta_{par}^{m})$ (r = -0.66, 95%CI: -0.70, -0.62).

There was no association between paranoia and $\Delta(\alpha_{par}^{m})$ (-0.05, 95%CI: 0.19, 0.42; Aux Model 1), and there was no interaction between paranoia and partner after controlling for age, sex, general cognition and task comprehension. Paranoia was associated with lower $\Delta(\beta_{par}^{m})$ across the board (-0.16, 95%CI: -0.27, -0.05; Aux Model 2), and there was an interaction between paranoia and partner such that those more paranoid changed their beliefs more with prosocial versus competitive partners (0.18, 95%CI: 0.02, 0.33; Aux Model 2) and more with individualist versus competitive partners (0.17, 95%CI: 0.02, 0.32; Aux Model 2), although no difference between prosocial and individualist partners, after controlling for age, sex, general cognition and task comprehension.

3.4.2.3. Intention Attributions

Paranoia and α^{σ} were both positively associated with harmful intent attributions, whereas baseline similarity was negatively associated with harmful intent attributions (Table E.1; Model 9a). Predictive accuracy, general cognition and β^{σ} were not associated with harmful intent attributions (Model 9a).

Regressing attributions of harmful intent against partner policy, general cognition, age, sex, and task comprehension generated an r^2 of 0.584 (p<0.001). Including predictive accuracy in this equation offered no additional explanatory power (r^2 = 0.584, p<0.001; F = 0.09, p = 0.76). Including paranoia significantly improved the model (r^2 = 0.588, p<0.001; F = 6.05, p< 0.014): paranoia was positively associated with harmful intent (non-averaged estimate = 0.06, 95%CI: 0.01, 0.11). Including baseline similarity also significantly improved the model (r^2 = 0.599, p<0.001; F = 18.26, p<0.001); and baseline similarity was negatively associated with harmful intent (non-averaged estimate: -0.17, 95%CI: -0.25, -0.09). Including α^{σ} or β^{σ} significantly improved the model (r^2= 0.603, p<0.001; F = 3.66, p = 0.03), although only α^{σ} (whose recovery is least accurate; Figure 3H) was positively associated with harmful intent (non-averaged estimate: 0.08, 95%CI: 0.01, 0.14).

We found a relationship between $\Delta(\alpha_{par}^{m})$ and harmful intent attributions, such that the more inferences about participants' beliefs about their partner moved away from $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$, the larger the attributions of harmful intent estimated by the participant (0.31, 95%CI: 0.24 ,0.38; Aux

Model 3a), after controlling for general cognition, age, sex, and task comprehension. The same was also true for $\Delta(\beta_{par}^{m})$ - the further participants needed to move away from their $\beta_{\mathrm{ppt}}^{\mathrm{m}}$ the larger their attributions of harmful intent (0.37, 95%CI: 0.30, 0.44; Aux Model 3b).

Self-interest attributions were positively associated with predictive accuracy and negatively associated with β^{σ} (Table E.1; Model 9b). Paranoia, general cognition and α^{σ} were not associated with self interest attributions (Model 9b).

Regressing attributions of self-interest against partner policy, general cognition, age, sex, and task comprehension generated an r^2 of 0.231 (p<0.001). Including predictive accuracy significantly improved the model's explanatory power (r^2 = 0.248, p<0.001; F = 15.75, p<0.001). Specifically, predictive accuracy was positively associated with self-interest (non-averaged estimate: 0.17, 95%CI: 0.08, 0.25). Including α^{σ} or β^{σ} significantly improved the model (r^2 = 0.267, p < 0.001; F = 8.66, p < 0.001), although only β^{σ} was negatively associated with self-interest attributions (non-averaged estimate: -0.19, 95%CI: -0 .29, -0.10). Neither paranoia (r^2 = 0.24, p<0.001; F = 0.5, p=0.48) nor baseline similarity significantly improved the model's explanatory power (r^2 = 0.249, p < 0.001; F = 0.04, p=0.85).

We found no relationship between $\Delta(\alpha_{par}^{m})$ or $\Delta(\beta_{par}^{m})$ and self-interest attributions (Aux Model 4a; Aux Model 4b), after controlling for general cognition, age, sex, and task comprehension.

4. Discussion

How do people learn about others when given little information? Core social-cognitive theory (Anderson & Chen, 2002; Anderson & Glassman, 1996) suggests that they use themselves as a starting point. Previous experimental evidence suggests that the neural regions participants use to estimate the choices others will make are like those they use to make choices for themselves (Behrens et al., 2008; Nicolle et al., 2012), with the right temporal parietal junction (Zhang & Glascher, 2020) and anterior cingulate gyrus (Chang et al., 2013) implicated in the integration of self and other choice information. We explored this hypothesis by asking whether participants' own social preferences influenced the way they learned about the social preferences of others, and how both traits were associated with paranoia. We found that people used their preferences as a prior for learning about others and that they were more accurate at predicting the choices of partners who had similar social preferences to them. This accuracy could not be attributed to baseline participant-partner similarity, nor willingness to adapt. Similarity between the preferences of participants and their partners reduced harmful intent attributions. More paranoid individuals were less flexible in adapting to a partner's prosocial or competitive nature, supporting the idea that paranoia involves alterations to key social computational processes that update representations of interaction partners.

Participants' predictions about their partners' social preferences were best fit by a Bayesian learning model incorporating their preferences as the central tendency of their priors about their partners, and parameters governing the flexibility of these beliefs. This model outperformed alternatives in which the participants' own preferences played either no, or a diminished, role in modelling their partners' preferences, and variants (based on Tarantola et al., 2017) in which participants' own preferences persistently influenced predictions (for instance, predicting options by ignoring their partner's social preferences, or predicting options based on the best outcome for themselves). Unlike Tarantola et al. (2017), the money received by participants here depended on their partner's choices which may have influenced the salience of the predictions. Alternatively, participants' preferences for snack foods could be more pronounced than for the subtler social value choices, implying that they exerted a more substantial effect. The Bayesian model also outperformed heuristic alternatives in which participants learned that their partners made choices in one of the simple three categories associated with SVO. This suggests that the nuance afforded by the full social preference model is beneficial.

Together, our data and model provide mechanistic insights into how interpersonal similarity affects social interaction. Participants perceived partners who were more like themselves (including more competitive ones) as being less intentionally harmful. This was an unanticipated result that should ideally be replicated before we draw firm conclusions, although it is consistent with prior theory. For instance, behavioural or psychological similarity to a social partner can foster social bonding and perception of friendship quality (Redcay & Schilbach, 2019, Bolis et al. 2021). These empirical findings are framed within the 'dialectical misattunement hypothesis' (Bolis et al., 2017): mismatch between two individuals through communication misalignment and unpredictability of action reduces the efficiency of information transfer and bonding and may increase distrust in social relationships. In one virtual reality study, unfamiliarity, feeling out of place and feeling like an outsider increased participants' perceptions of being judged by avatars, as well as increased perceptions of these avatars aiming to cause emotional distress (Riches et al., 2020). Likewise, epidemiological work has observed increased psychosis risk in people who are marginalised, and 'othered' in a community (el Bouhaddani et al., 2019; Kirkbride et al., 2017). Here we present evidence and a formal model that suggests the similarity between a participant and partner in non-clinical populations facilitates more precise predictions about a partner which may, in turn, reduce attributions of harmful intent.

Another possibility is that this relationship stemmed from participants' failure to recognise when their own decisions reflected harmful intent (specifically, when making choices that caused a larger discrepancy in earnings between themselves and their partner). Unfortunately, participants did not provide self-assessments of their intentions, so we cannot test whether participants recognised that their own competitive decisions reflected spiteful (and potentially harmful) motives. This would be fruitful to explore in any future work.

Paranoia was positively associated with preferences for more unequal payoffs in Phase 1, and with less flexible predictions and belief updating about a partner's preferences for relative payoffs in Phase 2, regardless of the baseline similarity. The relationship between paranoia and competitive preferences is consistent with prior work (Raihani et al., 2021; Raihani & Bell, 2018; Schaerer et al., 2021). Our observations of belief rigidity and less social belief updating in paranoia is also consistent with previous experimental evidence in those with schizophrenia and borderline personality disorder (Diaconescu et al., 2020; Henco et al., 2020; Wellstein et al., 2020), and suggests that this reduction in belief flexibility may specifically impinge upon the ability to form and update stable beliefs about the partner's preferences for equality or inequality. Interestingly, as more paranoid people were more likely to make competitive decisions in Phase 1 and attributed more harmful intent to

their partners in Phase 2 (replicating Barnby et al. 2020a; Greenburgh et al. 2019; Raihani & Bell 2017, Saalfeld et al. 2018), the increased baseline similarity should have helped participants be more accurate in predicting the decisions of competitive partners in Phase 2. Nevertheless, more paranoid individuals were not more accurate in predicting the decisions of competitive partners, which implicates reduced belief flexibility as a potential cause. It may be that the degree of similarity between more paranoid participants and competitive partners was insufficient to overcome their reduced β^{σ}. This combination of participant-partner similarity and belief flexibility should be explored in future work. Given that autistic-like traits have also shown associations with a reduction in social information processing (Sevgi et al, 2020) it is also important to measure autistic-like traits in the future use of this task to understand the unique contribution of pre-existing paranoia. Our and prior data are consistent with claims that positive symptoms of psychosis (which frequently involve paranoia) may stem from a general inability to reconcile incoming information with current predictions (Fletcher & Frith, 2009), and this may be particularly applicable in social contexts (Bolis & Schilbach, 2017).

We note two main limitations. First, we recruited a sample that exclusively resided in the UK. While our sample was diverse in age, self-reported ethnicity, and sex, given the disparity in neuroimaging evidence between how familiar and unfamiliar individuals are represented in the brain (Ng et al., 2010; Zhu et al., 2007) it might be that the model for representing social others will transfer cross-culturally, although the degree to which our beliefs and values are integrated into our priors over others may vary. Second, we focussed on a particular type of prosocial behaviour, where people show preferences for equal outcomes rather than preferring the partner to earn more than themselves. Meta-analysis suggests that different types of prosocial behaviour cluster together and are implemented with different subtleties in neural activation maps (Rhoads et al., 2021), and therefore it is unclear whether different forms of prosocial behaviour would produce similar learning effects, or perhaps show more specific persistent choice biases as reported in Tarantola et al., (2017).

In sum, we used an SVO task that involved real financial incentives to examine whether and how people use their own social preferences to learn about the social preferences of others. Consistent with accounts of social learning, we found that people used their own social preferences as a prior for learning about the social preferences of a partner and that impressions were updated in a Bayesian manner. More paranoid participants held more rigid beliefs about the partner's preference for inequality, updating their posterior beliefs less across the board. Finally, participants adopted a relative notion of harm, rating choices consistent with their own preferences as being less intentionally harmful.

CREDIT

Conceptualisation: JMB, NR & PD. Data Curation: JMB. Formal Analysis: JMB & PD. Funding Acquisition: JMB. Investigation: JMB. Methodology: JMB, NR & PD. Project Administration: JMB. Resources: JMB. Software: JMB. Supervision: NR & PD. Validation: JMB & PD. Visualisation: JMB. Writing – Original Draft: JMB. Writing – Review & Editing: JMB, NR & PD.

Conflict of Interest

None to declare.

Funding

JMB was supported by the UK Medical Research Council (MR/N013700/1) and King's College London member of the MRC Doctoral Training Partnership in Biomedical Sciences for this work. PD is supported by the Max Planck Society and the Alexander von Humboldt foundation.

Data and Code Availability

All data, model code, analysis scripts, and R-Markdown workbook to reproduce the regression analyses are available on GitHub: https://github.com/josephmbarnby/Barnby_etal_2021_SVO.

5. References

Andersen, S. M., & Chen, S. (2002). The relational self: an interpersonal social-cognitive theory. *Psychological review*, *109*(4), 619.

Andersen, S. M., & Glassman, N. S. (1996). Responding to significant others when they are not there: Effects on interpersonal inference, motivation, and affect.

Barnby, J. M., Bell, V., Mehta, M. A., & Moutoussis, M. (2020a). Reduction in social learning and increased policy uncertainty about harmful intent is associated with pre-existing paranoid beliefs: Evidence from modelling a modified serial dictator game. *PloS computational biology*, *16*(10), e1008372.

Barnby, J. M., Deeley, Q., Robinson, O., Raihani, N., Bell, V., & Mehta, M. A. (2020b). Paranoia, sensitization, and social inference: Findings from two large-scale, multi-round behavioural experiments. *Royal Society open science*, *7*(3), 191525.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. Doi:10.18637/jss.v067.i01.

Bartoń, K. (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245-249.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2017). Beyond autism: Introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology*, *50*(6), 355-372.

Bolis, D., Lahnakoski, J. M., Seidel, D., Tamm, J., & Schilbach, L. (2021). Interpersonal similarity of autistic traits predicts friendship quality. *Social cognitive and affective neuroscience*, *16*(1-2), 222-231.

El Bouhaddani, S., van Domburgh, L., Schaefer, B., Doreleijers, T. A., & Veling, W. (2019). Psychotic experiences among ethnic majority and minority adolescents and the role of discrimination and ethnic identity. *Social psychiatry and psychiatric epidemiology*, *54*(3), 343-353.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in cognitive sciences*, *11*(2), 49-57.

Burnham, K. P., & Anderson, D. R. (1998). Practical use of the information-theoretic approach. In *Model selection and inference* (pp. 75-117). Springer, New York, NY.

Burnham, K. P., & Anderson, D. R. (2002). A practical information-theoretic approach. *Model selection and multimodal inference*, *2*, 70-71.

Chang, S. W., Gariépy, J. F., & Platt, M. L. (2013). Neuronal reference frames for social decisions in primate frontal cortex. Nature neuroscience, 16(2), 243-250.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, *10*(7), 287-291.

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1379-1390.

Diaconescu, A. O., Wellstein, K. V., Kasper, L., Mathys, C., & Stephan, K. E. (2020). Hierarchical Bayesian models of social inference for probing persecutory delusional ideation. *Journal of Abnormal Psychology*, *129*(6), 556.

Fett, A. K. J., Shergill, S. S., Joyce, D. W., Riedl, A., Strobel, M., Gromann, P. M., & Krabbendam, L. (2012). To trust or not to trust: the dynamics of social interaction in psychosis. *Brain*, *135*(3), 976-984.

Freeman, D., & Garety, P. A. (2000). Comments on the content of persecutory delusions: does the definition need clarification?. *British Journal of Clinical Psychology*, *39*(4), 407-414.

Freeman, D., Loe, B. S., Kingdon, D., Startup, H., Molodynski, A., Rosebrock, L., … & Bird, J. C. (2021). The revised Green et al., Paranoid Thoughts Scale (R-GPTS): psychometric properties, severity ranges, and clinical cut-offs. *Psychological Medicine*, *51*(2), 244-253.

Hertz, U. (2021). Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms. *Proceedings of the Royal Society B*, *288*(1952), 20210293.

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS computational biology*, *11*(6), e1004254.

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLoS computational biology*, *14*(2), e1005935.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS computational biology*, *7*(4), e1002028.

Jaya, E. S., Hillmann, T. E., Reininger, K. M., Gollwitzer, A., & Lincoln, T. M. (2017). Loneliness and psychotic symptoms: The mediating role of depression. *Cognitive therapy and research*, *41*(1), 106-116.

Jones, E. E. & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior.* New York: General Learning Press.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *science*, *321*(5890), 806-810.

Kirkbride, J. B., Hameed, Y., Ioannidis, K., Ankireddypalli, G., Crane, C. M., Nasir, M., ... & Jones, P. B. (2017). Ethnic minority status, age-at-immigration and psychosis risk in rural environments: evidence from the SEPEA study. *Schizophrenia Bulletin*, *43*(6), 1251-1261.

Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of personality and social psychology*, *110*(5), 660.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: an ineradicable and egocentric bias in social perception. *Journal of personality and social psychology*, *67*(4), 596.

Lincoln, T. M., Peter, N., Schäfer, M., & Moritz, S. (2010). From stress to paranoia: an experimental investigation of the moderating and mediating role of reasoning biases. *Psychological Medicine*, *40*(1), 169-171.

Moutoussis, M., Dolan, R. J., & Dayan, P. (2016). How people use social information to find out what to want in the paradigmatic case of inter-temporal preferences. *PLoS computational biology*, *12*(7), e1004965.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision making*, *6*(8), 771-781.

Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, *18*(1), 13-41.

Ng, S. H., Han, S., Mao, L., & Lai, J. C. (2010). Dynamic bicultural brains: fMRI study of their flexible neural representation of self and significant others in response to culture primes. *Asian Journal of Social Psychology*, *13*(2), 83-91.

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. (2012). An agent independent axis for executed and modelled choice in medial prefrontal cortex. *Neuron*, *75*(6), 1114-1121.

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLoS computational biology*, *15*(6), e1007043.

Plitt, M. H., & Giocomo, L. M. (2021). Experience-dependent contextual codes in the hippocampus. *Nature Neuroscience*, *24*(5), 705-714.

Pronin, E. (2008). How we see ourselves and how we see others. *Science*, *320*(5880), 1177-1180.

Raihani, N. J., & Bell, V. (2017). Paranoia and the social representation of others: a large-scale game theory approach. *Scientific Reports*, *7*(1), 1-9.

Raihani, N. J., & Bell, V. (2018). Conflict and cooperation in paranoia: a large-scale behavioural experiment. *Psychological Medicine*, *48*(9), 1523-1531.

Raihani, N., Martinez-Gatell, D., Bell, V., & Foulkes, L. (2020). Social reward, punishment, and prosociality in paranoia. *Journal of abnormal psychology*.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495-505.

Rhoads, S. A., Cutler, J., & Marsh, A. A. (2021). A feature-based network analysis and fMRI meta-analysis reveal three distinct types of prosocial decisions. *bioRxiv*, 2020-12.

Riches, S., Bird, L., Chan, N., Garety, P., Rus-Calafell, M., & Valmaggia, L. (2020). Subjective experience of paranoid ideation in a virtual reality social environment: A mixed methods cross-sectional study. *Clinical psychology & psychotherapy*, *27*(3), 337-345.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and social psychology review*, *9*(1), 32-47.

Saalfeld, V., Ramadan, Z., Bell, V., & Raihani, N. J. (2018). Experimentally induced social threat increases paranoid thinking. *Royal Society open science*, *5*(8), 180569.

Schaerer, M., Foulk, T., du Plessis, C., Tu, M. H., & Krishnan, S. (2021). Just because you're powerless doesn't mean they aren't out to get you: Low power, paranoia, and aggression. *Organizational Behavior and Human Decision Processes*, *165*, 1-20.

Sevgi, M., Diaconescu, A. O., Henco, L., Tittgemeyer, M., & Schilbach, L. (2020). Social Bayes: using Bayesian modeling to study autistic trait–related differences in social cognition. *Biological Psychiatry*, *87*(2), 185-193.

Suzuki, S., Jensen, E. L., Bossaerts, P., & O'Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences*, *113*(14), 3755-3760.

Tarantola, T., Kumaran, D., Dayan, P., & De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature communications*, *8*(1), 1-14.

The International Cognitive Ability Resource Team. (2014). The International Cognitive Ability Resource. https://icar-project.com/.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. Current Opinion in Behavioral Sciences, 38, 110-115.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22.

Wellstein, K. V., Diaconescu, A. O., Bischof, M., Rüesch, A., Paolini, G., Aponte, E. A., ... & Stephan, K. E. (2020). Inflexible social inference in individuals with subclinical persecutory delusional tendencies. *Schizophrenia research*, *215*, 344-351.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, H., François, R., Henry, L., and Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS computational biology*, *8*(12), e1002841.

Zhang, L., & Gläscher, J. (2020). A brain network supporting social influences in human decision-making. Science advances, 6(34), eabb4159.

900 Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *Neuroimage*, *34*(3), 1310-1316.

Appendix

Table A.1: Model descriptions and free parameters. The winning model is highlighted in bold. NB: Our winning model estimated individual parameters simultaneously for both phases 1 and 2, since the participants' own preferences play a critical role in their learning about their partners. Since we directly perform analyses on the participants' own parameters, $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$, we sought to ensure that this simultaneous fit was not corrupting our estimates of these quantities. Therefore, we estimated each participant's $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$ exclusively from trials 1-18 (using the Phase 1 Only model) and assessed the correlations with the values estimated by the winning model based on all 18+36=54 trials. The correlations were indeed high ($\alpha_{\mathrm{ppt}}^{\mathrm{m}}$: r = 0.94, 95%CI: 0.93, 0.95; $\beta_{\mathrm{ppt}}^{\mathrm{m}}$: r = 0.98, 95%CI: 0.98, 0.99). We also assessed whether $\alpha_{\mathrm{ppt}}^{\mathrm{m}}$ and $\beta_{\mathrm{ppt}}^{\mathrm{m}}$ inferred only from Phase 2 data were correlated with parameters estimated from the winning model. Correlations were significant, but significantly lower ($P_{permuted} \sim 0$) than correlations with winning model parameters and Phase 1-only estimated parameters ($\alpha_{\mathrm{ppt}}^{\mathrm{m}}$: r = 0.60, 95%CI: 0.55, 0.65; $\beta_{\mathrm{ppt}}^{\mathrm{m}}$: r = 0.43, 95%CI: 0.37, 0.49).

Type	Model	Free parameters	Description
Phase 1 Only		$\alpha_{ppt}^{m}\beta_{ppt}^{m}$	This model quantifies the preference of a participant for the relative discrepancy between their earnings and their partner's earnings β_{ppt}^{m} and the preference a participant holds over their own absolute earnings, α_{ppt}^{m} using phase 1 data only.
Phase 2 Only		$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\ \beta^{\sigma}$	This model quantifies the preference of a participant for the relative discrepancy between their earnings and their partners earnings β_{ppt}^{m} and the preference a participant holds over their own absolute earnings, α_{ppt}^{m}, in addition to the flexibility with which they hold these beliefs $[\alpha^{\sigma}, \beta^{\sigma}]$ using phase 2 data only.
Bayesian Updating	1	$\beta_m\beta_v$	This model only quantifies the preference of a participant for the relative discrepancy between their earnings and their partners earnings β_{ppt}^{m}. In phase 2, this preference forms the central tendency for their beliefs about a partner's preference of relative payoff discrepancy which is held with flexibility β^{σ}.
	2	$\boldsymbol{\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\ \beta^{\sigma}}$	This model quantifies the preference of a participant for the relative discrepancy between their earnings and their partner's earnings β_{ppt}^{m} and the preference a participant holds over their own absolute earnings, α_{ppt}^{m} . In phase 2, the participant's relative payoff preference and preference for absolute payoffs form the central tendency for a participant's prior beliefs about their partner, which are held with standard deviation $\beta^{\sigma}\ \alpha^{\sigma}$. This model uses simultaneous estimation of participant preferences from their decisions and predictions across both phase 1 and phase 2.

	3	$\alpha_{ppt}^{m}\beta_{ppt}^{m}$	In both phases 1 and 2, this model assumes that the participant ignores their partner's preferences and only predicts what option their partner may choose in line with their own preference for relative payoffs β_{ppt}^{m} and absolute payoffs α_{ppt}^{m}.
	4	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ ω	Identical to model 2, with the addition of a parameter ω in phase 2 that shrinks the value of a participant's preferences toward 0 when integrating it into their prior beliefs about a partner.
	5	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ ζ	Identical to model 2, with the addition of a parameter ζ in phase 2 that quantifies how much a participant may over or under match predictions about their partner.
	6	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ κ	Identical to model 2, with the addition of a parameter κ in phase 2 that quantifies whether a participant is biased to learn about options a partner decided upon that are more favourable for the participant.
	7	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ ε	Identical to model 2, with the addition of a single lapse parameter in phase 2.
	8	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ $\varepsilon_{con}\varepsilon_{incon}$	Identical to model 2, with the addition of two lapse parameters in phase 2 that quantifies whether a participant is biased to persistently predict options their partner might choose that are congruent ε_{con} or incongruent ε_{incon} with their own SVO.
	9	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ $\rho_{con}\,\rho_{incon}$	Identical to model 2, with the addition of two lapse parameters in phase 2 that quantifies whether a participant is persistently biased to learn about *and* predict a partner's choices that are congruent ρ_{con} or incongruent ρ_{incon} with their own SVO.
	10	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\alpha^{\sigma}\beta^{\sigma}$ $\alpha_{ppt}^{2}\beta_{ppt}^{2}$	Identical to model 2, except that the model assumes participants do not use their prior preferences to form beliefs about their partner in phase 2 and instead use new central tendencies for their priors over a partner's preference for absolute payoffs α_{ppt}^{2} and relative payoffs β_{ppt}^{2} .
Heuristic/ Q - learning	11	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\,\tau$ λ	Like all Bayesian models, this model quantifies the participant's preferences for relative payoffs β_{ppt}^{m} and absolute payoffs α_{ppt}^{m}, although phase 2 uses an associative, model-free Q-learning framework based around categorical SVO predictions to quantify the learning rate λ and decision temperature τ of each participant.

	12	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\ \tau$ $\lambda_{pos}\lambda_{neg}$	Identical to model 11, except that separate learning rates for positive λ_{pos} and negative λ_{neg} prediction errors are quantified.
	13	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\ \tau$ $\lambda_{P}\lambda_{I}\lambda_{C}$	Identical to model 11, except that separate learning rates for prosocial λ_{P}, individualistic λ_{I}, and competitive λ_{C} categorical predictions about a partner are quantified.
	14	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\ \tau$ $\lambda_{con}\lambda_{icon}$	Identical to model 11, except that separate learning rates for congruent λ_{con} and incongruent λ_{incon} actions by a partner relative to what a participant would have chosen are quantified (calculated using a participant's $\alpha_{ppt}^{m}\beta_{ppt}^{m}$).
	15	$\alpha_{ppt}^{m}\beta_{ppt}^{m}\ \tau$ $\lambda\ \omega$	Identical to model 11, except that a consonance parameter ω is added that initiates Q_a^0 values according to the closest paradigmatic partner type to the participant's own preferences, and $(1-\omega)/2$ to the Q_a^0 of the two other partners.

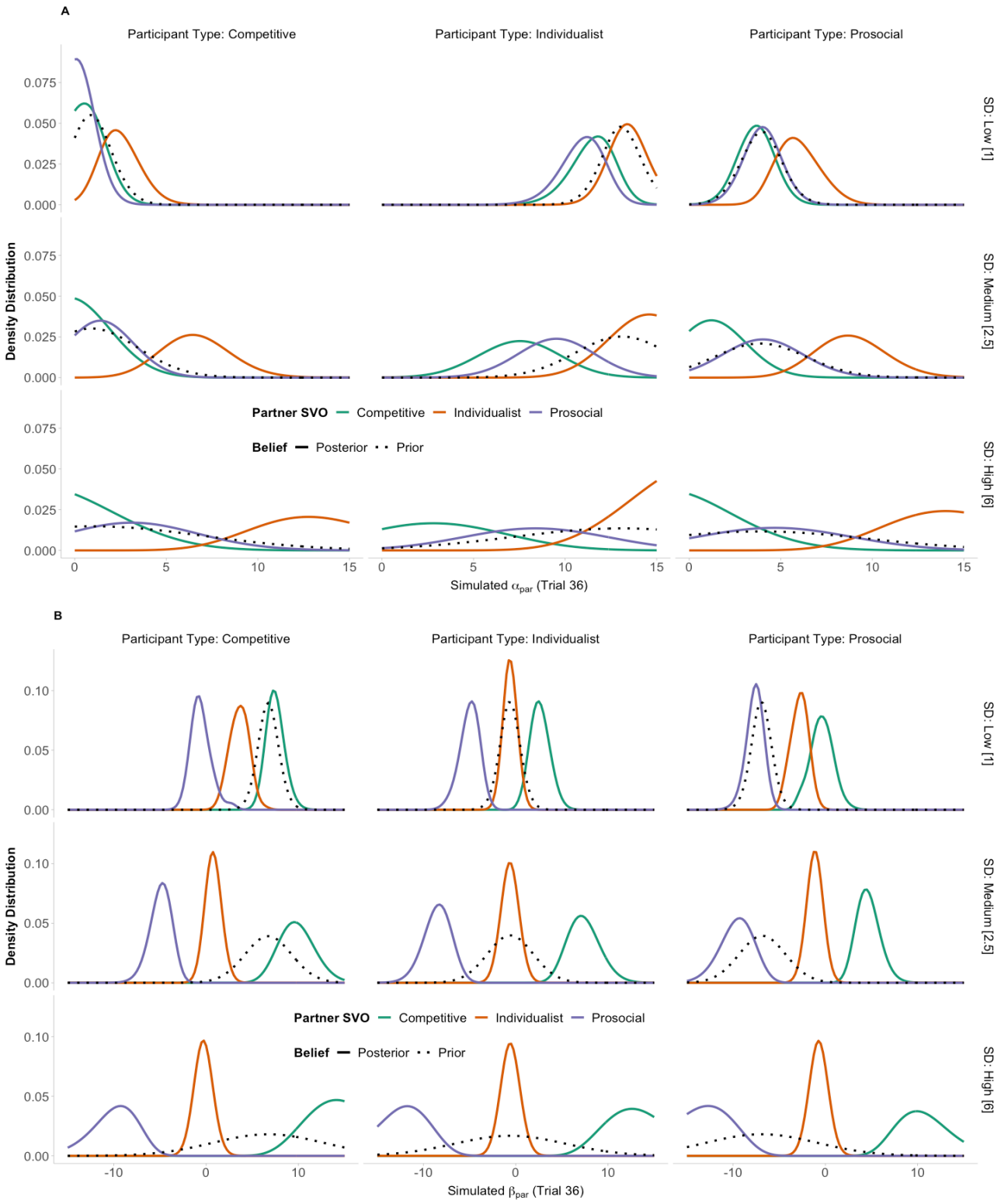


Figure A.1: Generative model simulations.

Simulations of prior (t = 0) and posterior (t = 36) beliefs over α_{par} (A) and β_{par} (B) of synthetic prosocial, competitive, and individualist participants (x-axis facets) that have either low, medium, or high σ over their prior beliefs (y-axis facets) about partners of different SVO types.

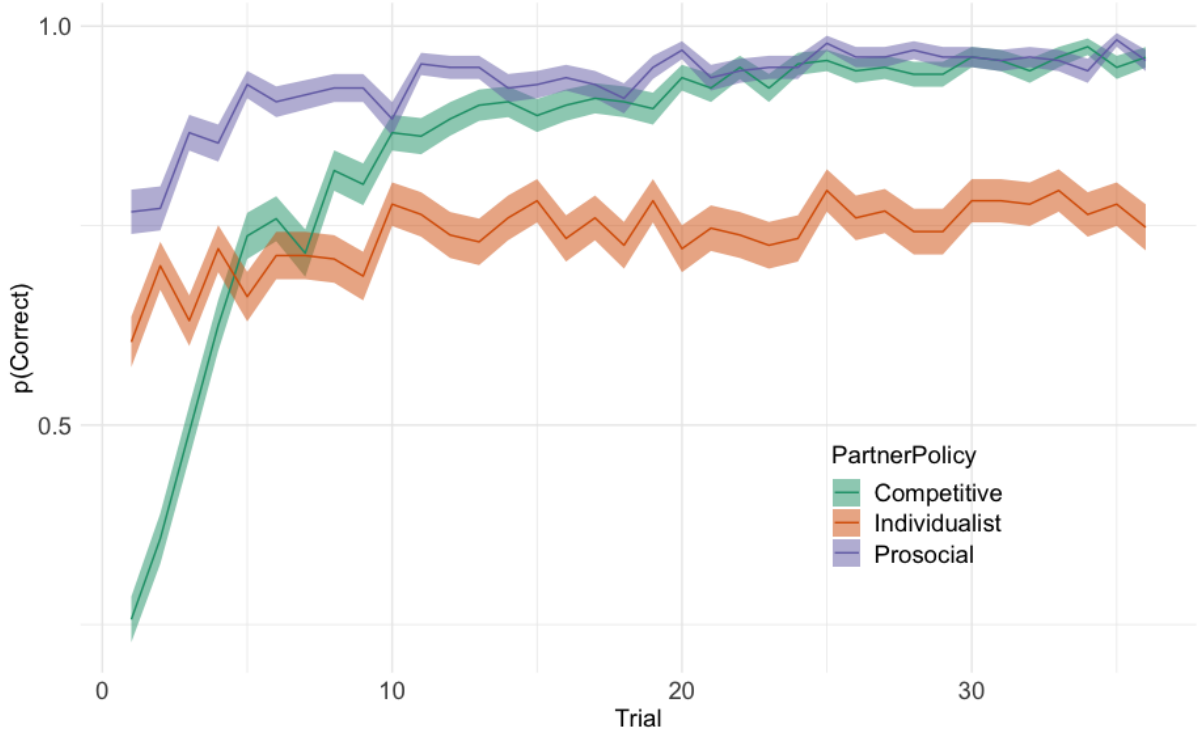


Figure B.1: Proportion of correct predictions made across the population for each partner type and trial in phase 2.

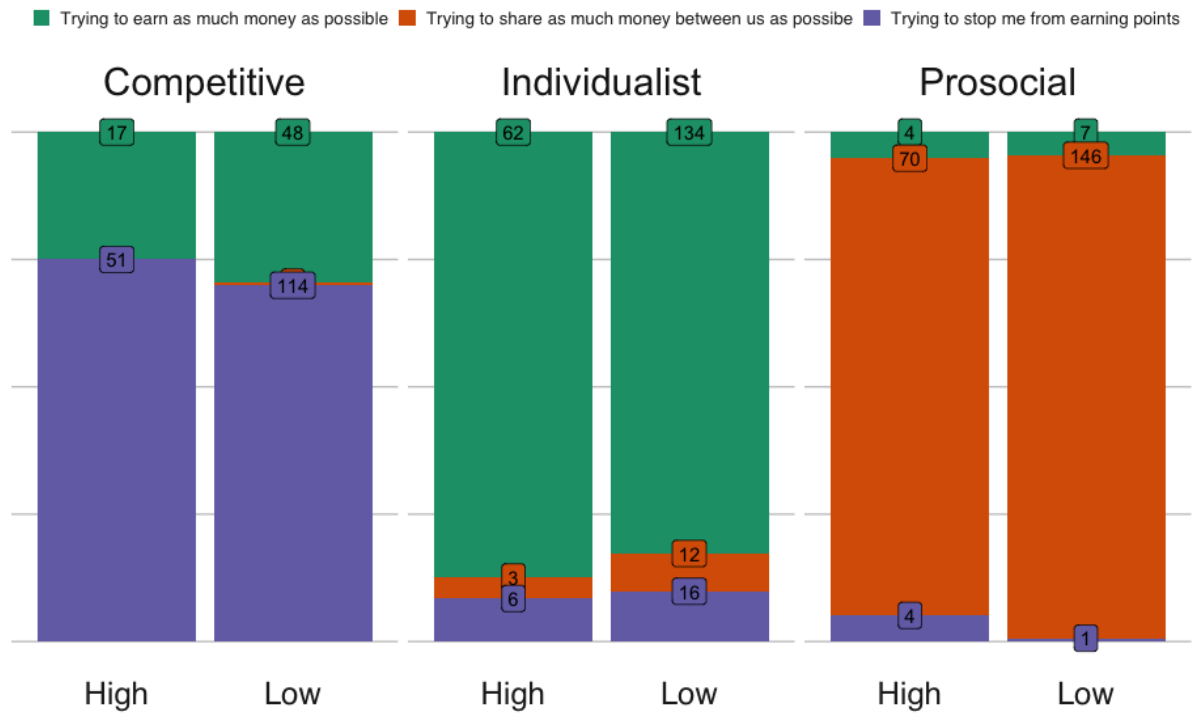


Figure C.1: Participants' classification of their partners' social preferences after Phase 2. High = high paranoia, Low = low paranoia. Numbers = count per group.

Table B.1 : Option pairs (participant, partner) presented to participants in Phase 1.

Participants made 18 choices in three categories. All participants saw all option pairs, but these were presented in a random order and their position on screen was counter-balanced, e.g., within prosocial-competitive option dyads, prosocial options appeared on the left three times and on the right three times. In Phase 1, participants acted in the role of decider in a modified SVO task. Participants chose between two options that determined the allocation of points between themselves and their partner (the receiver). Specifically, participants made six choices between prosocial and individualistic / competitive options, six choices between prosocial and competitive options, and six choices between individualistic and competitive options.

Prosocial-Individualistic/Competitive	**Prosocial**	**Individualist/Competitive**
1	6,6	10,5
2	7,7	10,5
3	8,8	10,5
4	8,8	12,5
5	9,9	12,5
6	10,10	12,5
Prosocial-Competitive	**Prosocial**	**Competitive**
1	6,6	6,2
2	7,7	7,2
3	8,8	8,5
4	8,8	8,2
5	9,9	9,5
6	10,10	10,5
Individualistic-Competitive	**Individualist**	**Competitive**
1	8,5	6,2
2	8,5	7,2
3	9,5	8,2
4	10,6	8.2
5	11,6	9,2
6	12,6	10,2

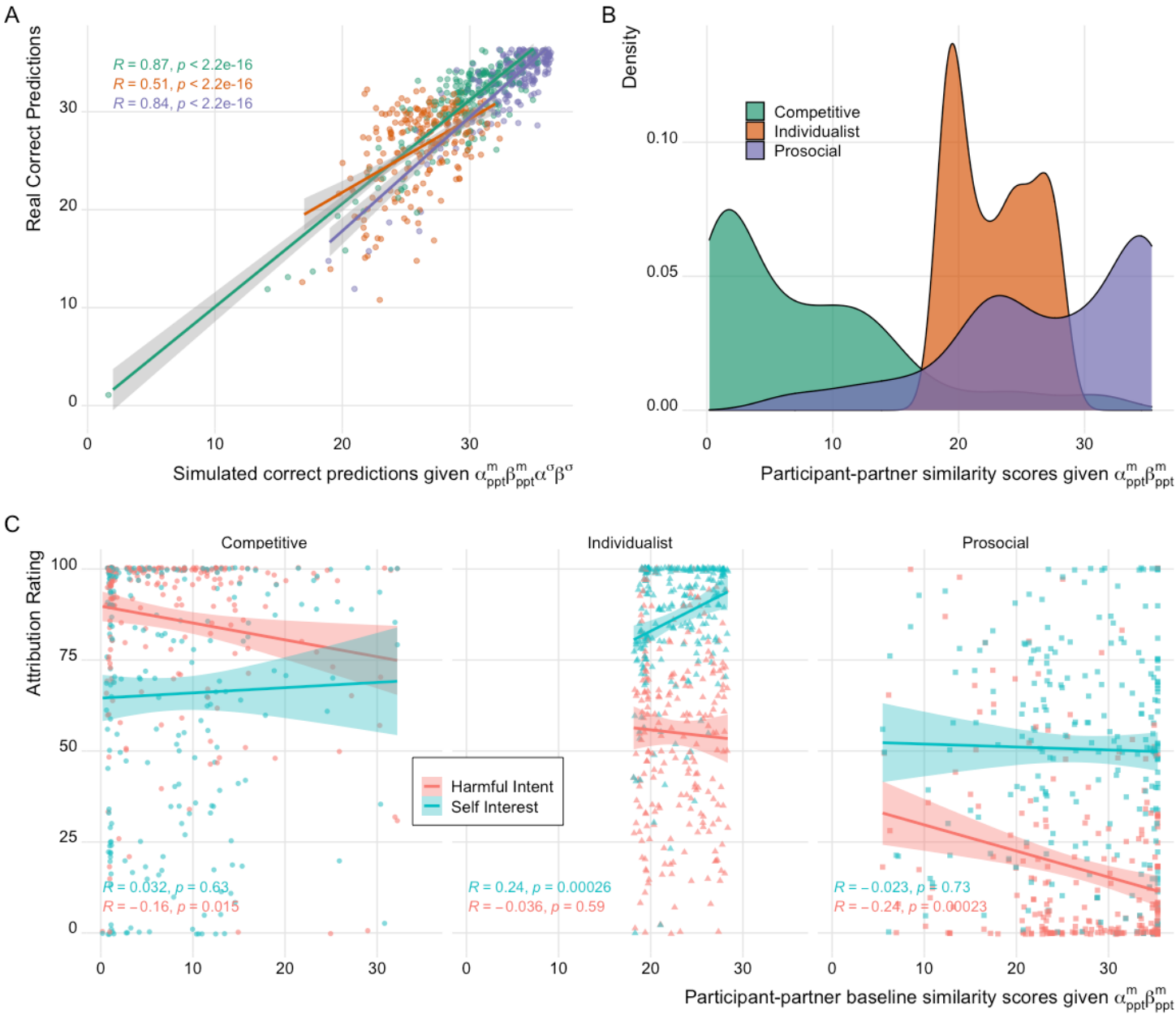
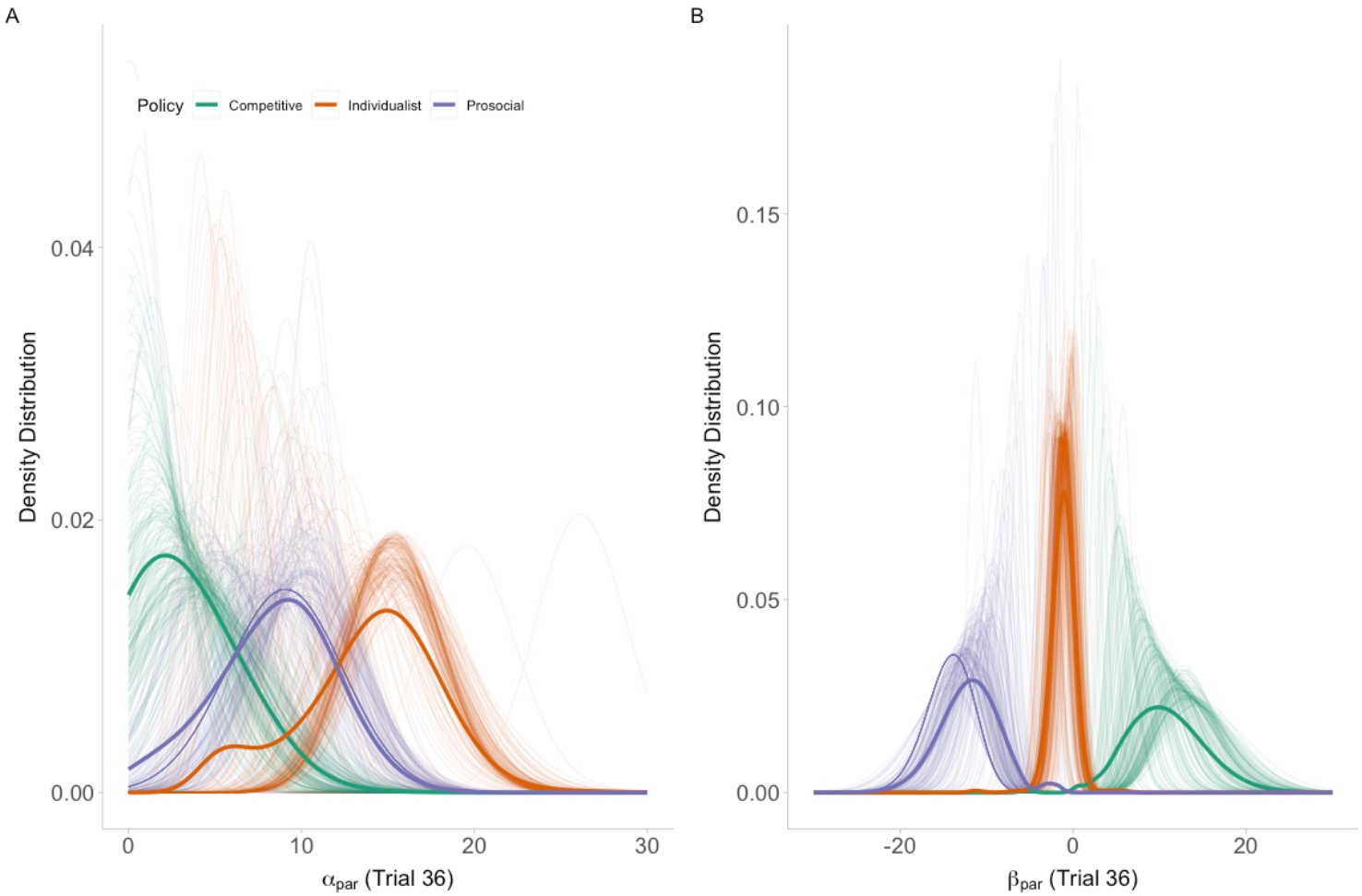


Figure D.1: Simulated and real correct predictions

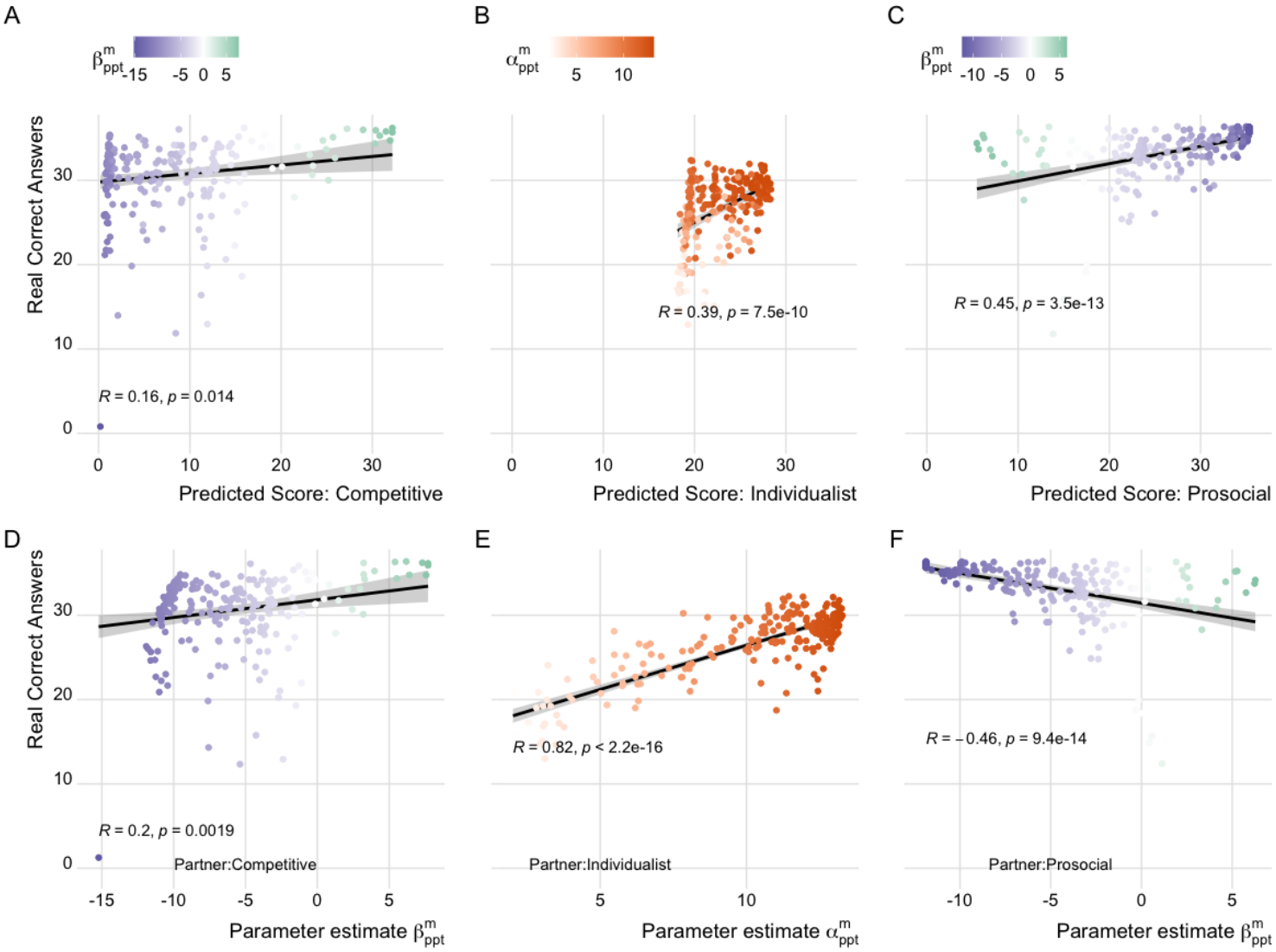
(A) Simulated sum of correct predictions generated for each participant's parameters ($\theta_{\alpha,\beta}$) by the real sum of correct predictions observed in the data for each partner. (B) Density distribution over participant-partner similarity scores in phase 2. This is calculated by estimating from the model the probability of each prediction a notional participant would have chosen in phase 2 given their $\alpha_{ppt}\beta_{ppt}$ (estimated from phase 1 alone without any learning). (C) Raw correlation between attributions and baseline similarity scores for each participant and faceted within each partner.



960

Figure E.1: Simulated posterior marginal distributions of the belief participants held about their partner's competitive and individualistic preferences at the end of phase 2 ($\theta_{\alpha,\beta}^{t=36}$).

965 Simulated marginal distributions of p(α_{par}) [A] and p(β_{par}) [B] for each partner policy. Thick lines denote the group average distribution. Thin lines denote individual participant distributions.



970

Figure F.1: Interaction between participant-partner similarity scores, parameters, and real correct answers.

975 (A-C) Relationship between real correct answers made by participants against their participant-partner baseline similarity given $\alpha_{ppt}^{m}\beta_{ppt}^{m}$ (predicted score), colour graded by parameter values. (D-F) Relationship between real correct answers made by participants against predicted score and parameter values, colour graded by parameter values.

Table C.1: Variables associated with participants' utilities concerning absolute (α^{m}_{ppt}) and relative (β^{m}_{ppt}) payoffs.

All estimates are following model averaging. Higher β^{m}_{ppt} values imply a stronger preference for earning more than the partner. Bold highlight signifies that the 95%CI does not cross 0. Model number refers to the model signifier in text and on Github.

Model 5a: α^{m}_{ppt}					
			CI (95%)		
Parameter	**Estimate**	**Std. Error**	**Lower**	**Upper**	**Importance**
(Intercept)	-0.44	0.12	-0.67	-0.21	
Task Comp.	**0.27**	**0.06**	**0.14**	**0.39**	**1.00**
ICAR	**0.22**	**0.04**	**0.15**	**0.29**	**1.00**
Paranoia	0.00	0.02	-0.08	0.06	0.28
Sex (M|F)	-0.02	0.05	-0.21	0.08	0.20
Age	**-0.09**	**0.04**	**-0.16**	**-0.02**	**1.00**
β^{m}_{ppt}	**-0.13**	**0.04**	**-0.20**	**-0.06**	**1.00**
Model 5b: β^{m}_{ppt}					
			CI (95%)		
Parameter	**Estimate**	**Std. Error**	**Lower**	**Upper**	**Importance**
(Intercept)	0.09	0.08	-0.07	0.26	
Paranoia	**0.09**	**0.04**	**0.01**	**0.16**	**0.12**
Task Comp.	0.01	0.03	-0.09	0.17	0.19
ICAR	0.00	0.02	-0.06	0.09	0.17
Age	-0.01	0.02	-0.10	0.05	0.21
α^{m}_{ppt}	**-0.13**	**0.04**	**-0.21**	**-0.06**	**1.00**
Sex (M|F)	**-0.16**	**0.08**	**-0.32**	**-0.01**	**1.00**

Table D.1: Factors associated with predictive accuracy

All estimates are following model averaging. We analysed the data once including all partner types, and then as three separate models, where we segregated data according to partner type. Estimates are averaged across the top model set. NA = parameter not included in the final top model. Task Comp = Task Comprehension. ICAR = International Cognitive Ability Resource (Matrix Reasoning). CI = Confidence Interval. Bold highlight signifies that the 95%CI does not cross 0. Model number refers to the model signifier in text and on Github.

$$
\begin{aligned}
&Predictive\ Accuracy \\
&= \beta 0 + (\beta 1 * \alpha^{\sigma}) + (\beta 2 * \beta^{\sigma}) + (\beta 3 * Baseline\ Similarity) \\
&+ (\beta 4 * Paranoia) + (\beta 5 * General\ Cognition) + (\beta 6 * Age) + (\beta 7 * Sex) \\
&+ (\beta 8 * Task\ Comprehension) + (\beta 9 * [\alpha^{\sigma} * \beta^{\sigma}]) \\
&+ (\beta 10 * [\alpha^{\sigma} * Baseline\ Similarity]) + (\beta 11 * [\beta^{\sigma} * Baseline\ Similarity]) \\
&+ (\beta 12 * [\alpha^{\sigma} * Baseline\ Similarity * \beta^{\sigma}])
\end{aligned}
$$

Model 7: Predictive Accuracy					
			CI (95%)		
Parameter	**Estimate**	**Std. Error**	**Lower**	**Upper**	**Importance**
(Intercept)	-0.28	0.09	-0.47	-0.10	
Baseline Similarity	**0.62**	**0.03**	**0.56**	**0.68**	**1.00**
α^{σ}	**0.49**	**0.03**	**0.43**	**0.55**	**1.00**
β^{σ}	**0.47**	**0.04**	**0.40**	**0.54**	**1.00**
α^{σ}: **Baseline Similarity:** β^{σ}	**0.23**	**0.04**	**0.16**	**0.31**	**1.00**
α^{σ}:β^{σ}	**0.16**	**0.03**	**0.09**	**0.22**	**1.00**
α^{σ}: **Baseline Similarity**	**0.12**	**0.04**	**0.04**	**0.19**	**1.00**
Sex (M|F)	**0.11**	**0.05**	**0.01**	**0.21**	**1.00**
Task Comp.	**0.06**	**0.05**	**0.00**	**0.17**	**0.78**
ICAR	**0.04**	**0.03**	**0.00**	**0.10**	**0.77**
Age	0.01	0.02	-0.02	0.08	0.38
Paranoia	0.00	0.01	-0.07	0.03	0.12
β^{σ}: **Baseline Similarity**	**-0.31**	**0.03**	**-0.38**	**-0.25**	**1.00**

Table E.1: Factors associated with harmful intent and self-interest attributions

All estimates are following model averaging. NA = parameter not included in the final top model. Task Comp = Task Comprehension. ICAR = International Cognitive Ability Resource (Matrix Reasoning). P = prosocial partner. C = competitive partner. I = individualistic partner. CI = Confidence Interval. Bold highlight signifies that the 95%CI does not cross 0. Model number refers to the model signifier in text and on Github.

$$Attribution = \beta 0 + (\beta 1 * \alpha^{\sigma}) + (\beta 2 * \beta^{\sigma}) + (\beta 3 * Baseline\ Similarity) + (\beta 4 * Paranoia) + (\beta 5 * Predictive\ Accuracy) + (\beta 6 * Partner\ Policy) + (\beta 7 * General\ Cognition) + (\beta 8 * Age) + (\beta 9 * Sex) + (\beta 10 * Task\ Comprehension)$$

Model 9a: Harmful Intent Attributions					
			CI (95%)		
Parameter	**Estimate**	**Std. Error**	**Lower**	**Upper**	**Importance**
(Intercept)	0.66	0.07	0.51	0.81	
α^{σ}	**0.08**	**0.03**	**0.03**	**0.14**	**1.00**
Paranoia	**0.06**	**0.02**	**0.01**	**0.11**	**1.00**
β^{σ}	0.00	0.02	-0.08	0.03	0.23
Age	0.00	0.00	0.00	0.00	0.10
Sex (M|F)	0.00	0.02	-0.12	0.08	0.10
Task Comp.	0.00	0.02	-0.06	0.11	0.11
ICAR	-0.02	0.03	-0.08	0.01	0.66
Baseline Similarity	**-0.17**	**0.04**	**-0.24**	**-0.09**	**1.00**
Partner (I|C)	**-0.50**	**0.09**	**-0.68**	**-0.33**	**1.00**
Partner (P|C)	**-1.48**	**0.10**	**-1.67**	**-1.29**	**1.00**
Predictive Accuracy	NA	NA	NA	NA	NA
Model 9b: Self-Interest Attributions					
			CI (95%)		
Parameter	**Estimate**	**Std. Error**	**Lower**	**Upper**	**Importance**
(Intercept)	0.12	0.14	-0.15	0.39	
Partner (I|C)	**0.84**	**0.13**	**0.58**	**1.10**	**1.00**

Predictive Accuracy	**0.29**	**0.06**	**0.18**	**0.40**	**1.00**
Sex (M|F)	0.07	0.08	-0.03	0.25	0.62
Paranoia	0.00	0.01	-0.05	0.08	0.07
Age	NA	NA	NA	NA	1.00
α^{σ}	-0.02	0.04	-0.14	0.03	0.39
ICAR	-0.04	0.04	-0.12	0.01	0.65
Baseline Similarity	-0.11	0.07	-0.24	0.00	0.93
Task Comp.	**-0.14**	**0.06**	**-0.25**	**-0.02**	**1.00**
$\boldsymbol{\beta^{\sigma}}$	**-0.19**	**0.05**	**-0.28**	**-0.09**	**1.00**
Partner (P|C)	**-0.62**	**0.14**	**-0.89**	**-0.36**	**1.00**

1015

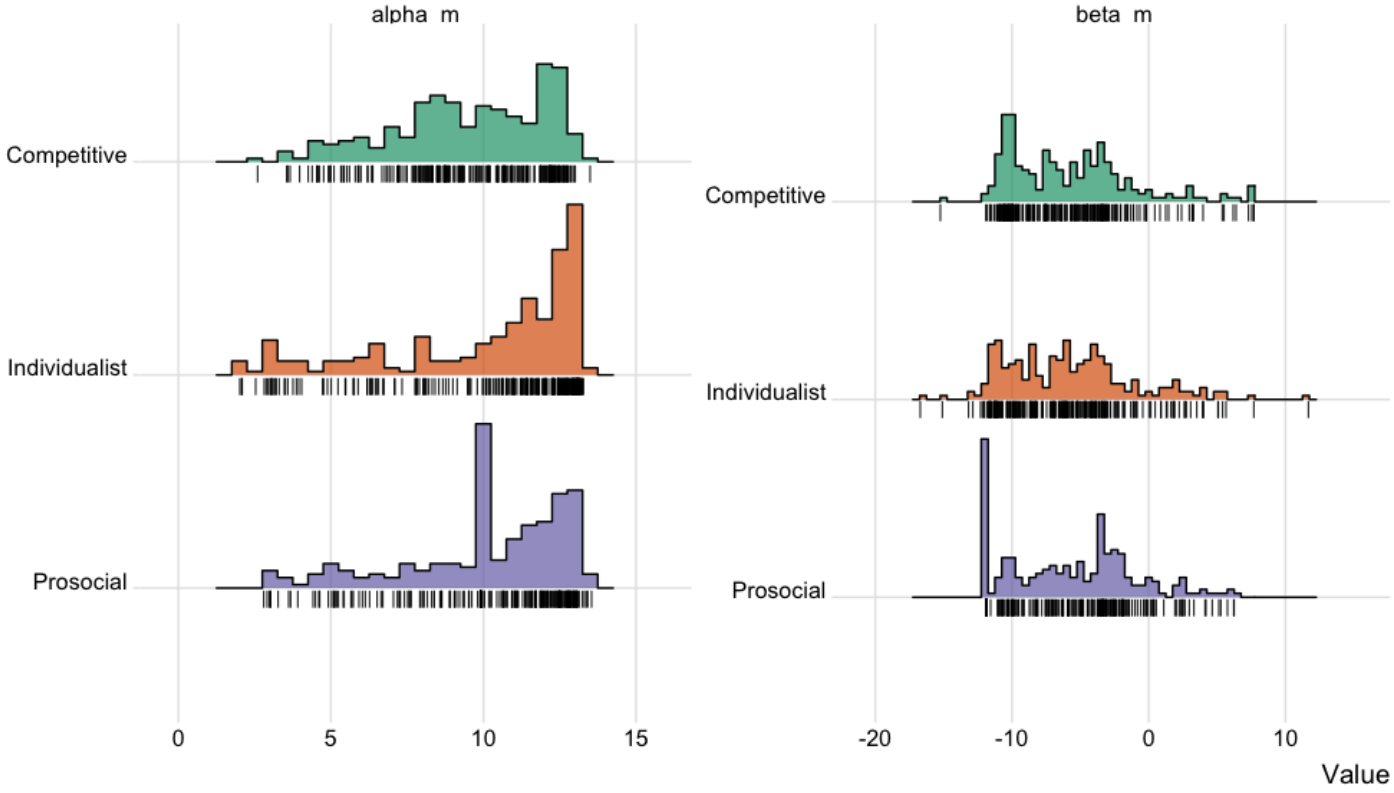
Table F.1: Summary of the inferred preferences and beliefs of participants

Participant preferences are those estimated by the winning model. All Phase 2 parameters for each participant are mean approximations of the belief distribution following trial 36 of Phase 2.

Parameter	**Partner**				
	Prosocial	**Individualist**	**Competitive**	Test Statistic $F_{(694)}$ =	P-value
n	232	233	232		
Inferences about the participant's preferences					
$\boldsymbol{\alpha}^{\mathbf{m}}_{\mathbf{ppt}}$					
mean [sd]	10.10 [2.63]]	10.10 [3.23]	9.56 [2.49]	3.13	0.044
min	2.80	2.02	2.61		
max	13.5	13.3	13.5		
$\boldsymbol{\beta}^{\mathbf{m}}_{\mathbf{ppt}}$					
mean [sd]	-5.69 [4.67]	-5.89 [4.59]	-5.90 [4.44]	0.16	0.853
min	-11.9	-16.7	-15.2		
max	6.27	11.70	7.70		
Inferences about the participant's belief about the partner					
				$F_{(2)}$ =	
$\widehat{\boldsymbol{\alpha}}^{\mathbf{m}}_{\mathbf{par}}$					
mean[sd]	16.5 [4.20]	26.7 [5.89]	5.10 [3.25]	1290	~0
$\widehat{\boldsymbol{\beta}}^{\mathbf{m}}_{\mathbf{par}}$					
mean[sd]	-22.8 [4.65]	-2.15 [2.07]	20.0 [6.43]	4570	~0
$\Delta(\alpha^{m}_{par})$					
mean[sd]	0.26 [0.14]	0.70 [0.16]	1.21 [1.40]	77.9	~0
$\Delta(\beta^{m}_{par})$					
mean[sd]	0.76 [0.74]	0.20 [0.14]	8.03 [9.00]	161.3	~0

Figure G.1: Distribution of participant preferences (from Phase 1) within each partner type

1025



Text A.1: Competing Bayesian model formalisms for phase 2.

Beta-only model (Model 1)

Phase 1 – **estimating participants' social preferences**

These equations were applied to all models for phase 1, aside from the 'beta-only' model. We modelled participant SVO as ranging along one dimension: relative payoffs (β_{ppt}) – how much participants preferred creating discrepancy between themselves and their partner.

In Phase 1, participants made 18 choices c^t, t = {1…T} about whether option 1 or option 2 should be chosen given the utility (U^t) of each, such that the likelihood of choosing option one was:

(A.1)

$$U_{\beta_{ppt}}(\mathbf{R}^{t;\mathbf{1}}) = \beta_{ppt} * \max(R^{t;1}_{\text{self}} - R^{t;1}_{\text{other}}, 0)$$

$$\Delta U_{\beta_{ppt}}(\mathbf{R}^{t}) = U_{\beta_{ppt}}(\mathbf{R}^{t;\mathbf{1}}) - U_{\beta_{ppt}}(\mathbf{R}^{t;\mathbf{2}})$$

$$p(c^t = 1|\beta_{ppt};\ \mathbf{R}^{t}) = \sigma(\Delta U_{\beta_{ppt}}(\mathbf{R}^{t})) \text{ or}$$

$$p(c^t|\beta_{ppt};\ \mathbf{R}^{t}) = \sigma((2c^t - 1)\,\Delta U_{\beta_{ppt}}(\mathbf{R}^{t}))$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic sigmoid.

Phase 2 – how participants learned about their partners' social preferences

Over 36 trials, participants made binary predictions $\hat{d}^t$, t = {1…T} about whether option 1 or option 2 would be chosen by their partner given the returns $\mathbf{R}^{t}$=$\{\mathbf{R}^{t;\mathbf{1}}; \mathbf{R}^{t;\mathbf{2}}\}$ = $\{R^{t;1}_{\text{self}}, R^{t;1}_{\text{other}};\ R^{t;2}_{\text{self}}, R^{t;2}_{\text{other}}\}$ of each pair of offers. They then discovered what the partner actually chose, which we write as d^t.

The participant assumed that the partner chose the same way that they did themselves, but with SVO parameters β_{par}, which they needed to infer from observation. That is, the log likelihood that the partner chose d^t is $LL = \log(\ p(d^t|\beta_{par};\ \mathbf{R}^{t})\ )$ using the same formula as equation 8.

The partner's decisions $D^t = \{d^1, d^2, \ldots, d^t\}$ were used to update a participant's beliefs about a partner's β_{par} , written as $p(\beta_{par}|D^t)$. The starting point for these beliefs (written as $p(\beta_{par}|D^0)$ was the participant's prior. For model 1, we assumed that this was a factorised distribution with each parameter centred on the participant's own preference β^m_{ppt} but with a standard deviation parameter β^σ that characterised the extent to which the participant thought their partner might differ from themselves. Therefore, we had

(A.2)

$$p(\beta_{par}|D^0) \sim N\left(\beta_{par}; \beta^m_{ppt}, \beta^\sigma\right)$$

We then assumed that a participant's posterior beliefs about their partner from trials t = 1…36 given a partner's decisions followed Bayes rule:

(A.3)

$$p\left(\beta_{par} \,\middle|\, D^t\right) = \frac{p(d^t|\beta_{par};\ \mathbf{R}^{t})\, p\left(\beta_{par}|D^{t-1}\right)}{p(d^t|D^{t-1})}$$

For efficiency, we could conveniently represent $p\left(\beta_{par} \,\middle|\, D^t\right)$ by a vector over the fixed grid of β values, $\theta^t_{\beta_{par}}$. We then calculated the participant's beliefs about their partner's SVO preferences for each trial:

(A.4)

$$\theta^t_{\beta_{par}} = \frac{p(d^t|\beta_{par};\, \mathbf{R^t})\theta^{t-1}_{\beta_{par}}}{\sum_{\beta'_{par}} p(d^t|\beta'_{par};\, \mathbf{R^t})\theta^{t-1}_{\beta'_{par}}}$$

The model then stated that the participant predicted their partner's choice by calculating the probability determined by the utility differences $\Delta U_{\beta_{par}}(\mathbf{R^{t+1}})$ as in equation (1), summed over the distribution $\theta^t_{\beta_{par}}$ over the partner parameters

(A.5)

$$p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right) = \sum_{\beta_{par}} \theta^t_{\beta_{par}} \cdot \sigma(\Delta U_{\beta_{par}}(\mathbf{R^{t+1}}))$$

$$p\left(d^{t+1} = 2 \middle| D^t; \mathbf{R^{t+1}}\right) = 1 - p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right)$$

and then performed probability matching, so that

$$p\left(\hat{d}^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right) = p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right).$$

Participant ignores the partner (Model 3)

We fitted the data based on the rules that the participant predicts their partner's decision based on their own α_{ppt} & β_{ppt} values rather than considering their partner's preferences. In this model, participants did not update their inferences based on feedback about whether they were correct or incorrect.

(A.6)

$$\alpha_{par} = \alpha_{ppt}\,;\ \beta_{par} = \beta_{ppt}$$

$$\Delta U_{\alpha_{par},\beta_{par}} = U_{\alpha_{par},\beta_{par}}\left(\mathbf{R^{t+1;1}}\right) - U_{\alpha_{par},\beta_{par}}(\mathbf{R^{t+1;2}})$$

$$p(d^{t+1} = 1|\alpha_{par}, \beta_{par};\, \mathbf{R^{t+1}}) = \sigma(\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^{t+1}}))$$

$$p\left(\hat{d}^{t+1} = 1 \middle| D^t, \mathbf{R^{t+1}}\right) = p\left(d^{t+1} = 1 \middle| \alpha_{par}, \beta_{par}; \mathbf{R^{t+1}}\right)$$

Shrinkage Model (Model 4)

We introduced a version of the winning model with a parameter that pulled the central tendency of a participant's prior beliefs about a partner's α_{par} & β_{par} toward zero using a shrinkage parameter (ω) between phase 1 and phase 2. This replaces equation 3.

(A.7)

$$\alpha_{ppt}^{m} = \alpha_{ppt} \cdot \omega$$

$$\beta_{ppt}^{m} = \beta_{ppt} \cdot \omega$$

$$p(\alpha_{par} \mid D^0) \sim N(\alpha_{par}; \alpha_{ppt}^{m}, \alpha^{\sigma})$$

$$p(\beta_{par} \mid D^0) \sim N\left(\beta_{par}; \beta_{ppt}^{m}, \beta^{\sigma}\right)$$

$$p\left(\alpha_{par}, \beta_{par} \mid D^0\right) = p\left(\alpha_{par} \mid D^0\right) p\left(\beta_{par} \mid D^0\right)$$

Probability over/under matching model (Model 5)

In this model we allowed for a parameter ζ to account for probability under or over-matching by a participant about a predicted decision, although inference occurred as per equation 5. This equation replaces equation 7.

(A.8)

$$p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right) = \sum_{\alpha_{par}, \beta_{par}} \theta^{t}_{\alpha_{par}, \beta_{par}} \cdot \sigma\left(\Delta U_{\alpha_{par}, \beta_{par}}(\mathbf{R^{t+1}})\right)$$

$$p\left(d^{t+1} = 2 \middle| D^t; \mathbf{R^{t+1}}\right) = 1 - p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right)$$

$$p\left(\hat{d}^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right) = \frac{\left[p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right)\right]^{\zeta}}{\left[p\left(d^{t+1} = 1 \middle| D^t; \mathbf{R^{t+1}}\right)\right]^{\zeta} + \left[p\left(d^{t+1} = 2 \middle| D^t; \mathbf{R^{t+1}}\right)\right]^{\zeta}}$$

Favourable bias (Model 6)

This model assumed that participants may have selectively learned about a partner's actions that were more favourable to the participant, where $\mathbf{R^t_{self}}$ are the returns to the participant. This equation replaces equation 5.

(A.9)

$$\pi_1^t = \sigma\left(\Delta U_{\alpha_{par}, \beta_{par}}(\mathbf{R^t})\right) \exp\left(\kappa \cdot R^{t;1}_{\text{self}}\right)$$

$$\pi_2^t = \sigma\left(-\Delta U_{\alpha_{par}, \beta_{par}}(\mathbf{R^t})\right) \exp\left(\kappa \cdot R^{t;2}_{\text{self}}\right)$$

$$p\left(d^t = 1 \middle| \alpha_{par}, \beta_{par}; \mathbf{R^t}\right) = \frac{\pi_1^t}{\pi_1^t + \pi_2^t}$$

$$\theta^{t}_{\alpha_{par}, \beta_{par}} = \frac{p(d^t \mid \alpha_{par}, \beta_{par}; \mathbf{R^t}) \theta^{t-1}_{\alpha_{par}, \beta_{par}}}{\sum_{\alpha'_{par}, \beta'_{par}} p(d^t \mid \alpha'_{par}, \beta'_{par}; \mathbf{R^t}) \theta^{t-1}_{\alpha'_{par}, \beta'_{par}}}$$

Lapse rate allowance (Model 7)

Model 7 allowed for the participant to make lapse errors during their predictions of a partner using a single parameter ε and therefore replaces equation 7:

(A.10)

$$p\left(d^{t+1}=1\middle|D^t;\mathbf{R^{t+1}}\right)=\sum_{\alpha_{par},\beta_{par}}\theta^t_{\alpha_{par},\beta_{par}}\cdot\left[(1-\varepsilon)\cdot\sigma\left(\Delta U_{\alpha_{par},\beta_{par}}\left(\mathbf{R^{t+1}}\right)\right)+\frac{\varepsilon}{2}\right]$$

$$p\left(\hat{d}^{t+1}=1\middle|D^t;\mathbf{R^{t+1}}\right)=p\left(d^{t+1}=1\middle|D^t;\mathbf{R^{t+1}}\right)$$

Participant choice congruency bias (Model 8)

In model 6, the participant's own return influenced the way that they learned about their partner. In model 8, we considered the case that learning proceeds normally, but the participant evaluated the probability that the partner chose a particular option in a way that is biased by the participant's preferences. To this, we considered whether a potential partner prediction is congruent with the participant's preferences (having a greater utility for the participant based on the participant's SVO parameters) or incongruent. We implemented this using two lapse-like parameters, each bounded between 0 and 2. We therefore replaced equation 7 as:

(A.11)

$$p\left(d^{t+1}=1\middle|D^t;\mathbf{R^{t+1}}\right)=\sum_{\alpha_{par},\beta_{par}}\theta^t_{\alpha_{par},\beta_{par}}\cdot\begin{cases}(1-\varepsilon_c)\cdot\sigma\left(\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^t})\right)+\dfrac{\varepsilon_c}{2} & \text{if } U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;1}}\right)>U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;2}}\right)\\ (1-\varepsilon_i)\cdot\sigma\left(\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^t})\right)+\dfrac{\varepsilon_i}{2} & \text{if } U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;1}}\right)\le U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;2}}\right)\end{cases}$$

$$p\left(\hat{d}^{t+1}=1\middle|D^t;\mathbf{R^{t+1}}\right)=p\left(d^{t+1}=1\middle|D^t;\mathbf{R^{t+1}}\right)$$

Participant learning congruency bias (Model 9)

In model 9, we allowed congruency to affect learning as well as predictions. Therefore, we replaced equation 5 as:

(A.12)

$$p(d^t=1\mid\alpha_{par},\beta_{par};\,\mathbf{R^t})=\begin{cases}(1-\rho_c)\cdot\sigma\left(\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^t})\right)+\dfrac{\rho_c}{2} & \text{if } U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;1}}\right)>U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;2}}\right)\\ (1-\rho_i)\cdot\sigma\left(\Delta U_{\alpha_{par},\beta_{par}}(\mathbf{R^t})\right)+\dfrac{\rho_i}{2} & \text{if } U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;1}}\right)\le U_{\alpha_{ppt},\beta_{ppt}}\left(\mathbf{R^{t;2}}\right)\end{cases}$$

$$\theta^t_{\alpha_{par},\beta_{par}}=\frac{p(d^t|\alpha_{par},\beta_{par};\,\mathbf{R^t})\theta^{t-1}_{\alpha_{par},\beta_{par}}}{\sum_{\alpha'_{par},\beta'_{par}}p(d^t|\alpha'_{par},\beta'_{par};\,\mathbf{R^t})\theta^{t-1}_{\alpha'_{par},\beta'_{par}}}$$

Participant does not use their own SVO to learn about the partner (Model 10).

The decision $D^t = \{d^1, d^2, \dots, d^t\}$ a partner made is used to update a participant's beliefs about a partner's α and β, $\theta^t_{\alpha_{par},\beta_{par}}$. However, in contrast to previous models, a participant's prior over a partner's initial probabilistic parameters α_{par} & β_{par} was parametrised by a new α & β which formed the central tendency of their beliefs (α_2^m & β_2^m). This new joint probability was given a particular standard deviation along each dimension (α^σ & β^σ). Together, this formed the adjusted initial joint distribution of their beliefs about a partner's SVO, although all inference and probability matching occurred as usual. This replaces equation 3.

(A.13)

$$p(\alpha_{par}|D^0) \sim N(\alpha_{par}; \alpha_2^m, \alpha^\sigma)$$

$$p(\beta_{par}|D^0) \sim N\left(\beta_{par}; \beta_2^m, \beta^\sigma\right)$$

$$p\left(\alpha_{par}, \beta_{par} \,\middle|\, D^0\right) = p\left(\alpha_{par}|D^0\right) p(\beta_{par}|D^0)$$

Text B.1: Competing model heuristic formalism.

We calculated a participant's preferences using the same framework as models 1-10.

We then constructed a variation on the classic Q-learning model (Watkins, 1989; Watkins & Dayan, 1992) that computes the subjective internal value of choice types in the environment in phase 2. The classic model computes an option-value for each option Q_a^t , in our case where $a \in \{P, I, C\}$ ranges over the three possible categorical social-value decisions a partner might make (prosocial, individualistic, competitive), rather than the actual values of the options. The values were initialized to $Q_a^0 = 1/3$ (the mean reward expected given that each potential SVO choice has equal probability of giving a 1 – correct - or 0 – incorrect - outcome). Then if the participant predicted option $\hat{a}^t$ on trial t and the partner chose option a^t, the value of the chosen option is updated according to:

(B.1)

$$Q_{\hat{a}_t}^{t+1} = Q_{\hat{a}_t}^t + \lambda * (r^t - Q_{\hat{a}_t}^t)$$

where $r^t = 1$ if $\hat{a}^t = a^t$ (i.e., the participant's prediction was correct) and $r^t = 0$ otherwise and λ is the learning rate. In model 11, we applied the same learning rate to all 36 trials in phase 2. In model 12, we allowed for different learning rates according to whether the participant was incorrect or correct [λ_{neg}, λ_{pos}]. In model 13, we had different learning rates [λ_P, λ_I, λ_C] for each of the three possible choices $\hat{a}^t$ of the participant on trial t. In model 14, there were separate learning rates [λ_c, λ_i] according to whether the partner's choice was congruent or incongruent with the participant's SVO, using the Phase 1 inference process of the Bayesian model to infer the participant's preferences. Finally, in model 15 we added a consonance parameter (ω; where 0 < ω < 1) that initialised Q_a^0 according to the paradigmatic partner type the participant was most like, given the mode of their categorical decisions in Phase 1, and $(1-\omega)/2$ to the Q_a^0 of the other two partners.

The prediction of the participant for trial $t+1$ was then calculated using a softmax function of the values Q_a^t subject to a decision temperature, $\boldsymbol{\tau}$:

(B.2)

$$p(\hat{a}^{t+1} = a|D^t) = \frac{\exp\left(\frac{Q_a^t}{\tau}\right)}{\sum_{a' \in \{P,I,C\}} \exp\left(\frac{Q_{a'}^t}{\tau}\right)}$$