Self-Other Generalisation Shapes Social Interaction and Is Disrupted in Borderline Personality Disorder

Barnby, J.M.* [1,2], Nguyen, J.[1], Griem, J.[3,4], Wloszek, M. [4], Burgess, H.[5], Richards, L.[5], Kingston, J.[1], Cooper, G.[1], London Personality and Mood Disorders Consortium, Montague, P. R.[6], Dayan, P.[7,8], Nolte, T. [3,4], Fonagy, P. [3,4]

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

[2]School of Psychiatry and Clinical Neuroscience, University of Western Australia, AU

[3]Department for Clinical, Educational, and Health Psychology, Division of Psychology and Language Sciences, University College London, London, UK

[4]Anna Freud, London, UK

[5]Department of Neuroscience, Washington University in St. Louis, St. Louis, MO, USA

[6]Centre for Human Neuroscience Research, Virginia Tech, VA, USA

[7]Max Planck Institute of Biological Cybernetics, Tübingen, DE,

[8]University of Tübingen, Tübingen, DE

*corresponding: joseph.barnby@rhul.ac.uk

Abstract. Generalising information from ourselves to others, and others to ourselves allows for both a dependable source of navigation and adaptability in interpersonal exchange. Disturbances to social development in sensitive periods can cause enduring and distressing damage to lasting healthy relationships. However, identifying the mechanisms of healthy exchange has been difficult. We introduce a theory of self-other generalisation tested with data from a three-phase social value orientation task - the Intentions Game. We involved individuals with (n=50) and without (n=53) a diagnosis of borderline personality disorder and assessed whether self-other information generalisation may explain interpersonal (in)stability. Healthy controls initially used their preferences to predict others and were influenced by their partners, leading to self-other convergence. In contrast, individuals with borderline personality disorder maintained distinct self-other representations, generating a new neutral prior to begin learning. Both groups steadily reduced their updating over time, with healthy participants showing increased sensitivity to update beliefs. Furthermore, we explored theory-driven individual differences underpinning learning. Overall, the findings provide a clear explanation of how self-other generalisation constrains and assists learning, how childhood adversity disrupts this through separation of internalised beliefs and makes clear predictions about the mechanisms of social information integration under uncertainty.

Keywords: Interpersonal Generalisation; Computational Modelling; Social Learning; Mentalizing; Borderline Personality Disorder; Trauma; Maltreatment

Open data and code:
https://github.com/josephmbarnby/SocialTransfer_Barnby_etal_2024

Highlights

- Humans use self-to-other transfer to constrain initial predictions about the social behaviour of others
- Information is transferred from other-to-self following observation, calibrated by the precision of beliefs
- Learning about others is characterized by initially fast updating followed by slower updating schedules.
- Those diagnosed with BPD do not engage in self-other information transfer, instead keeping self and other representationally distinct
- Higher reported childhood trauma, paranoia, and poorer trait mentalizing all diminish other-to-self information transfer irrespective of diagnosis

Introduction

Social animals have evolved sophisticated mechanisms for cooperation. To coordinate, individuals exchange information, enabling independent and group regulation (Emerson, 1956; Wheeler, 1911). Interaction provides the self with insights about others and, simultaneously, about its own state within the environment. In humans, this psychological differentiation begins in utero (Ciaunica et al., 2021) and continues into adulthood. This prolonged differentiation facilitates the distinction between self and others, allowing grounded social orientation and flexible adaptation. Early psychological theorists highlighted the importance of healthy social exchange in humans for integrating constructive relational blueprints, noting that disruptions during sensitive periods can impair the formation of trusting, safe social bonds (e.g., Fairbairn, 1952; 1994) and foster less adaptable interpersonal beliefs (Young et al., 2006).

At the individual level, humans must manage information exchange to navigate and reduce relational uncertainty. When faced with external uncertainty about others' characteristics, prior knowledge can swiftly and effectively guide predictions. When uncertainty arises regarding one's own internal state, external cues can provide anchoring information to support self-through-other calibration. To reduce uncertainty about others, theories of the relational self (Anderson & Chen, 2002) suggest that the self is the most extensive and well-grounded representation available, leading to a readily accessible initial belief (Kreuger & Clement, 1994) that can be projected or integrated into social learning. Conversely, to address uncertainty about the self, individuals can generalize information from others to themselves—this social contagion facilitates adaptation to social groups and is a crucial component of interpersonal cohesion that relies on trust (Frith & Frith, 2012).

Computational modelling has advanced our understanding of how humans engage in self-insertion and social contagion to achieve efficient prediction and adaptation. By integrating self-preferences into prior beliefs, self-insertion models account for increased reaction times at the onset of learning and enhanced predictive accuracy as a function of interpersonal similarity (Tarantola et al., 2017; Barnby et al., 2022). Conversely, uncertain self-preferences can be influenced by observing others' intertemporal discounting behavior (Garvert et al., 2015; Moutoussis et al., 2016; Thomas et al., 2022).

However, critical questions remain: How do humans adjudicate between self-insertion and contagion during interaction to manage interpersonal generalization? Does the uncertainty in self-other beliefs affect their generalizability? How can disruptions in interpersonal exchange during sensitive developmental periods (e.g., childhood maltreatment) inform models of psychiatric disorders? Understanding the computational processes humans use in social exchange has broad implications for theories of healthy childhood development, group cohesion, reputation management, interpersonal synchrony, and the breakdown of social bonds.

In this study, we present a formal account of self-other generalization in healthy individuals and those seeking psychiatric support, tested through an interactive social economic paradigm—the Intentions Game—and a computational model that concurrently allows for self-insertion and social contagion. We evaluate our model

using data from matched participants with and without a diagnosis of Borderline Personality Disorder (BPD). BPD is characterized by interpersonal sensitivity, relational instability, emotional dysregulation, impulsivity, paranoia, and severe fear of abandonment (e.g., Gunderson et al., 2018; Euler et al., 2021). It is strongly associated with early childhood adversity, such as psychological, physical, and sexual abuse or neglect (Afifi et al., 2011; Bateman et al., 2023), and inconsistent parenting experiences (Crawford et al., 2009), particularly during sensitive developmental periods (Fonagy & Bateman, 2008). Early adversity interacts with pre-existing variations in stress-homeostasis mechanisms, exacerbating interpersonal disruptions (Pratt et al., 2017), underpinned by difficulties in mentalizing and trust appraisal, leading to disruptions in social learning (e.g., Fonagy & Luyten, 2009; Nolte et al., 2023) and representations of self and other (Hanegraaf et al., 2021).

Computational models have probed social processes in BPD, linking the BPD phenotype to a potential over-reliance on social versus internal cues (Henco et al., 2020), ‘splitting’ of social latent states that encode beliefs about others (Story et al., 2023), negative appraisal of interpersonal experiences with heightened self-blame (Mancinelli et al., 2024), inaccurate inferences about others’ irritability (Hula et al., 2018), and reduced belief adaptation in social learning contexts (Siegel et al., 2020). Previous studies have typically overlooked how self and other are represented in tandem, prompting further investigation into why any of these BPD phenotypes manifest.

We propose a theory with testable predictions to begin addressing this question, outlining that information generalization is foundational to healthy and evolving social bonds, and that the BPD phenotype may arise from an infraction to this process. To foreshadow our results we discover that healthy participants employ a mixed process of self-insertion and contagion to predict and align with the beliefs of their partners in the Intentions Game. In contrast, individuals with BPD exhibit distinct, disintegrated representations of self and other, despite showing similar average accuracy in their predictions about partners. Our model and data suggest that the previously observed computational characteristics in BPD, such as reduced self-anchoring during ambiguous learning and a relative impermeability of the self, arise from the failure of information about others to transfer and inform the self. By integrating separate computational findings, we provide a foundational model and a concise, dynamic paradigm to investigate uncertainty, generalization, and regulation in social interactions. Additionally, we examine the extent to which self-reported complex trauma and its sequelae are linked to these computational processes.

Results

Healthy participants (CON; n=53) and participants diagnosed with BPD (n=50), matched on age, gender, education, and social deprivation indices (Table 1), were invited to participate in a three-phase social value orientation paradigm—the Intentions Game (**Figure 1A**)—with virtual partners. In the first phase, participants made forced choices between two options for splitting points with an anonymous partner. In the second phase, participants learned to predict the decisions of a new anonymous partner using the same forced-choice setup, receiving feedback on the accuracy of their successive predictions. Notably, using a novel server architecture (Burgess et al., 2023), partners in phase 2 were configured to be approximately 50% different from the participants in terms of their choices, ensuring that all participants had to learn about their phase 2 partner. The third phase mirrored the first, with participants informed that they were matched with a third anonymous partner, unrelated to those in phases 1 and 2. Detailed descriptions of the task can be found in the methods section.

Psychometric and Behavioural Results

Participants with BPD, compared to CON, reported significant childhood trauma, epistemic disruptions (including mistrust and credulity), elevated referential and persecutory beliefs, and demonstrated ineffective trait mentalizing (Table 1). The groups did not differ in trait measures of certainty regarding self and others' mental states, nor in elevated trust.

We analyzed the 'types' of choices participants made in each phase (**Supplementary Table 1**). For example, a participant could make prosocial (self=5; other=5) versus individualistic (self=10; other=5) choices, or prosocial (self=10; other=10) versus competitive (self=10; other=5) choices. There were 12 types of pairs in phases 1 and 3 (individualistic vs. prosocial; prosocial vs. competitive; individualistic vs. competitive).

In phase 1, both CON and BPD participants made prosocial choices over competitive choices with similar frequency (CON=9.67[3.62]; BPD=9.60[3.57]). However, CON participants made significantly fewer prosocial choices when individualistic choices were available (CON=2.87[4.01]; BPD=5.22[4.54]; t=2.75, p=0.007). Both groups favored individualistic over competitive choices with similar frequency (CON=11.03[1.95]; BPD=10.34[2.63]). Examining reaction times (in milliseconds; ms) in phase 1 by choice type revealed that, compared to competitive choices, individualistic choices were made faster (linear estimate = -880.60, 95%CI: -1385.42, -376.2; t = -3.42, p < 0.001), and prosocial choices were made fastest (linear estimate = -1171.1, 95%CI: -1701.97, -640.71; t = -4.32, p < 0.001) irrespective of the type of choice pair. Prosocial choices were made significantly faster than individualistic choices (linear estimate = -290.70, 95%CI: -548.50, -32.91; t = -2.21, p = 0.027). There was no difference in reaction times between CON and BPD participants in phase 1.

In phase 2 each group showed good predictive accuracy (CON=77.2%[13.9%]; BPD=72.7%[15.6%]). There was no difference in overall predictive accuracy between BPD and CON (linear estimate=2.44, 95%CI: -0.67, 5.54; t=1.56; p=0.12), nor when analysed on a trial-by-trial basis (linear estimate=0.26, 95%CI: -0.06, 0.59; z=1.61, p=0.11) using a random effects models. All participants showed an effect of time on

accuracy, such that participants became more accurate in predicting their partner over the course of phase 2 (linear estimate=0.013, 95%CI: 0.008, 0.017; z=6.01; p<0.001). There was no impact of choice type on reaction times in phase 2.

In phase 3, both CON and BPD participants continued to make equally frequent prosocial versus competitive choices (CON=9.15[3.91]; BPD=9.38[3.31]), and CON participants continued to make significantly less prosocial versus individualistic choices (CON=2.03[3.45]; BPD=3.78 [4.16]; t=2.31, p=0.02). Both groups chose equally frequent individualistic versus competitive choices (CON=10.91[2.40]; BPD=10.18[2.72]). Reaction times in phase 3 revealed that compared to competitive choices, individualistic choices were made faster (linear estimate = -528.50, 95%CI: -943.60, -114.6; t = -2.50, p = 0.012), and prosocial choices were made fastest (linear estimate = -693.5, 95%CI: -1137.65, -250.39; t = -3.07, p = 0.002), irrespective of the type of choice pair. Prosocial choices were no longer executed significantly faster than individualistic choices. All participants made faster choices in phase 3 compared to phase 1 (linear estimate = -242.02, 95% CI: -332.64, -151.41; t = -5.24, p = 0.001). There was no significant effect of group on reaction times between phases 1 and 3, nor within phase 3 when analyzed independently.

Table 1. Demographics of participants. CTQ=Childhood Trauma Questionnaire, MZQ = Mentalisation Questionnaire, RGPTSB=Revised Green Paranoid Thoughts Scale (Persecutory Subscale), RGPTSA=Revised Green Paranoid Thoughts Scale (Referential Subscale), CAMSQ=Certainty About Mental States Questionnaire. ETMCQ=Epistemic Trust, Mistrust and Credulity Questionnaire, M=Male, F=Female, O=Other. For continuous variables, all means are stated with corresponding standard deviations in brackets. Significant differences are highlighted in **bold.**

	BPD	**CON**	**Test**
DEMOGRAPHICS			
n	50	53	
Age	31.2[11.16]	30.0[8.64]	t=-0.61; p=0.54
Gender (M:F:O)	8:39:2	14:36:3	t=-1.25; p=0.21
Education (Years)	14.39(3.38)	14.8[3.10]	t=0.62; p=0.53
Soc. Dep. Index	12834.94[7911]	11967.63[7567]	t=-0.54; p=0.59
PSYCHOMETRICS			
RGPTSB	14.53[1.07]	5.33[0.72]	**t=4.67; p<0.001**
RGPTSA	16.79[0.96]	7.29[0.70]	**t=6.47; p<0.001**
CTQ (Trauma)	64.88[0.90]	42.27[0.78]	**t=6.48; p<0.001**
CAMSQ (Self)	4.95[1.42]	5.02[0.22]	t=-0.07; p=0.94
CAMSQ (Other)	5.32[1.42]	5.15[0.17]	t=-0.71; p=0.48
MZQ (Mentalizing)	55.94[0.59]	38.16[0.85]	**t=9.39; p<0.001**
ETMCQ (Mistrust)	5.25[0.93]	4.21[0.84]	**t=5.27; p<0.001**
ETMCQ (Trust)	4.86[1.13]	5.04[0.86]	t=-0.87, p=0.39
ETMCQ (Credulity)	4.35[1.06]	3.34[0.77]	t**=4.14; p<0.001**

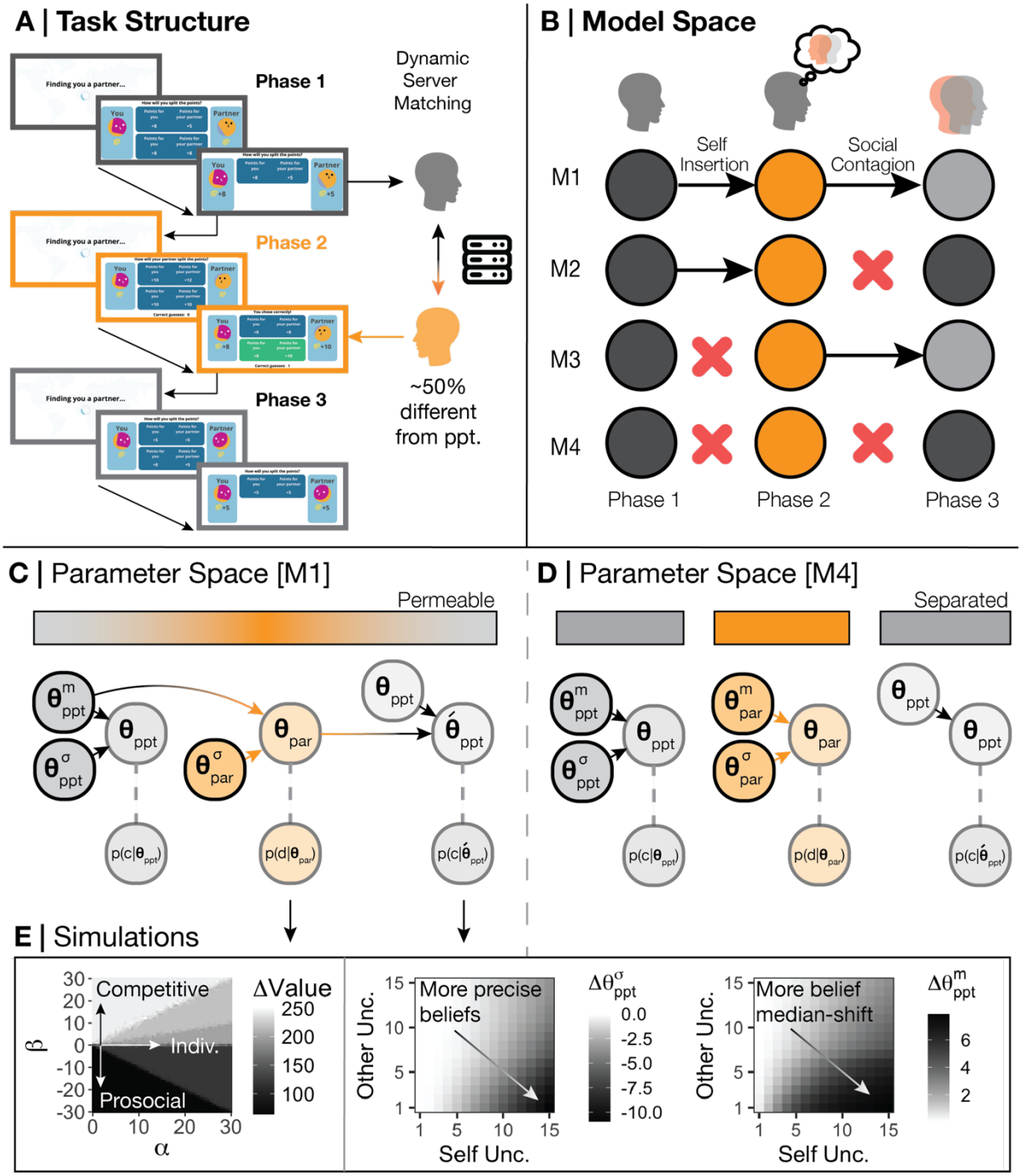

A | Task Structure
Phase 1
Phase 2
Phase 3
Dynamic Server Matching
~50% different from ppt.
B | Model Space
Self Insertion
Social Contagion
M1
M2
M3
M4
Phase 1
Phase 2
Phase 3
C | Parameter Space [M1]
Permeable
D | Parameter Space [M4]
Separated
E | Simulations
Competitive
Indiv.
Prosocial
ΔValue
250
200
150
100
β
α
More precise beliefs
More belief median-shift
Other Unc.
Self Unc.

Figure 1. Task and Model Space. (A) Participants were invited to play a three-phase, repeated social value orientation paradigm—the Intentions Game—with virtual partners. Phase 1 of the Intentions Game lasted 36 trials and asks participants to make a forced choice between two options as to how to split points with an anonymous virtual partner. An example of a prosocial-individualistic pair of options could be (self=5, other=5) or (self=10, other=5) – if the participant chooses option 1 they could be viewed as less individualistic and more prosocial as the outcomes to the other do not change, but the self would earn less. In phase 2, lasting 54 trials, participants were asked to predict the decisions of a new anonymous partner using the same two-forced choice set-up and the same option pairs; participants were given feedback on whether they were correct or incorrect in their prediction. We used Amazon Web Services to create a novel server architecture to match participants and (virtual) partners (Burgess et al., 2023). Partners in phase 2 were matched to be approximately 50% different from the participant with respect to their choices in phase 1 to ensure all participants needed to learn about their phase 2 partner, and to provide a mechanism to examine whether beliefs about partners had an effect on the self. Phase 3 was identical to phase 1, although participants were informed that they were matched with a third anonymous partner, unconnected to the partners in phase 1 and 2. At the end of the game, if participants collected over 1000 points overall, they were entered into a lottery to win a bonus. **(B)** We created four models that may explain the data and to test theories of social generalization. Model M1 assumes participants are subject to both self-insertion and social-contagion, that is, participants used their own preferences as a prior about their partner in phase 2, and partner behaviour subsequently influenced participant's preferences in phase 3. Model M4 assumes participants are subject to neither self-insertion nor social contagion, instead forming a novel prior around the phase 2 partner rather than using their own preferences and failing to be influenced by their partner after observation. Models M2 and M3 suggest participants are only explained by either self-insertion or social-contagion, not both. **(C)** We assume that participants choices in phase 1 are governed by both a median (θ_{ppt}^{m}) and standard deviation (θ_{ppt}^{σ}). Participants insert their median preferences (θ_{ppt}^{m}) into their prior beliefs over their partner in phase 2, but with a different standard deviation to allow for flexibility and learning (θ_{ref}^{σ}). The combination of the prior and posterior belief uncertainty about the partner (θ_{ref}^{σ}; θ_{par}^{σ}), the precision participants have over their own preferences (θ_{ppt}^{σ}), and the median posterior of the participant and partner (θ_{ppt}^{m}; θ_{par}^{m}) form the new median and standard deviation over participant preferences in phase 3 ($\acute{\theta}_{ppt}$). **(D)** In contrast to M1, M4 generates a new central tendency over the partner in phase 2 ($\bar{\theta}_{par}^{m}$) which disconnects participant preferences and prior beliefs. M4 also assumes that the same parameters that generated participant choices in phase 1 also generate choices in phase 3. **(E)** Simulating our model demonstrates how different combinations of α (preferences for absolute self-reward) and β (preferences for relative reward; prosocial-competitiveness) lead to changes in the discrepancy of value between participants and partners (left panel). We also show how increasing uncertainty over self-beliefs, and higher precision over partners, causally draws participants more toward the beliefs of their partner in phase 3 and increases their precision over their phase 3 beliefs (Moutoussis et al., 2016).

Computational Analysis

Over all three phases, we assumed participants and their partners used a Fehr-Schmidt utility function (Fehr & Schmidt, 1999) to calculate the utility of two options ($\mathbf{U}_{\alpha,\beta} = \{\mathrm{U}^1_{\alpha,\beta}, \mathrm{U}^2_{\alpha,\beta}\}$), based on the joint rewards available for both the participant, $\mathbf{R}_{ppt}$ = $\{\mathrm{r}^1_{ppt},\ \mathrm{r}^2_{ppt}\}$ and their partner, $\mathbf{R}_{par} = \{\mathrm{r}^1_{par},\ \mathrm{r}^2_{par}\}$. The utility of each option was weighted based on absolute-reward gain α (how much participants care about self-earnings) and relative reward β along a prosocial-competitive axis (how much participants care about equality of outcomes); $\beta < 0$ is prosocial, whereas $\beta > 0$ is competitive.

$$\mathbf{U}_{\alpha,\beta} = \alpha * \mathbf{R}_{ppt} + \beta * \max(\mathbf{R}_{ppt} - \mathbf{R}_{par}, 0)$$

We then constructed four models to explain how participants used their own preferences ($\theta^m_{ppt} = \{\alpha^m_{ppt}, \beta^m_{ppt}\}$) and uncertainty over these preferences ($\theta^\sigma_{ppt} = \{\alpha^\sigma_{ppt}, \beta^\sigma_{ppt}\}$) to predict and learn about the preferences of their partner (θ_{par}; **Figure 1B**; see methods). Model M1 (**Figure 1C**) suggests that participants initially use their own preferences as a prior belief about their partner (self-insertion), which is gradually diminished during the learning process in phase 2. M1 also posits that, following learning, the inferred beliefs about a partner will influence participants' own preferences, making them more similar to their partner's preferences following observation (social contagion). According to this model, participants shift towards their partner based on their uncertainty about self and others (**Figure 1E**): greater uncertainty over self-preferences and increased precision in representing the other cause stronger social contagion effects.

Model M4 (**Figure 1D**), on the other hand, suggests that participants do not engage in these generalization processes: predictions about others are not grounded in the self, and observing others does not alter self-preferences. Models M2 and M3 allow for either self-insertion or social contagion to occur independently. Consistent with prior research, we also constructed a model that assumes the same insertion and contagion processes as M1, but along a single prosocial-competitive axis ('Beta model'; Barnby et al., 2022). The 'Beta model' accommodates the possibility that participants might only consider a single dimension of joint reward allocation, which is typically emphasized in previous studies (e.g., Hula et al., 2018).

Table 2. Parameter and model specification. Grey shading = parameters relevant to representations of the self (ppt). Orange shading = parameters relevant to representations of the other (par). Free = parameters are random variables to fit through model inversion. Derived = parameter is calculated from latent values within the model. SD = standard deviation.

	M1	M2	M3	M4	Beta	Description	Type	Phase
α_{ppt}^{m}	X	X	X	X		Median of absolute reward preferences	Free	1
β_{ppt}^{m}	X	X	X	X	X	Median of relative reward preferences	Free	1
α_{ppt}^{σ}	X	X	X	X		SD of absolute reward preferences	Free	1
β_{ppt}^{σ}	X	X	X	X	X	SD of relative reward preferences	Free	1
$\bar{\alpha}_{par}^{m}$			X	X		Prior beliefs median over absolute reward preferences	Free	2
$\bar{\beta}_{par}^{m}$			X	X		Prior beliefs median over relative reward preferences	Free	2
α_{par}^{ref}	X	X	X	X		Prior beliefs SD over absolute reward preferences	Free	2
β_{par}^{ref}	X	X	X	X	X	Prior beliefs SD over relative reward preferences	Free	2
α_{par}^{m}		X	X	X		Posterior belief median over absolute reward preferences	Derived	2
β_{par}^{m}		X	X	X	X	Posterior belief median over relative reward preferences	Derived	2
α_{par}^{σ}		X	X	X		Posterior belief SD over absolute reward preferences	Derived	2
β_{par}^{σ}		X	X	X	X	Posterior belief SD over relative reward preferences	Derived	2
$\acute{\alpha}_{ppt}^{m}$	X		X			Shifted median of absolute reward preferences	Derived	3
$\acute{\beta}_{ppt}^{m}$	X		X		X	Shifted median of relative reward preferences	Derived	3
$\acute{\alpha}_{ppt}^{\sigma}$	X		X			Shifted SD of absolute reward preferences	Derived	3
$\acute{\beta}_{ppt}^{\sigma}$	X		X		X	Shifted SD of relative reward preferences	Derived	3
No. Free	6	6	8	8	3			

Model Comparison – BPD Participants Hold Disintegrated Self-Other Beliefs

We found that CON participants were best fit at the group level by M1 (Frequency = 0.59, Protected Exceedance Probability = 0.98), whereas BPD participants were best fit by M4 (Frequency = 0.54, Protected Exceedance Probability = 0.86; **Figure 2A**). Consequently, we analyzed common parameters between groups, assuming that M1 and M4 best fitted the CON and BPD groups, respectively. It is worth noting that a minority of participants were best fit by M3 (CON: Frequency = 0.32; BPD: Frequency = 0.39), which assumes that participants were influenced by their partner but were not subject to self-insertion biases. We therefore also examined the change in beliefs between phases 1 and 3 under the assumption that M3 was accurate. Anticipating this analysis, we find that our main conclusions hold, showing that BPD participants were significantly less influenced by their partner compared to CON participants.

Generative Accuracy and Recovery

We simulated data for each participant using their individual parameters from the winning model within each group and refitted our models using this simulated data. Model comparison yielded very similar results (**Figure 3A**), with CON synthetic participants best fit at the group level by M1 (Frequency = 0.58, Protected Exceedance Probability = 0.98) and BPD synthetic participants best fit by M4 (Frequency = 0.57, Protected Exceedance Probability = 0.85). The simulated data closely matched the actions of participants across all three phases (median accuracy = 0.8, SD = 0.12). In phase 2, the model-predicted total correct scores were not significantly different from observed scores (**Figure 3E**). Both model responsibility and common parameters within each dominant model were strongly and significantly associated (model confusion $\rho = 0.46–0.97$, $p < 0.001$; parameter recovery $\rho = 0.70–0.94$, $p < 0.001$; **Figure 3C**). Given the very good to excellent performance of the models, we continued to analyze individual parameters and simulations across each group.

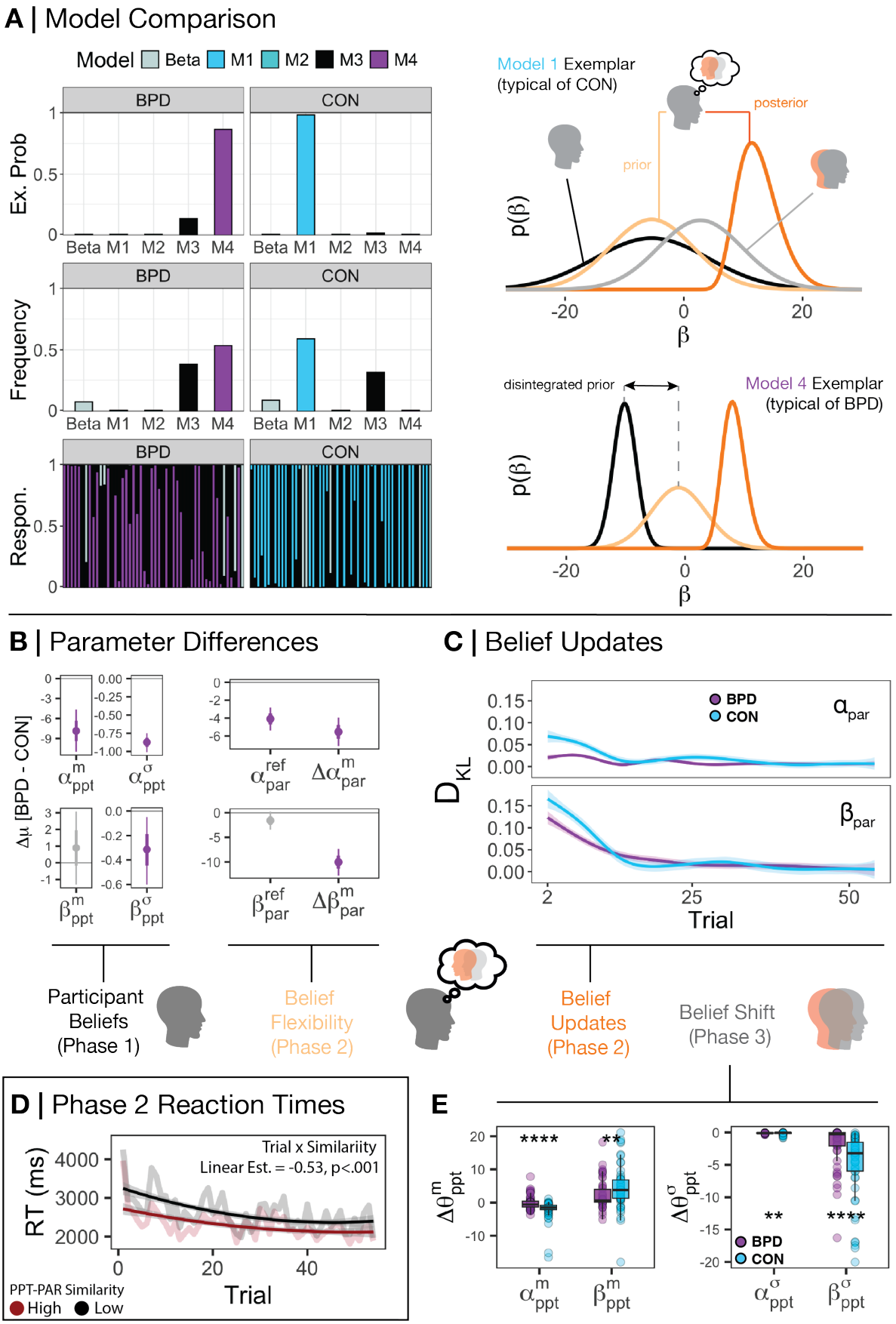

A | Model Comparison
Model Beta M1 M2 M3 M4
BPD
CON
Ex. Prob
Frequency
Respon.
Beta M1 M2 M3 M4
Model 1 Exemplar
(typical of CON)
posterior
prior
p(β)
β
-20
0
20
disintegrated prior
Model 4 Exemplar
(typical of BPD)
B | Parameter Differences
Δμ [BPD - CON]
Participant
Beliefs
(Phase 1)
Belief
Flexibility
(Phase 2)
C | Belief Updates
BPD
CON
Trial
Belief
Updates
(Phase 2)
Belief Shift
(Phase 3)
D | Phase 2 Reaction Times
RT (ms)
Trial x Similariity
Linear Est. = -0.53, p<.001
PPT-PAR Similarity
High Low
E
BPD
CON

Figure 2. Beliefs between groups and within phases. (A) We used random-effects hierarchical model fitting and comparison to jointly estimate group level and individual level parameters based on real data from participants (Piray et al., 2021). CON participants were best fit by M1, whereas BPD participants were best fit by M4 on a group level. Looking within each model by simulating the beliefs of each participant reveals that – as expected – CON participants use the median of their self-preferences (black distribution) as a basis for their prior beliefs about partners (light orange distribution), and that the precision of their posterior beliefs about partners (dark orange distribution) and the precision of their own self preferences leads to a shifted model of the self (grey distribution). BPD participants on the other hand have a disintegrated prior over their partner which is not subject to their own self representation. Likewise, there is no change in self-preferences following learning, and thus an absence of the light grey distribution. For illustration, we focus on beliefs over relative preferences (β) and use real individual participants as exemplars for illustration. **(B)** Across models we extracted the common parameters that generate the behaviour of both CON and BPD participants – that is, their median and standard deviation over both α (absolute reward preferences) and β (relative value preferences), the flexibility over participants' prior beliefs about their partners over each dimension, and the absolute change in posterior beliefs in phase 2 over each dimension ($\Delta\alpha^{m}_{par}$;$\Delta\beta^{m}_{par}$). Using hierarchical Bayesian t-tests we demonstrated the mean difference in parameter values between groups. Purple values lower than 0 indicate that the BPD participants had significantly smaller parameter values. Here we find that BPD participants were less individualistic, equally prosocial, and more certain about their self-preferences. BPD participants were also less flexible over their beliefs about a partner's absolute reward preferences and updated their beliefs less across the board. **(C)** We also calculated the Kullback-Leibler divergence (D_{KL}) of beliefs between each trial (*t*-1 vs *t*) on each trial during phase 2. We observed three things: 1. All participants display more sensitive updates initially, 2. all participants 'cool off' in their sensitivity over the course of phase 2, and 3. BPD participants make significantly less sensitive updating throughout the course of phase 2 vs. CON participants. **(D)** Examining reaction times of participants over phase 2 revealed that participants became faster at making predictions as phase 2 continued. We also find that participant similarity interacted with trial to change reaction times, such that higher participant-partner similarity reduced reaction times in earlier trials but this difference was attenuated over time. Participant(PPT)-partner(PAR) similarity was calculated as the combined distance between participant and partner parameters determined by server matching along absolute (α) and relative reward (β) axes. Similarity was visualised as dichotomous for illustration but treated as a continuous variable in our analyses) **(E)** Examining participants under a blanket assumption that participants in both BPD and CON groups were influenced by their partner revealed that BPD participants were significantly less influenced by their partner across the board, both with respect to their phase 3 median and standard deviation of beliefs. Kruskal-Wallis tests were used between groups within the visualisation. *=p<0.05, **=p<0.01, ***=p<0.001, ****=p<0.0001.

Phase 1 – BPD Participants Are More Certain About Themselves

We first examined self-representations of participants in phase 1. CON participants and BPD participants were equally prosocial (CON mean$[\beta^m_{ppt}]$ = -7.50; BPD mean$\left[\beta^m_{ppt}\right]$ = -6.59; $\Delta\mu[\beta^m_{ppt}]$ = 0.92, 95%HDI: -1.24, 3.12) – both groups valued equal allocation of reward between themselves and their partners. BPD participants had lower preferences for earning higher absolute rewards (CON mean$[\alpha^m_{ppt}]$=18.41; BPD mean$\left[\alpha^m_{ppt}\right]$ = 10.57; $\Delta\mu[\alpha^m_{ppt}]$ = -7.83, 95%HDI: -11.06, -4.75). BPD participants were also more certain about both types of preference ($\Delta\mu[\alpha^\sigma_{ppt}]$ = -0.89, 95%HDI: -1.01, -0.75; $\Delta\mu[\beta^\sigma_{ppt}]$ = -0.32, 95%HDI: -0.60, -0.04) versus CON participants (Figure 2B).

Phase 2 – BPD Participants Use Neutral Priors And Form Rigid Beliefs

We next assessed how participants generated their prior beliefs about a partner in phase 2. CON participants were best fit by M1 which assumes the same median belief participants use in phase 1 is identical to their median prior belief about their partners. In contrast, BPD participants were best fit by M4 and generated a new median prior belief about their partners.

In BPD participants, only new beliefs about the relative preferences of partners (prosocial-competitive axis) differed - new median priors were larger than median preferences in phase 1 (mean$[\beta^m_{par}]$ = -0.47; $\Delta\mu[\beta^m_{ppt} - \beta^m_{par}]$ = -6.10, 95%HDI: -7.60, -4.60). BPD priors about their partner's relative preferences were also centred closely around 0 ($\Delta\mu[0 - \beta^m_{par}]$ = -0.39, 95%HDI: -0.77, -0.05), suggesting that BPD participants entered into the interaction with very neutral priors about their partner.

BPD participants were equally flexible around their prior beliefs about a partner's relative reward preferences ($\Delta\mu[\beta^{ref}_{par}]$ = -1.60, 95%HDI: -3.42, 0.23), and were less flexible around their beliefs about a partner's absolute reward preferences ($\Delta\mu[\alpha^{ref}_{par}]$=-4.09, 95%HDI: -5.37, -2.80), versus their CON counterparts. (**Figure 2B**).

Belief updating in phase 2 was substantially less adaptive in BPD participants across the board. The median change in beliefs (from priors to posteriors) about a partner's preferences was lower versus. controls ($\Delta\mu[\Delta\alpha^m_{par}]$ = -5.53, 95%HDI: -7.20, -4.00; $\Delta\mu[\Delta\beta^m_{par}]$ = -10.02, 95%HDI: -12.81, -7.30). Posterior beliefs about partner were also more rigid in BPD versus CON ($\Delta\mu[\alpha^\sigma_{par}]$ = -0.94, 95%HDI: -1.50, -0.45; $\Delta\mu[\beta^\sigma_{par}]$ = -0.70, 95%HDI: -1.20, -0.25). This is perhaps unsurprising given the disintegrated priors of the BPD group, meaning they need to 'travel less' and thus have longer to converge on the beliefs of their partner.

Analysing belief updating on a more granular trial-by-trial basis revealed axial and group differences in belief refinement over the course of phase 2 (**Figure 2C**). We examined this by analysing the Kullback-Leibler divergence (D_{KL}) of beliefs between each trial in Phase 2, from *t*-1 to *t* over trial 1-54, using random-effect linear models.

Across both groups and belief types, the magnitude of belief updating reduced over time (linear estimate$[\mathrm{D}_{KL}]$ = -0.007, 95%CI: -0.008, -0.005; t = -7.60, p < 0.001). Beliefs about a partner's relative reward preferences were updated more vs. absolute reward

preferences (linear estimate= 0.54, 95%CI: 0.47, 0.62; t = 14.00, p < 0.001). These interacted, such that initial flexibility over relative vs. absolute beliefs reduced over the course of phase 2 (linear estimate = -0.013, 95%CI: -0.015, -0.011; t = -10.81, p < 0.001).

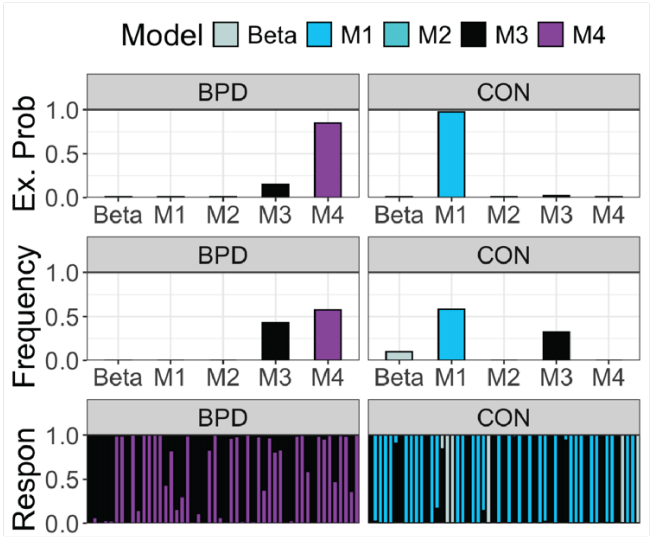
CON participants remained more flexible than BPD participants along both axes (linear estimate$\left[\mathrm{D}_{KL}(\alpha_{par}^{m})\right]$ = 0.40, 95%CI: 0.29, 0.51, t = 7.18, p < 0.001; linear estimate$\left[\mathrm{D}_{KL}(\beta_{par}^{m})\right]$ = 0.17, 95%CI: 0.29, 0.51, t=3.06, p=0.002). This interacted over time, such that the difference between groups, and the magnitude of belief updates decreased over time (linear estimate$\left[\mathrm{D}_{KL}(\alpha_{par}^{m})\right]$ = -0.009, 95%CI:-0.012,-0.006, t = -5.30, p < 0.001; linear estimate$\left[\mathrm{D}_{KL}(\beta_{par}^{m})\right]$ = -0.004, 95%CI:-0.008,-0.001, t = -2.78, p = 0.005) - CON participants and BPD participants eventually converged to an equivalent updating schedule (**Figure 2C**). Analyses of phase 2 belief updating suggests posterior beliefs are generally less sensitive to change in BPD versus CON.

Beliefs in phase 2 penetrated into reaction times (**Figure 2D**): all participants were slower at the start of phase 2 and sped up over time (linear estimate = -15.03, 95%CI: -21.06, -8.99; t = -4.88, p < 0.001). Baseline participant-partner similarity did not have an overall effect on reaction time but did interact with trial – as participant-partner similarity increased, reaction times early in phase 2 were significantly slower and this effect attenuated over time (linear estimate = -0.53, 95%CI: -0.75, -0.32; t = -4.91, p < 0.001). Reaction time did not vary between groups: both BPD and CON participants displayed the same effect. We also show that reaction times and belief updates in phase 2 were significantly coupled, such that larger shifts in posterior beliefs along both axes were associated with larger reaction times (linear estimate$\left[\mathrm{D}_{KL}(\alpha_{par}^{m})\right]$ = 0.044, 95%CI: 0.027, 0.06, t = 5.01, p < 0.001; linear estimate$\left[\mathrm{D}_{KL}(\beta_{par}^{m})\right]$ = 0.021, 95%CI: 0.005, 0.039, t = 2.49, p = 0.012; **Figure S9**).

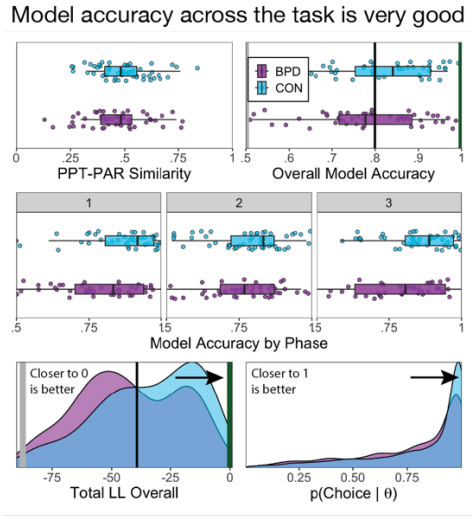
Phase 3 – BPD Participants Are Less Influenced by Partners

In the dominant model for the BPD group—M4—participants are not influenced in their phase 3 choices following exposure to their partner in phase 2. To further confirm this conclusion, we also analyzed participants under the assumption that M3 was the dominant model for both groups, considering that a minority of participants were best explained by this model. This analysis aligns with our primary model comparison (**Figure 2E**). CON participants altered their absolute median beliefs in phase 3 (linear estimate = 1.75, 95% CI: 0.73, 2.79; t = 3.36, p < 0.001) and increased their precision (linear estimate = 1.53, 95% CI: 0.65, 2.40; t = 3.43, p < 0.001) more than BPD participants. There was also an interaction with the type of belief: CON participants changed their median beliefs about relative reward along the prosocial-competitive axis more than their beliefs about absolute reward (linear estimate = 2.13, 95% CI: 0.09, 4.18; t = 2.06, p = 0.041), and became more precise along the same axis (linear estimate = 3.01, 95% CI: 1.30, 4.71; t = 3.47, p < 0.01), compared to BPD participants. This suggests that relative reward preferences are particularly resistant to change in BPD participants.

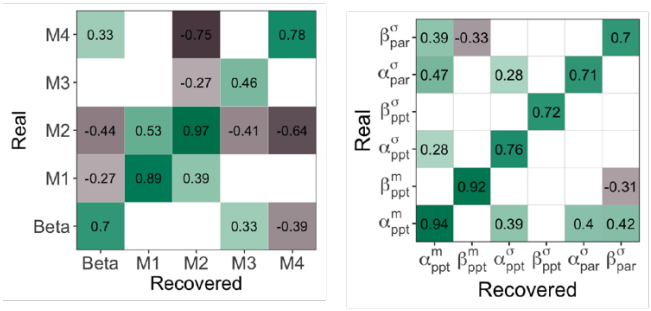
A | Model Comparison Recovery



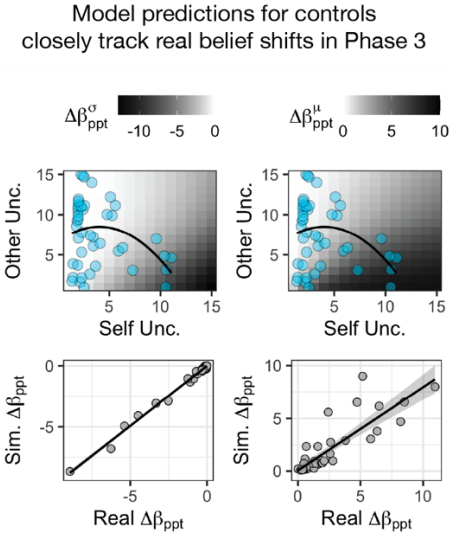
B | Generative Accuracy



C | Model Confusion & Parameter Recovery



D | Phase 3 Model Accuracy



E | Behavioural Recovery

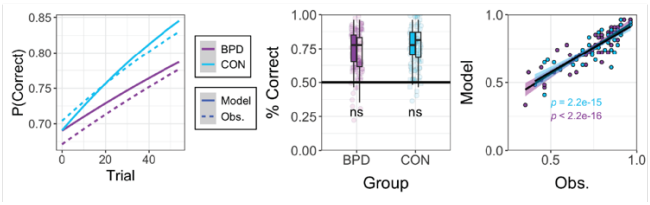


Figure 3. Model Accuracy. (A) We used random-effects hierarchical model fitting and comparison to jointly estimate group level and individual level parameters on simulated data (Piray et al., 2019). CON participants were best fit by M1, whereas BPD participants were best fit by M4 **(B)** Server matching between participant and partner in phase two was successful, with participants being approximately 50% different to their partners with respect to the choices each would have made on each trial in phase 2 (mean similarity=0.49, SD=0.12). Model accuracy across the task was very high (mean accuracy=0.8, SD=0.12). Model accuracy within each phase was very high (mean accuracy[phase1]=0.83, SD[phase1]=0.16; mean accuracy[phase2]=0.77, SD[phase2]=0.14; mean accuracy[phase3]=0.82, SD[phase3]=0.17). Loglikelihood values were also well above what would be expected had the model fitted the data by chance (median=-40.68, SD=22.7; chance value=-87.33). Choice probabilities generated by the model on each trial were also well above chance thresholds (median=0.91, SD=0.24; chance value=0.5). **(C)** The spearman association between the responsibility allocated for each participant during real and recovered model comparison was highly correlated on the diagonal. There was some correlation between M1-M2 but this was due to M2 being a nested model of M1, sharing similar free parameters; this was not worrying in light of excellent model identifiability overall in the synthetic comparison. Associations between real and recovered parameters from the dominant model within each BPD and CON participants was very high with few cross correlations on the off-diagonal. In both confusion and parameter recovery matrices, white spaces indicate insignificant associations at the $p > 0.01$ level. **(D)** *(top panel)* The relationship between uncertainty over the self and uncertainty over the other with respect to the change in the precision (left) and median-shift (right) in phase 3 beliefs. CON participant self and other uncertainty is overlaid onto the plot to demonstrate the degree to which their beliefs *should* change in phase 3 according to the model. *(bottom panel)* Correlating the model-predicted median shift in beliefs and derived change in beliefs between phase 1 and 3 demonstrates a very strong association ($r = 0.88$, $p < 0.001$). For the purposes of visualisation we cap real and simulated values <15 for compactness, although the true correlation reported is irrelevant to this visual constraint. **(E)** *(left panel)* We overlay model-predicted (solid line) and real observed (dashed line) trial-by-trial probabilities extracted from a linear model for a correct prediction by participants. For raw trial by trial updating see **Supplementary Figure 5.** Both closely match. *(middle panel)* There was no significant difference (ns) for BPD and CON participants with respect to their total correct answers over phase 2. *(right panel)* Model-predicted and real observations in phase 2 total scores were highly correlated in both groups (CON r=0.84, p<0.001; BPD r=0.89, p<0.001).

Parameter Associations with Reported Trauma, Paranoia, and Attributed Intent

We collected psychometric data from participants prior to entering the task, and then additionally asked participants to attribute explicit intentions to their partner after phase 2. Attributions varied along two axes: the degree to which they believed their partner was motived by harmful intent (HI) and self-interest (SI).

Reported persecutory ideation (RGPTSB) and childhood trauma (CTQ) across both groups were associated with lower self-preferences for absolute reward (RGPTSB: ρ = -0.27, p = 0.007, CTQ: ρ = -0.25, p = 0.001) and higher competitive self-preferences (RGPTSB: ρ = 0.29, p = 0.003, CTQ: ρ = 0.23, p = 0.02). Both the CTQ and RGPTSB were also associated with more rigid self-preferences about absolute reward (RGPTSB: ρ = -0.30, p = 0.003, CTQ: ρ = -0.50, p < 0.001), but not relative reward.

Reported CTQ was associated with a reduction in prior belief uncertainty ($\rho[\beta_{par}^{ref}]$ = -0.26, p = 0.008) and updating ($\rho[\Delta\beta_{par}^{m}]$ = -0.37, p < 0.001) about a partner's relative preferences (**Figure 4A**). There was no association of reported childhood trauma and beliefs about a partner's absolute preferences. In contrast, baseline reported RGPTSB was associated with a reduction in prior belief uncertainty and updating across the board ($\rho[\alpha_{par}^{ref}]$ = -0.28, p = 0.004; $\rho[\beta_{par}^{ref}]$ = -0.22, p = 0.03; $\rho[\Delta\alpha_{par}^{m}]$ = -0.30, p = 0.002; $\rho[\Delta\beta_{par}^{m}]$ = -0.25, p = 0.01).

Regarding participants' mentalizing capacity assessed by the MZQ (higher scores equate to worse trait mentalizing), total scores were negatively associated with belief updating in phase 2 ($\rho[\Delta\alpha_{par}^{m}]$ = -0.35, p < 0.001; $\rho[\Delta\beta_{par}^{m}]$ = -0.34, p < 0.001), but only increased prior belief uncertainty about a partner's absolute preferences ($\rho[\alpha_{par}^{ref}]$ = -0.25, p = 0.009). The MZQ was also negatively associated with social contagion along the prosocial-competitive axis ($\rho\left[\Delta\beta_{ppt}^{m}\right]$ = -0.43, p < 0.001). The credulity (ETMCQ) of participants was negatively associated with belief updating ($\rho[\Delta\alpha_{par}^{m}]$ = -0.25, p = 0.013; $\rho[\Delta\beta_{par}^{m}]$ = -0.35, p<0.001), but was not associated with prior belief uncertainty about self-other disparity, nor with social contagion. See **Supplementary Figure 7** for full correlations of parameters with all measures.

We also show how social contagion may be restricted as a result of trauma, paranoia, and less effective trait mentalizing. By assessing all participants under the assumption of M3 (where everyone is able to be influenced by their partner) allows a test of psychometric scores on preferences changes between phase 1 and 3 across our total population. We found a negative association between CTQ scores and absolute changes in self-preferences across the board ($\rho\left[|\Delta\alpha_{ppt}^{m}|\right]$ = -0.26, p = 0.010; $\rho\left[|\Delta\beta_{ppt}^{m}|\right]$ = -0.22, p = 0.031), whereas RGPTSA and RGPTSB scores were only negatively associated with changes in self-preferences about relative reward (RGPTSA: $\rho\left[|\Delta\beta_{ppt}^{m}|\right]$ = -0.33, p < 0.001; RGPTSB: $\rho\left[|\Delta\beta_{ppt}^{m}|\right]$ = -0.31, p = 0.002). The MZQ was also associated with reduced social contagion effects for relative reward preferences ($\rho\left[|\Delta\beta_{ppt}^{m}|\right]$ = -0.43, p < 0.001). No other scale was affiliated with social contagion under M3 (see **Supplementary Figure 8**). Controlling for trait mentalizing (MZQ) nullified the relationship between RGPTS scores and social contagion, as well as CTQ scores and

social contagion but only for relative reward. Thus, childhood trauma and trait paranoia may only result in less self-change when trait mentalizing is impacted.

We then tested parameter influences on explicit intentional attributions in Phase 2. Uncertainty about the self in phase 1 was not associated with either HI or SI attributions. Greater participant-partner disparity at baseline (before interaction) was distinctly associated with HI and SI (**Figure 4B**). Greater disparity of absolute preferences before learning was associated with reduced attributions of SI ($\rho[|\alpha^m_{par} - \alpha^m_{ppt}|]$ = -0.22, p = 0.03), and greater disparity of relative preferences before learning exaggerated attributions of HI ($\rho[|\beta^m_{par} - \beta^m_{ppt}|]$ = 0.22, p = 0.03). This is likely due to partners being significantly less individualistic and prosocial on average compared to participants ($\Delta\mu[\alpha]$ = -5.50, 95%HDI: -7.60, -3.60; $\Delta\mu[\beta]$ = 12, 95%HDI: 9.70, 14.00), thus partners are correctly recognised as less selfish and more competitive.

Greater prior uncertainty (before interaction) over a partner's relative preferences was associated with increased HI ($\rho[\beta^{ref}_{par}]$ = 0.26, p = 0.007) but not SI, according with prior work (Barnby et al., 2022). There was no association between prior uncertainty over absolute preferences with either attribution. Controlling for total belief updating about a partner's relative reward preferences ($\Delta\beta^m_{par}$) and baseline similarity ($|\beta^m_{par} - \beta^m_{ppt}|$) did not remove the association with HI ($\rho[\beta^{ref}_{par}]$ = 0.21, p = 0.03). This suggests that expectations of greater difference, irrespective of one's true difference between self-other, may exaggerate beliefs about the intentional harm of others.

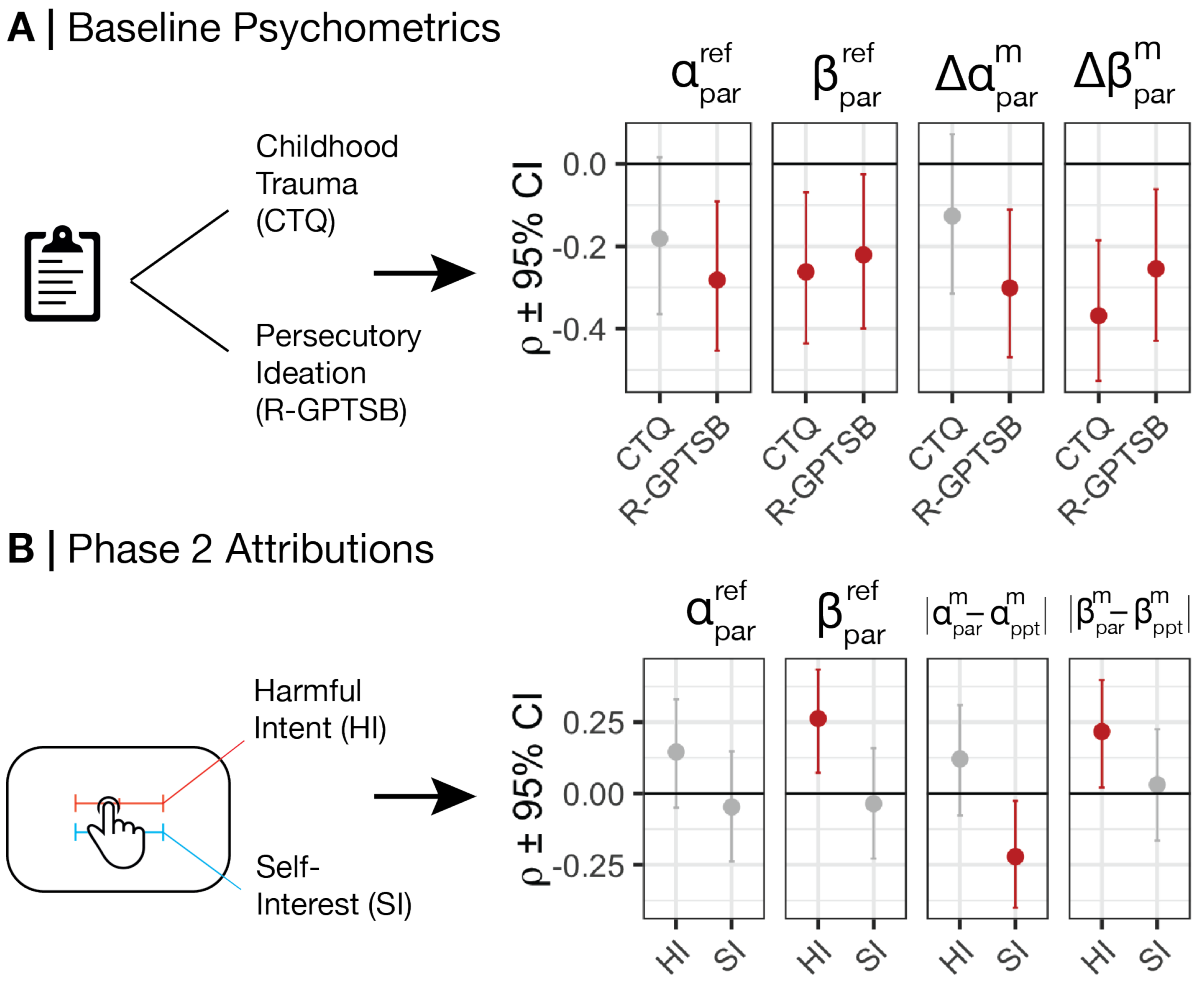


Figure 4. Psychometric correlations. (A) We conducted ranked spearman correlations between belief flexibility (α^{ref}_{par}; β^{ref}_{par}) and updating ($\Delta\alpha^{m}_{par}$; $\Delta\beta^{m}_{par}$) in phase 2 controlling for true baseline similarity with respect to server derived parameters. We found that childhood trauma was negatively associated with flexibility and updating over relative reward preferences. Persecutory ideation scores were negatively associated with belief flexibility and updating across the board. (B). We conducted ranked spearman correlations between belief flexibility (α^{ref}_{par}; β^{ref}_{par}) and absolute partner-participant dissimilarity ($|\alpha^{m}_{par} - \alpha^{m}_{ppt}|$; $|\beta^{m}_{par} - \beta^{m}_{ppt}|$) – with respect to server-derived parameters - in phase 2. Only flexibility over relative reward preferences in phase 2 was associated with harmful intent attributions. Increased absolute participant-partner dissimilarity was associated with lower self-interest attributions, and increased relative participant-partner dissimilarity was associated with high harmful intent attributions.

Discussion

We built and tested a theory of interpersonal generalization in a population of matched participants with (BPD) and without (CON) a diagnosis of borderline personality disorder using the Intentions Game, a three-phase social value orientation task. Both groups demonstrated equivalent behavioral accuracy but employed opposite strategies. CON participants used a process of self-other generalization to predict and align with their partners, while BPD participants maintained distinct representations of self and other. In phase 2, CON participants exhibited greater belief sensitivity to new information during observational learning, eventually adopting a similar updating regime to those with BPD. Our findings also indicate that reported childhood trauma and persecutory beliefs were linked to reduced flexibility when learning about partners, diminishing the influence of a partner's behavior on self-change. Collectively, our results integrate prior computational and behavioral findings in BPD and provide a formal account of social information generalization in humans, alongside a concise social paradigm to test these processes.

The data replicate models of social generalization that have focused on individual processes of self-insertion and contagion, extending these theories by demonstrating both processes in conjunction. Models of self-insertion directly map participant preferences onto prior beliefs about others, which has been used to explain increased reaction times in observational learning of others' snack food preferences (Tarantola et al., 2017), as well as improved predictive accuracy when matched with individuals of similar social values (Barnby et al., 2022). Both findings are replicated in this study. Although we did not explicitly model reaction times, we observed an interaction between reaction time reductions over time and interpersonal similarity at baseline. In tandem, models of social contagion have focused on intertemporal discounting and explain shifts in self-preferences as a function of uncertainty regarding self and others (Moutoussis et al., 2016). In both the dominant (M1) and sub-dominant (M3) models that best explained data in healthy participants, shifts in self-beliefs were also influenced by representational uncertainty of self and other: greater self-uncertainty and reduced other uncertainty led to larger shifts in self-beliefs.

The data also align with prior research on social impression formation, which suggests that humans form rapid evaluations of others that are refined over time (Bone et al., 2021; Moutoussis et al., 2023). This initial 'heating' and subsequent 'cooling' of beliefs corresponds to the computational complexity employed: model-based strategies are typically used early in interactions, transitioning to simpler, model-free computations once a partner's behavior becomes predictable (Gęsiarz & Crockett, 2015; Guennouni & Speekenbrink, 2022). Our findings support this framework, demonstrating initial variability early in interactions followed by steady, minimal updating. Notably, participants with a BPD diagnosis exhibit a less sensitive updating profile compared to CON participants.

Disruptions in self-to-other generalization provide an explanation for previous computational findings related to task-based mentalizing in BPD. BPD is characterized by early life adversity and neglect, which result in diminished representations of self and other (Fonagy & Bateman, 2008). Studies tracking observational mentalizing reveal that individuals with BPD, compared to those without, place greater emphasis on social over internal cues when learning (Henco et al., 2020; Fineberg et al., 2018)

and demonstrate reduced belief adaptation (Siegel et al., 2020), along with 'splitting' of latent social representations (Story et al., 2024). This heightened focus on others often leads to perceiving them as harmful. From the perspective of our model, those with BPD intensely focus on social information (Henco et al., 2020) due to the adoption of a new, neutral belief about others. The absence of constrained self-insertion may predispose them to 'split' beliefs (Story et al., 2024), as individuals with BPD reach the beliefs of their partner more rapidly, are less receptive to new information, and adopt greater belief precision (Siegel et al., 2020). This may represent an attempt to quickly reduce the uncertainty of a neutral prior. Essentially, individuals with BPD assume ambiguity and are quicker to settle on an explanation given limited data. Although self-insertion may intuitively seem counter to rational belief formation, it has important implications for sustaining trusting social bonds through moderation of information in the face of uncertainty.

Those with a diagnosis of BPD also show reduced permeability in generalising from other to self. While prior research has predominantly focused on how those with BPD use information to form impressions, it has not typically examined whether these impressions affect the self. In interactive trust paradigms, neural responses to monetary offers from others to the self were substantially blunted in individuals with BPD compared to those without (King-Casas et al., 2008). Similarly, in non-social reward tasks, those with BPD show reduced neural feedback-related negativity amplitudes, which obstructs feedback-related self-change (Stewart et al., 2019; Vega et al., 2013). Our results suggest a mechanistic basis for social contagion, indicating that self-rigidity prevents observed social behaviors from generalizing to the self, potentially exacerbated by childhood trauma, paranoia, and impaired mentalizing capabilities. Resistance to social influence may serve as a protective response but can also contribute to the pervasive loneliness experienced by individuals with BPD, even in the absence of social isolation (Liebke et al., 2017).

Clinical implications of our work underscores the importance of consistency and stability in clinical support for individuals with a diagnosis of BPD. Encouragingly, we found that those with BPD were not entirely impermeable to observed behavior, suggesting that consistent external models of trust could be internalized over time. Restoring a stable sense of self through social learning and effective mentalizing, along with a consistent focus on differentiating self from other (de Meulemeester et al., 2021), are central to mentalization-based therapies (Bateman & Fonagy, 2010; Smits et al., 2024) and other evidence-based treatments for BPD. We hope that our paradigm and model can offer insights into the effectiveness of these and other therapies in driving mechanistic psychological change.

More broadly, our model bridges formal theories of associative learning and social cognition. Reinforcement learning approaches have effectively organized theories around uncertainty navigation in non-social contexts (Piray et al., 2021; Zika, 2023). However, humans do not function in isolation. Bayesian models of internal and external social beliefs are better suited to capture the dynamic nature of time, context, and uncertainty during interactions (FeldmanHall & Nassar, 2021; Velez & Gweon, 2021). This is particularly important for understanding psychiatric disorders (Barnby et al., 2023). Our paradigm is concise, visually engaging, includes straightforward rules and instructions, and allows for tight experimental control over partner similarity to promote learning. Our model and paradigm effectively capture core social

psychological principles grounded in general computational approaches to learning and uncertainty, elucidating key aspects of human social interaction and exchange.

We note some limitations to our study. Primarily, we focused on the ability of individuals to integrate their self-concept into beliefs about others. It is also possible that humans possess strong, salient representations of others (or groups of others) that serve as dominant templates for learning. This may be particularly relevant for individuals with BPD, who will often have interpersonal experiences of abuse, neglect, or other forms of distress. The use of a salient, negative other-prior as a basis for learning was not measured in this study, but it may explain the ambivalent prior observed in phase 2, where a mixture of self and notional other influences belief formation, leading to rigid belief updating. Individuals with BPD may integrate priors from different sources as a mixture. We can simulate this by modelling a causal framework that incorporates priors based on both self and a strong memory impression of a notional other (**Figure S3**). However, a strength of our data is that we observed impression formation independent of valence—impressions were formed regardless of whether a partner was more or less prosocial or selfish than the participant (**Figure S4**). This supports our hypothesis that a vulnerable self-model and lack of self-insertion contribute to the formation of overly precise beliefs during learning as a means of rapidly reducing uncertainty. Even if a mixture model better explains the ambivalent prior in phase 2, it would still support a general hypothesis about the fractured concept of self and other in BPD.

Another strength of our work is demonstrating processes of self-insertion and contagion under minimal interaction conditions: simple observation alone was sufficient to elicit both processes. However, this is also a limitation. While we predict that these processes will apply in more naturalistic settings, this has yet to be tested, and it remains unclear whether these effects will persist in richer conditions, particularly when higher affective arousal and challenges to mentalising are present. Lastly, the action space and parameters governing choice in our study were quite simple—two actions influenced by two parameters. This was a deliberate computational choice to avoid overly complex action spaces that may be difficult to fit to real human data, and which might fail to capture how these mechanisms operate in the context of increasing action and model complexity.

Our findings open new possibilities for testing how social uncertainty across the lifespan, and in the context of ill-health, may explain the formation and maintenance of healthy social bonds as well as their disruption. We make two key predictions: 1. The self is an evolving and dynamic concept, particularly susceptible to peer influence during adolescence. We predict that adolescents will use self-insertion to a lesser degree (if at all) than adults in our sample, and that the extent of social contagion in our paradigm will correlate with reported peer influence in other areas. We further predict that the degree of social contagion will correspond to brain regions associated with self-processing (Sebastian, Burnett & Blakemore, 2008), and that these contagion effects will diminish as individuals progress into emerging and full adulthood. 2. Psychosis is conceptualized as a heightened absorption in self-generated conscious experiences, leading to a collapse of the internal and external boundaries of self and other (Humpston, 2018). We predict that in our paradigm, this will manifest as an exaggeration of self-insertion when predicting others, accompanied by low learning rates and minimal social contagion effects.

Materials and Methods

Participants

We used a case-control, between-subjects design with 103 participants: a control group from the general population (N = 53) and a clinical group diagnosed with BPD (N = 50). Both groups were recruited for a larger study investigating social exchanges in BPD and Anti-Social Personality Disorder (approved by the Research Ethics Committee for Wales, 12/WA/0283). The control and clinical groups were matched on age, sex, years in education, and the English Indices of Deprivation based on the 2019 census (IoD2019; Ministry of Housing, Communities & Local Government, 2019). Participants received £70 compensation for completing questionnaires and online tasks which included the Intentions Game. They also received a performance bonus if they were entered into the lottery for surpassing 1000 points over the course of the game.

Participants for the control group were recruited through an advertisement on the Call For Participants website (https://www.callforparticipants.com), local community services and adult schools. Inclusion criteria required control participants to have no pre-existing or current diagnoses of mental health disorders, neurological disorders, or traumatic brain injuries. Additionally, control participants must not have been currently in therapy or taking medication for any psychiatric disorders.

BPD participants were recruited through referrals by psychiatrists, psychotherapists, and trainee clinical psychologists within personality disorder services across 9 NHS Foundation Trusts in the London, and 3 NHS Foundation Trusts across England (Devon, Merseyside, Cambridgeshire). Participants were also recruited through the UCLH website, where the study was advertised. Individuals who discovered the study through this platform and were interested in participating initiated contact themselves. To be included in the study, all participants needed to have, or meet criteria for, a primary diagnosis of BPD (or emotionally-unstable personality disorder or complex emotional needs) based on a clinical assessment and be under the care of one of the trusts collaborating in recruitment or have a general practitioner whose details they were willing to provide. Clinical participants with recent psychotic episodes, severe learning disability, or current or past neurological disorders were excluded.

Psychometric Measures

Green et al. Paranoid Thought Scale (GPTS). The GPTS assesses paranoid thoughts, including ideas of social reference (scale A) and persecution (scale B), in both general and clinical populations (Green et al., 2008). Each item is scored from 0 (not at all) to 5 (totally) concerning endorsement of each item. We retained items from the GPTS that were consistent with the revised version outlined in Freeman et al., 2021 (Revised GPTS; R-GPTS). The R-GPTS has demonstrated excellent psychometric properties (Freeman et al., 2021), making it a reliable and valid tool for assessing trait paranoid thoughts in non-clinical and clinical populations.

Childhood Trauma Questionnaire (CTQ). The Childhood Trauma Questionnaire is used to screen for maltreatment history (Bernstein et al., 2003). Each item is scored from 1 (never true) to 5 (very often true). The CTQ has showed good internal consistency reliability across the five scales (Sacchi et al., 2018) and good construct validity based on significant associations with stress responsivity (McMahon et al., 2022), and dissociation (Nobakht et al., 2021).

Certainty About Mental States Questionnaire (CAMSQ). The CAMSQ assesses one's certainty in classifying the mental states of oneself and others at an abstract level (Müller et al., 2023), e.g. 'I know what other people think of me' and 'I know my feelings'. Each subscale is scored from 1 (never) to 7 (always). In US and German samples, the CAMSQ showed high internal consistency for Self-Certainty (ω = .90/.88) and Other-Certainty (ω = .91/.89) subscales, and high two-week test-retest reliability for Self-Certainty (r = .85), Other-Certainty (r = .78), and Other-Self-Discrepancy (r = .82) scores (Müller et al., 2023).

Mentalisation Questionnaire (MZQ). The MZQ is a 15-item questionnaire assessing an individual's trait mentalizing, i.e., one's ability to understand and interpret their own and others' mental states (Hausberg et al., 2012). The MZQ demonstrated good internal consistency (α = .81) and test-retest reliability (r = .76), and was sensitive to change over a 6-month follow-up period and showed good criterion-related validity, distinguishing individuals with BPD from those without BPD (Hausberg et al., 2012). A higher score reflects worse trait mentalizing.

Epistemic Trust, Mistrust and Credulity Questionnaire. The ETMCQ is a 15-item measure calibrated to assess trust (e.g. 'I usually ask people for advice when I have a personal problem), mistrust (e.g. 'I'd prefer to find things out for myself on the internet rather than asking people for information), and credulity (e.g. 'I am often considered naïve because I believe almost anything that people tell me'; Campbell et al., 2021). Each item is scored from 1(Strongly Disagree) to 7(Strongly Agree).

Paradigm, procedure and server architecture

The Intentions Game is a repeated social-value orientation paradigm with three phases.

In Phase 1 of the Intentions Game, participants take on the role of the decider with an anonymous partner over 36 trials. In each trial, participants choose between two options to distribute points between themselves and their partners. Participants make 12 choices each between prosocial and competitive (e.g. Option 1=[10,10], Option 2 = [10,5]) individualistic and competitive (e.g. Option 1=[10,5], Option 2=[8,1]), and prosocial and individualistic options (e.g. Option 1=[5,5], Option 2=[10,5]). Phase 1 choices allowed experimenters to classify participants' social preferences as prosocial (preferring equal outcomes), individualistic (maximising own payoff), or competitive (maximising relative payoff difference at the cost of lower self-gain).

We included a task environment that balanced each type of choice pair (see **Supplementary Table 1**).

In phase 2 of the game, participants were matched with a new anonymous partner and played the role of the recipient over 54 trials. In this phase, the participants predicted which of the two options their partner would choose on each trial. Trial numerical values for self and other were identical to Phase 1. Partners' decisions were determined via a dynamic algorithm (Burgess et al., 2023) to ensure partners were approximately ~50% different from the participants' based on participants' choices in phase 1. To surmise this architecture, we implemented a version of the client-server paradigm hosted on an Amazon Web Service (AWS) LightSail server, where the web-based behavioural task (implemented with JavaScript in Gorilla.sc) acted as the client and exchanged information with a remote AWS server. The server received all anonymised behavioural data following phase 1. The Application Programming Interface (API) to interact with the server used a customizable R script (v4.3) to process the received data from the participant, and additional R scripts were used to process and generate output for the participant. A function within the backend scripts first used Bayesian inference to approximate a participant's parameters for phase 1. It then simulated what choices the participant would have made in phase 2 had the participant been in the role of the partner. The algorithm then sought to find parameters that would be at least 50% dissimilar from participant parameters with respect to the generated choices of those parameters. This allowed the task behaviour of phase 2 to be dynamically updated in response to participant choices in phase 1. This facilitated tight control over the state of the task and enabled advanced computations to be performed on participant data beyond the capabilities of a web browser.

Participants were incentivised in phase 2 to predict accurately, as accurate predictions would contribute to their total point scores (total correct answers were multiplied by 10 and added to their points) and determined their entry into the lottery to win an extra £20 Amazon voucher. After participants had made their predictions, they were given feedback informed on whether their predictions were accurate.

At the end of phase 2, participants were asked to rate (1) the extent to which they thought their partner was driven by the desire to earn points in this task overall (self-interest) and (2) the extent to which they thought their partner was driven by the desire to reduce the participant's points in this task overall (attribution of harmful intent). The answers were presented using two separate sliders from 0 to 100; the sliders were initialised to be invisible until the participants made the first click.

Phase 3 was identical to phase 1 except that participants were matched with a new anonymous partner. Participants would take on the decider role similar to phase 1 which allowed experimenters to estimate whether the observation of their partner in phase 2 had an influence on participants in phase 3.

Behavioural Analysis

All analysis was conducted in R (v. 4.3.3) on a macbook pro (M2 Max; OS=Ventura13.5). All individual numeric values extraneous of statistical tests are reported with their mean and standard deviation (mean=XX, SD=YY). All statistical tests where dependent variables mapped one value to one participant (e.g. trait psychometric scores) were conducted as linear models, with the regression coefficient, 95% confidence interval (95%CI), t-value and p-value reported like so

(linear estimate=XX, 95%CI:AA,BB; t=CC, p=DD). When dependent variables mapped multiple values to each participant (e.g. trial-by-trial accuracy or reaction time) random-effects linear modelling was used. All correlations used Pearson estimates (*r*) unless distributions were non-normal, in which case Spearman-ranked correlations (ρ) were performed.

Model space and Computational Analysis

We apply four computational hypotheses (M1-M4) which could explain the data collected from the Intentions Game (**Figure 1**), centred around formal principles of self-insertion and social contagion. Self-insertion states that a self inserts their own preferences into their beliefs about others (Anderson & Chen, 2002; Kreuger & Clement, 1994); Social Contagion states that a self's preferences will change when exposed to the preferences of an other (Frith & Frith, 2012). In each case, cognitive representations of self and other are allowed to intermingle to form a new hybrid of the two for the purposes of computational efficiency and/or social bonding.

We note some important assumptions in our notation going forward. In dyadic social interaction, both parties are trying to estimate and predict the true state (θ) of the self (θ_s) and the other (θ_o). However, this estimation is inherently imperfect. Theories of social inference need to consider three sources of noisy estimation of this quantity: a self's (s) metacognitive model of their own state, $\bar{\theta_s}$, their partner's (o) state, $\bar{\theta_o}$, and finally the experimenter's approximation of both quantities, $\hat{\theta}_{s,o}$ (Barnby et al., 2024). In this work we consider the experimenter's approximation of the self's state $\hat{\theta}_s$ (phase1), the self's approximation of their other $\bar{\theta}_o$ (phase 2), and how exposure to a partner may influence $\hat{\theta}_s$ (phase 3). We term the self the participant (ppt) and the other the partner (par) and assume $\theta_{ppt} \equiv \hat{\theta}_s$ in phase 1, $\theta_{par} \equiv \bar{\theta}_o$ in phase 2, and $\acute{\theta}_{ppt}$ are the shifted participant preferences following exposure to the partner.

All models assumed a constricted Fehr-Schmidt utility function was used by participants and partners to calculate the utility of two options ($\mathbf{U}_{\alpha,\beta} = \{\mathrm{U}^1_{\alpha,\beta}, \mathrm{U}^2_{\alpha,\beta}\}$) in each trial within the task.

In phase 1, participants made binary choices $c^t, t = \{1 \ldots T\}$ about whether option 1 or option 2 should be chosen given the returns for each option pair, $\mathbf{R^t} = \{\mathbf{R^{t;1}}; \mathbf{R^{t;2}}\} = \{R^{t;1}_{ppt}, R^{t;1}_{par}, R^{t;2}_{ppt}, R^{t;2}_{par}\}$.

(1)

$$\mathbf{U}_{\alpha,\beta}(\mathbf{R^{t;1}}) = \alpha_{ppt} * \mathrm{R}^{t;1}_{ppt} + \beta_{ppt} * \max(\mathrm{R}^{t;1}_{ppt} - \mathrm{R}^{t;1}_{par}, 0)$$

$$\Delta\mathrm{U}_{\alpha,\beta}(\mathbf{R^t}) = \mathrm{U}_{\alpha,\beta}(\mathbf{R^{t;1}}) - \mathrm{U}_{\alpha,\beta}(\mathbf{R^{t;2}})$$

Here, α_{ppt} describes the weight a participant places on their own payoff (in one reduced model we set α_{ppt} = 1), and β_{ppt}, the weight a participant places on their payoff relative to the payoff of their partner. Large positive or negative values of β_{ppt} 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 efficiency we discretised states of α_{ppt} from 0-30 (increments of 0.125) and β_{ppt} from -30 to 30 (increments of 0.25).

Over this state space we can construct a belief that participants are estimated to hold which generate their choices, $\boldsymbol{C}$. Herein, we refer to this belief as θ_{ppt}, where θ_{ppt} is a matrix over a fixed grid of α_{ppt} and β_{ppt} values. In the models, θ_{ppt} is drawn from a normal distribution made from a central tendency, θ^{m}_{ppt}, and a standard deviation, θ^{σ}_{ppt}. The standard deviation around the central tendency allows for stochastic choice behaviour consistent with random utility models (Block, 1974; McFadden, 1974). We invert the model to estimate θ_{ppt} based on a participant's choices given their likelihood of choosing $c^t = 1$:

(2)

$$p(\theta_{ppt}|\boldsymbol{C}) \sim N(\theta_{ppt};\ \theta^{m}_{ppt}, \theta^{\sigma}_{ppt})$$

$$p\left(c^t = 1 \middle| \theta_{ppt}, \mathbf{R^t}\right) = \sum_{\theta_{ppt}} \sigma(\Delta \mathrm{U}_{\alpha,\beta}(\mathbf{R^t})) \cdot \theta_{ppt}$$

$$LL = \log\left[p\left(c^t = 1 \middle| \theta_{ppt}, \mathbf{R^t}\right)\right]$$

When θ^{σ}_{ppt} is larger, a participant's choices in Phase 1 are estimated to be less deterministic and more stochastic – i.e. they are less sure about their preferences along each dimension. This consideration will become important for choices made in phase 3.

In phase 2, over 54 trials, we then model the participants binary predictions $\bar{d}^t, t = \{1 \ldots T\}$ about whether option 1 or 2 would be chosen by their partner given the returns $\mathbf{R^t} = \{\mathbf{R^{t;1}}; \mathbf{R^{t;2}}\} = \left\{R^{t;1}_{ppt}, R^{t;1}_{par}, R^{t;2}_{ppt}, R^{t;2}_{par}\right\}$ for each pair of options. They then were given feedback about the partner's true decision which we note as d^t. We assumed the participant predict the partner in the same way they would themselves, ranging along two dimensions, α_{par} and β_{par} which was needed to be inferred through observation, using a likelihood for d^t of $LL = \log\left[p\left(d^t = 1 \middle| \alpha_{par}, \beta_{par}, \mathbf{R^t}\right)\right]$ using the same formula as phase 1. We note the belief about α_{par} and β_{par} together as θ_{par}, represented as a matrix over a fixed grid of α_{par} and β_{par} values.

The partner decisions, $D^t = \{d^1, d^2 \ldots, d^T\}$ are then used to update the participants beliefs about the partner, written as $p(\theta_{par}|D^t)$, starting with prior $p(\theta_{par}|D^0)$. Both M1 and M2 assume participants use their own central tendency, θ^{m}_{ppt}, as a starting point for their prior beliefs about their partner as theoretically outlined as a self-insertion bias (Barnby et al., 2024) which draws from past computational work (Barnby et al., 2022; Tarantola et al., 2017). We also assumed participants used a new standard deviation θ^{ref}_{par} which allowed for participants to believe their partner may be different from them (belief flexibility). Therefore we have:

(3)

$$p(\theta_{par}|D^0) \sim N(\theta_{par};\ \theta^m_{ppt}, \theta^{ref}_{par})$$

In models M3 and M4, we assume participant's may instead use a new central tendency (rather than their own) as prior beliefs over their partner. This are free parameters to be approximated, $\bar{\alpha}^m_{par}, \bar{\beta}^m_{par}$.

In all cases, we assume participants update their beliefs about their partner's social preferences given their partner's decisions ***D*** along trials 1-54 according to Bayes rule:

(4)

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

We can then marginalise over θ^t_{par} to calculate the belief participants had over their participant's social value preferences.

We assume that participants predict the partner's decision in the next trial by calculating the probability determined by the utility differences $\Delta U_{\alpha,\beta}(\mathbf{R^{t+1}})$ as in phase 1, summed over the joint distribution of partner parameters, θ^t_{par}:

(5)

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

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

And then performed probability matching, so that:

(6)

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

In the third phase participants are once again asked to make choices for themselves and a new anonymous partner over 36 trials with an assumed identical utility function as in phase 1. In model M1 and M3 we assume participants use a combination of their own preferences and the posterior beliefs about their partner to form a new distribution to select between the two options available on each trial. This draws from the same formulation used previously (Moutoussis et al. 2016). In essence, we state that participants know their true preferences in phase 1 but are unsure about them. The inferred partner beliefs θ^t_{par} provides information to the participant about some common preference distribution both share, which in turn informs the participant's own choices $\acute{c}^t, t = \{1 \ldots T\}$ in the form of an adjusted belief along each dimension for phase

3, $\acute{\alpha}_{ppt}$ and $\acute{\beta}_{ppt}$ (eq. 7), using a log likelihood of $LL = \log\left[p\left(\acute{c}^t = 1 \middle| \acute{\alpha}_{ppt}, \acute{\beta}_{ppt}, \mathbf{R^t}\right)\right]$. We refer to $\acute{\alpha}_{ppt}$ and $\acute{\beta}_{ppt}$ together as $\acute{\theta}_{ppt}$ for convenience, where $\acute{\theta}_{ppt}$ is a matrix over a grid of fixed values of $\acute{\alpha}_{ppt}$ and $\acute{\beta}_{ppt}$. To note: models M2 and M4 do not assume participants undergo this change, and instead use their original phase 1 beliefs to make choices $LL = \log\left[p\left(\acute{c}^t = 1 \middle| \alpha_{ppt}, \beta_{ppt}, \mathbf{R^t}\right)\right]$.

(7)

$$p(\acute{\theta}_{ppt}|\boldsymbol{C}) \sim N(\theta_{ppt};\ \acute{\theta}^m_{ppt}, \acute{\theta}^\sigma_{ppt})$$

$$\acute{\theta}^\sigma_{ppt} = \ (\theta^{\sigma^{-2}}_{ppt} + (2\theta^{ref^2}_{par} + \theta^{\sigma^2}_{par})^{-1})^{-1}$$

$$\acute{\theta}^m_{ppt} = \ \acute{\theta}^{\sigma^2}_{ppt}[(\theta^{\sigma^{-2}}_{ppt} * \ \theta^m_{ppt}) + \left(2\theta^{\sigma^2}_{ref} + \theta^{\sigma^2}_{par}\right)^{-1} * \theta^m_{par}]$$

Where θ^σ_{par} and θ^m_{par} are the standard deviation and central tendency of the final posterior inference about the partner, $\theta^{t;54}_{par}$.

All computational models were fitted using a Hierarchical Bayesian Inference (HBI) algorithm which allows hierarchical parameter estimation while assuming random effects for group and individual model responsibility (Piray et al., 2019). During fitting we added a small noise floor to distributions ($2.22e^{-16}$) before normalisation for numerical stability. Parameters were estimated using the HBI in native space drawing from broad priors ($\mu_{\mathbf{M}}=0$, $\sigma^2_{\mathbf{M}} = 6.5$; where **M**={M1, M2, M3, M4}). This process was run independently for each group. Parameters were transformed into model-relevant space for analysis. All models and hierarchical fitting was implemented in Matlab (Version R2022B). All other analyses were conducted in R (version 4.3.3; arm64 build) running on Mac OS (Ventura 13.0).

To conduct model recovery we simulated synthetic participants (CON=53; BPD=50) using their fitted parameters from the dominant model of the group (CON=M1; BPD=M4). We then performed model fitting with an identical procedure to the real behavioural data. We tested associations between model responsibility and individual parameters for the real and recovered models, as well as the association between choices and predictions made by the model from simulation and the choices and predictions made by participants in each trial.

Differences between groups for individual-level parameters were estimated using hierarchical Bayesian t-tests (Bååth, 2014). This used JAGS as a backend MCMC sampler; differences in mean between groups ($\Delta\mu$) are additionally reported with their corresponding posterior 95% High Density Interval (95%HDI). Belief updates were calculated as the Kullback-Leibler Divergence between probabilities (P) from trial *t-1* to *t*, marginalised along all possible states, ***S***={s^1,s^2,…,s^n}: $D_{KL}(P^t||P^{t-1}) = \sum_s^{\boldsymbol{S}} P^t(s) \log \frac{P^t(s)}{P^{t-1}(s)}$.

CRediT

JMB: Conceptualisation, Data Curation, Investigation, Formal Analysis, Methodology, Project Administration, Software, Supervision, Visualisation, Writing – Original Draft, Writing – Review and Editing. **JN:** Investigation, Methodology, Writing – Original Draft, Writing – Review and Editing. **JG:** Conceptualisation, Investigation, Project Administration, Resources, Writing – Review and Editing. **MW:** Project Administration. **HB:** Software, Writing – Review and Editing. **LR:** Resources, Writing – Review and Editing. **GC:** Validation, Writing - Review and Editing. **JK:** Supervision, Writing – Review and Editing. **PRM:** Resources, Writing – Review and Editing. **PD:** Conceptualisation, Formal Analysis, Writing - Review and Editing. **TN:** Conceptualisation, Project Administration, Resources, Supervision, Writing – Review and Editing. **PF:** Conceptualisation, Resources, Supervision, Writing – Review and Editing.

Funding

JMB is supported by a Wellcome Trust award (228268/Z/23/Z) and as a scholar within the FENS-Kavli Network of Excellence. Funding for PD was from the Max Planck Society and the Humboldt Foundation. PD is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764 and of the Else Kroner Medical Scientist College "ClinbrAIn: Artificial Intelligence for Clinical Brain Research”.

Conflict of Interest

None to declare.

Acknowledgements

We would like to greatly thank all participants who took part in the research.

References

Afifi, T. O., Mather, A., Boman, J., Fleisher, W., Enns, M. W., MacMillan, H., & Sareen, J. (2011). Childhood adversity and personality disorders: Results from a nationally representative population-based study. *Journal of psychiatric research*, *45*(6), 814-822.

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

Bååth, R. (2014). Bayesian first aid: A package that implements Bayesian alternatives to the classical*. test functions in R. *Proceedings of useR*, *2014*, 2.

Barnby, J. M., Raihani, N., & Dayan, P. (2022). Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*, *225*, 105098.

Barnby, J. M., Dayan, P., & Bell, V. (2023). Formalising social representation to explain psychiatric symptoms. *Trends in cognitive sciences*, *27*(3), 317-332.

Barnby, J. M., Alon, N., Bellucci, G., Schilbach, L., Frith, C. Bell, V. (2024; preprint). A Standard Framework for Social Cognition: Interoperable algorithms for inference and representation. *PsyArXiv.* https://doi.org/10.31234/osf.io/cmgu7

Bateman, A., & Fonagy, P. (2010). Mentalization based treatment for borderline personality disorder. *World psychiatry*, *9*(1), 11.

Bateman, A., Rüfenacht, E., Perroud, N., Debbané, M., Nolte, T., Shaverin, L., & Fonagy, P. (2023). Childhood maltreatment, dissociation and borderline personality disorder: Preliminary data on the mediational role of mentalizing in complex post-traumatic stress disorder. *Psychology and Psychotherapy: Theory, Research and Practice*.

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., ... & Zule, W. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child abuse & neglect*, *27*(2), 169-190.

Block, H. D. (1974). Random orderings and stochastic theories of responses (1960). In *Economic Information, Decision, and Prediction: Selected Essays: Volume I Part I Economics of Decision* (pp. 172-217). Dordrecht: Springer Netherlands.

Bone, J., Pike, A. C., Lewis, G., Lewis, G., Blakemore, S. J., & Roiser, J. Computational mechanisms underlying social evaluation learning and associations with depressive symptoms during adolescence.

Burgess, H., Barnby, J., Dayan, P., & Richards, L. (2023, October). Realizing Dynamic Cognitive Tasks with Cloud-based Computation. In *1st Annual Conference of the US Research Software Engineer Association (US-RSE 2023)*.

Campbell, C., Tanzer, M., Saunders, R., Booker, T., Allison, E., Li, E., ... & Fonagy, P. (2021). Development and validation of a self-report measure of epistemic trust. *PloS one*, *16*(4), e0250264.

Ciaunica, A., Constant, A., Preissl, H., & Fotopoulou, K. (2021). The first prior: from co-embodiment to co-homeostasis in early life. *Consciousness and cognition*, *91*, 103117.

Crawford, T. N., Cohen, P. R., Chen, H., Anglin, D. M., & Ehrensaft, M. (2009). Early maternal separation and the trajectory of borderline personality disorder symptoms. *Development and psychopathology*, *21*(3), 1013-1030.

Emerson, A. E. (1956). Regenerate behavior and social homeostasis of termites. *Ecology*, *37*(2), 248-258.

Euler, S., Nolte, T., Constantinou, M., Griem, J., Montague, P. R., Fonagy, P., & Personality and Mood Disorders Research Network. (2021). Interpersonal problems in borderline personality disorder: associations with mentalizing, emotion regulation, and impulsiveness. *Journal of Personality Disorders*, *35*(2), 177-193.

Fairbairn, W. R. D. (1994). *Psychoanalytic studies of the personality*. Psychology Press.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817-868.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045-1057.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B., Davies, M., Borus, J., ... & Rounsaville, B. (1995). The structured clinical interview for DSM-III-R personality disorders (SCID-II). Part II: Multi-site test-retest reliability study. *Journal of personality disorders*, *9*(2), 92-104.

Fonagy, P., & Bateman, A. (2008). The development of borderline personality disorder—A mentalizing model. *Journal of personality disorders*, *22*(1), 4-21.

Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Development and psychopathology*, *21*(4), 1355-1381.

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.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, *63*(1), 287-313.

Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E., & Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, *85*(2), 418-428.

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in behavioral neuroscience*, *9*, 135.

Green, C. E. L., Freeman, D., Kuipers, E., Bebbington, P., Fowler, D., Dunn, G., & Garety, P. A. (2008). Measuring ideas of persecution and social reference: the Green et al. Paranoid Thought Scales (GPTS). *Psychological medicine*, *38*(1), 101-111.

Guennouni, I., & Speekenbrink, M. (2022). Transfer of learned opponent models in zero sum games. *Computational Brain & Behavior*, *5*(3), 326-342.

Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nature reviews disease primers*, *4*(1), 1-20.

Hanegraaf, L., van Baal, S., Hohwy, J., & Verdejo-Garcia, A. (2021). A systematic review and meta-analysis of 'Systems for Social Processes' in borderline personality and substance use disorders. *Neuroscience & Biobehavioral Reviews*, *127*, 572-592.

Hausberg, M. C., Schulz, H., Piegler, T., Happach, C. G., Klöpper, M., Brütt, A. L., ... & Andreas, S. (2012). Is a self-rated instrument appropriate to assess mentalization in patients with mental disorders? Development and first validation of the Mentalization Questionnaire (MZQ). *Psychotherapy Research*, *22*(6), 699-709.

Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M. L., Hörmann, S., Hennings, J., ... & Mathys, C. (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS computational biology*, *16*(9), e1008162.

Hopkins, A. K., Dolan, R., Button, K. S., & Moutoussis, M. (2021). A reduced self-positive belief underpins greater sensitivity to negative evaluation in socially anxious individuals. *Computational Psychiatry*, *5*(1), 21.

Humpston, C. S. (2018). The paradoxical self: Awareness, solipsism and first-rank symptoms in schizophrenia. *Philosophical Psychology*, *31*(2), 210-231.

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.

Huprich, S. K., Paggeot, A. V., & Samuel, D. B. (2015). Comparing the personality disorder interview for DSM–IV (PDI–IV) and SCID–II borderline personality disorder scales: An item–response theory analysis. *Journal of Personality Assessment*, *97*(1), 13-21.

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.

Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., ... & Corlett, P. R. (2018). Differential valuation and learning from social and nonsocial cues in borderline personality disorder. *Biological psychiatry*, *84*(11), 838-845.

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.

Liebke, L., Bungert, M., Thome, J., Hauschild, S., Gescher, D. M., Schmahl, C., ... & Lis, S. (2017). Loneliness, social networks, and social functioning in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *8*(4), 349.

Lobbestael, J., Leurgans, M., & Arntz, A. (2011). Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I disorders (SCID I) and Axis II disorders (SCID II). *Clinical psychology & psychotherapy*, *18*(1), 75-79.

Maffei, C., Fossati, A., Agostoni, I., Barraco, A., Bagnato, M., Deborah, D., ... & Petrachi, M. (1997). Interrater reliability and internal consistency of the structured clinical interview for

DSM-IV axis II personality disorders (SCID-II), version 2.0. *Journal of personality disorders*, *11*(3), 279-284.

Mancinelli, F., Nolte, T., Griem, J., Lohrenz, T., Feigenbaum, J., King-Casas, B., ... & Mathys, C. (2024). Attachment and borderline personality disorder as the dance unfolds: A quantitative analysis of a novel paradigm. *Journal of Psychiatric Research*, *175*, 470-478.

McFadden, D (1974). "Conditional Logit Analysis of Qualitative Choice Behavior". In Zarembka, Paul (ed.). *Frontiers in Econometrics*. Academic Press. pp. 105–142.

McMahon, G., Griffin, S. M., Borinca, I., Bradshaw, D., Ryan, M., & Muldoon, O. T. (2022). Social integration: Implications for the association between childhood trauma and stress responsivity. *Psychological trauma: theory, research, practice, and policy*.

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.

Moutoussis, M., Barnby, J., Durand, A., Croal, M., Rutledge, R. B., & Mason, L. (2023). The role of serotonin and of perceived social differences in infering the motivation of others. *bioRxiv*, 2023-05.

Müller, S., Wendt, L. P., & Zimmermann, J. (2023). Development and validation of the Certainty About Mental States Questionnaire (CAMSQ): A self-report measure of mentalizing oneself and others. *Assessment*, *30*(3), 651-674.

Nobakht, H. N., Ojagh, F. S., & Dale, K. Y. (2021). Validity, Reliability and Internal Consistency of Persian Versions of the Childhood Trauma Questionnaire, the Traumatic Exposure Severity Scale and the Peritraumatic Dissociative Experiences Questionnaire. *Journal of Trauma & Dissociation*, *22*(3), 332-348.

Nolte, T., Hutsebaut, J., Sharp, C., Campbell, C., Fonagy, P., & Bateman, A. (2023). The role of epistemic trust in mentalization-based treatment of borderline psychopathology. *Journal of Personality Disorders*, *37*(5), 633-659.

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.

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature communications*, *12*(1), 6587.

Pratt, M., Apter-Levi, Y., Vakart, A., Kanat-Maymon, Y., Zagoory-Sharon, O., & Feldman, R. (2017). Mother-child adrenocortical synchrony; Moderation by dyadic relational behavior. *Hormones and behavior*, *89*, 167-175.

Sacchi, C., Vieno, A., & Simonelli, A. (2018). Italian validation of the Childhood Trauma Questionnaire—Short Form on a college group. *Psychological Trauma: Theory, Research, Practice, and Policy*, *10*(5), 563.

Sebastian, C., Burnett, S., & Blakemore, S. J. (2008). Development of the self-concept during adolescence. *Trends in cognitive sciences*, *12*(11), 441-446.

Siegel, J. Z., Curwell-Parry, O., Pearce, S., Saunders, K. E., & Crockett, M. J. (2020). A computational phenotype of disrupted moral inference in borderline personality disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(12), 1134-1141.

Stewart, J. G., Singleton, P., Benau, E. M., Foti, D., Allchurch, H., Kaplan, C. S., ... & Auerbach, R. P. (2019). Neurophysiological activity following rewards and losses among female adolescents and young adults with borderline personality disorder. *Journal of abnormal psychology*, *128*(6), 610.

Story, G. W., Smith, R., Moutoussis, M., Berwian, I. M., Nolte, T., Bilek, E., ... & Dolan, R. J. (2023). A social inference model of idealization and devaluation. *Psychological Review*.

Tarantola, T., Kumaran, D., Dayan, P., & De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature Communications*, *8*(1), 817.

Thomas, L., Lockwood, P. L., Garvert, M. M., & Balsters, J. H. (2022). Contagion of temporal discounting value preferences in neurotypical and autistic adults. *Journal of autism and developmental disorders*, 1-14.

Vega, D., Soto, À., Amengual, J. L., Ribas, J., Torrubia, R., Rodríguez-Fornells, A., & Marco-Pallarés, J. (2013). Negative reward expectations in Borderline Personality Disorder patients: Neurophysiological evidence. *Biological Psychology*, *94*(2), 388-396.

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.

Wheeler, W. M. (1911). The North American ants of the genus Camponotus MAYR. *Annals of the New York Academy of Sciences*, *20*(1), 295-354.

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2006). *Schema therapy: A practitioner's guide*. guilford press.

Zika, O. (2023). The relationship between latent state inference and (intolerance of) uncertainty. *Neuroscience and Biobehavioral Reviews*, *152*, 105321.

Supplementary Materials

Self-Other Generalisation Shapes Social Interaction and Is Disrupted in Borderline Personality Disorder

Barnby, J.M.* [1,2,3], Nguyen, J.[1], Griem, J.[4,5], Wloszek, M. [5], Burgess, H.[6], Richards, L.[6], Kingston, J.[1], Cooper, G.[1], London Personality and Mood Disorders Consortium, Montague, P. R.[7], Dayan, P.[8,9], Nolte, T. [4,5], Fonagy, P. [4,5]

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

[2]Social and Cultural Neuroscience Group, King's College London, London, UK

[3]School of Psychiatry and Clinical Neuroscience, University of Western Australia, AU

[4]Department for Clinical, Educational, and Healthy Psychology, Division of Psychology and Language Sciences, University College London, London, UK

[5]Anna Freud, London, UK

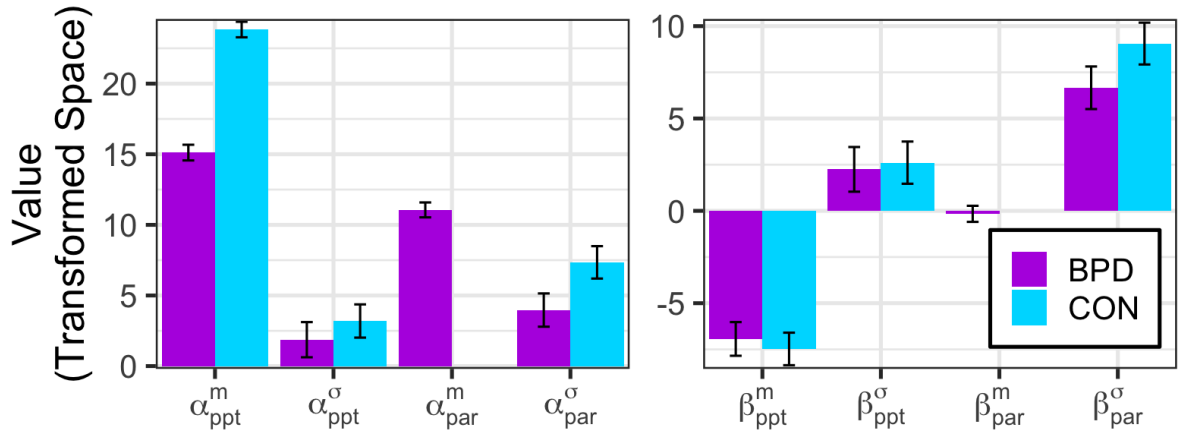
[6]Washington University in St. Louis, MO, USA

[7]Centre for Human Neuroscience Research, Virginia Tech, VA, USA

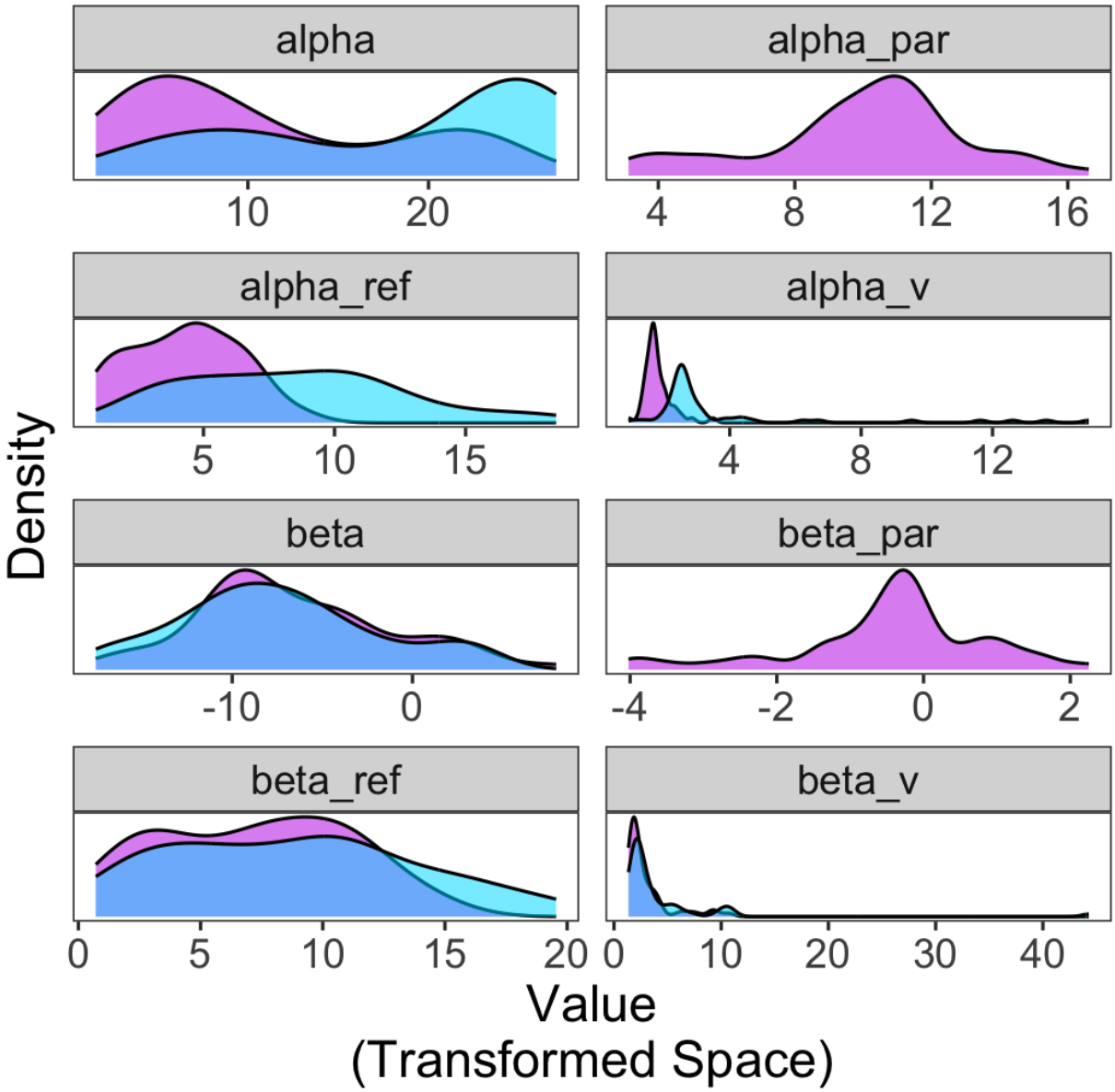
[8]Max Planck Institute of Biological Cybernetics, Tübingen, DE,

[9]University of Tübingen, Tübingen, DE

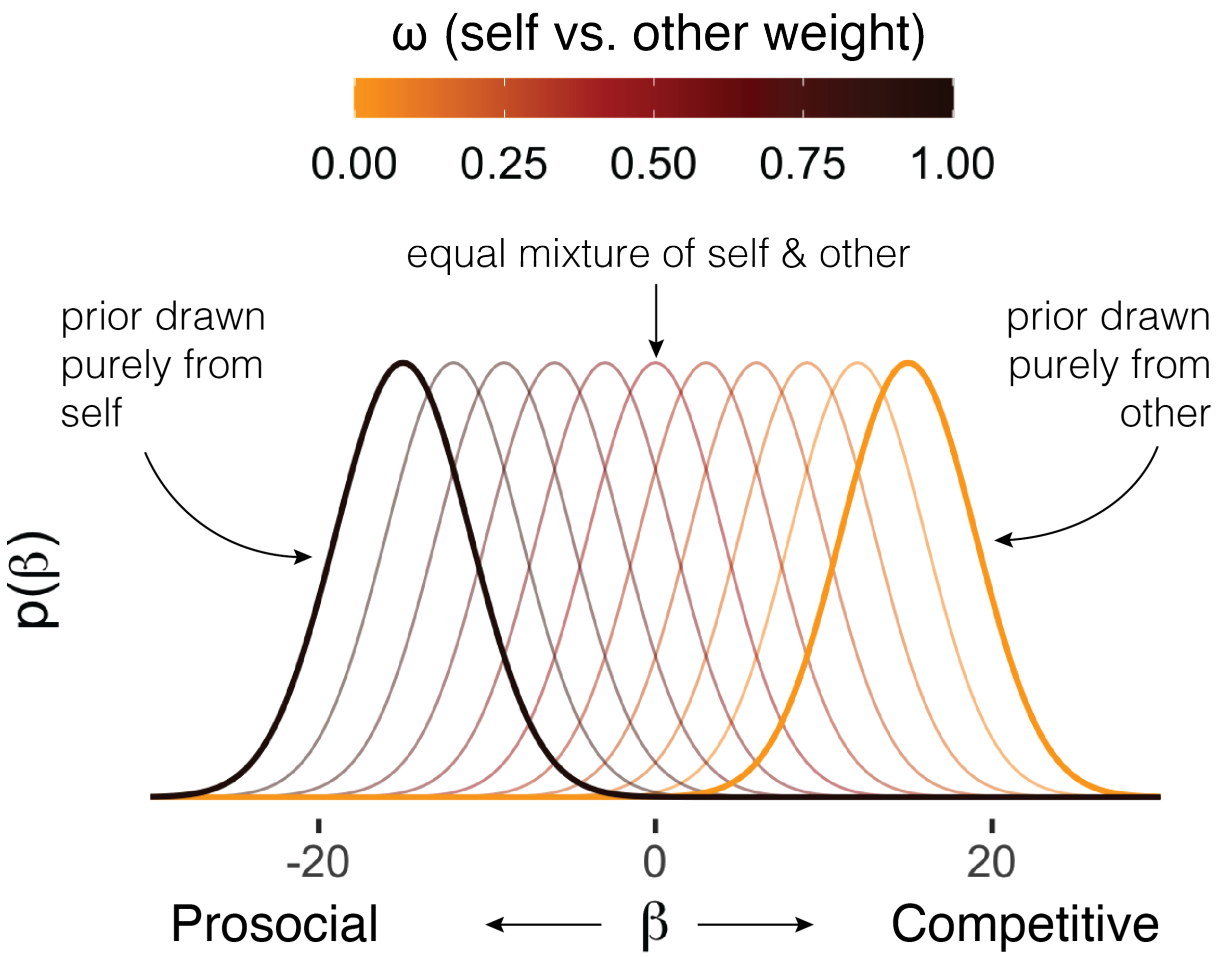
*corresponding: joseph.barnby@rhul.ac.uk



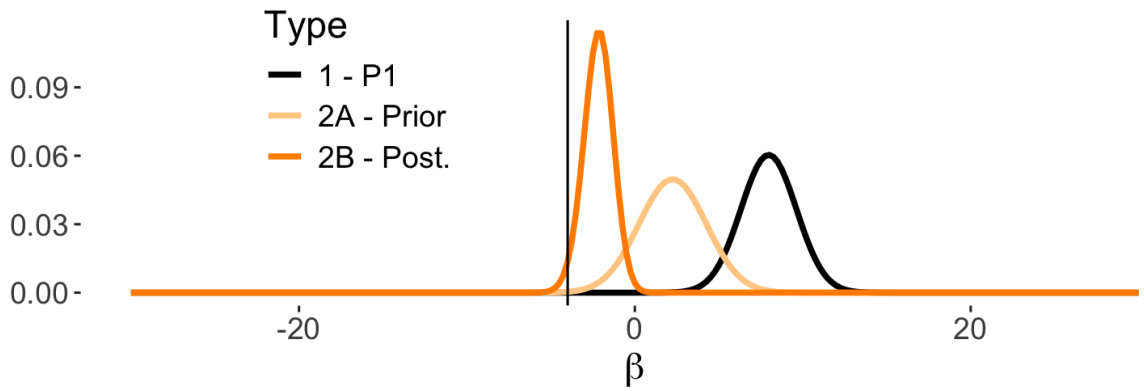
Supplementary Figure 1. Group Level Parameter Values. BPD participants were explained by M4 which has two extra free parameters than CON participants who were best explained by M1.



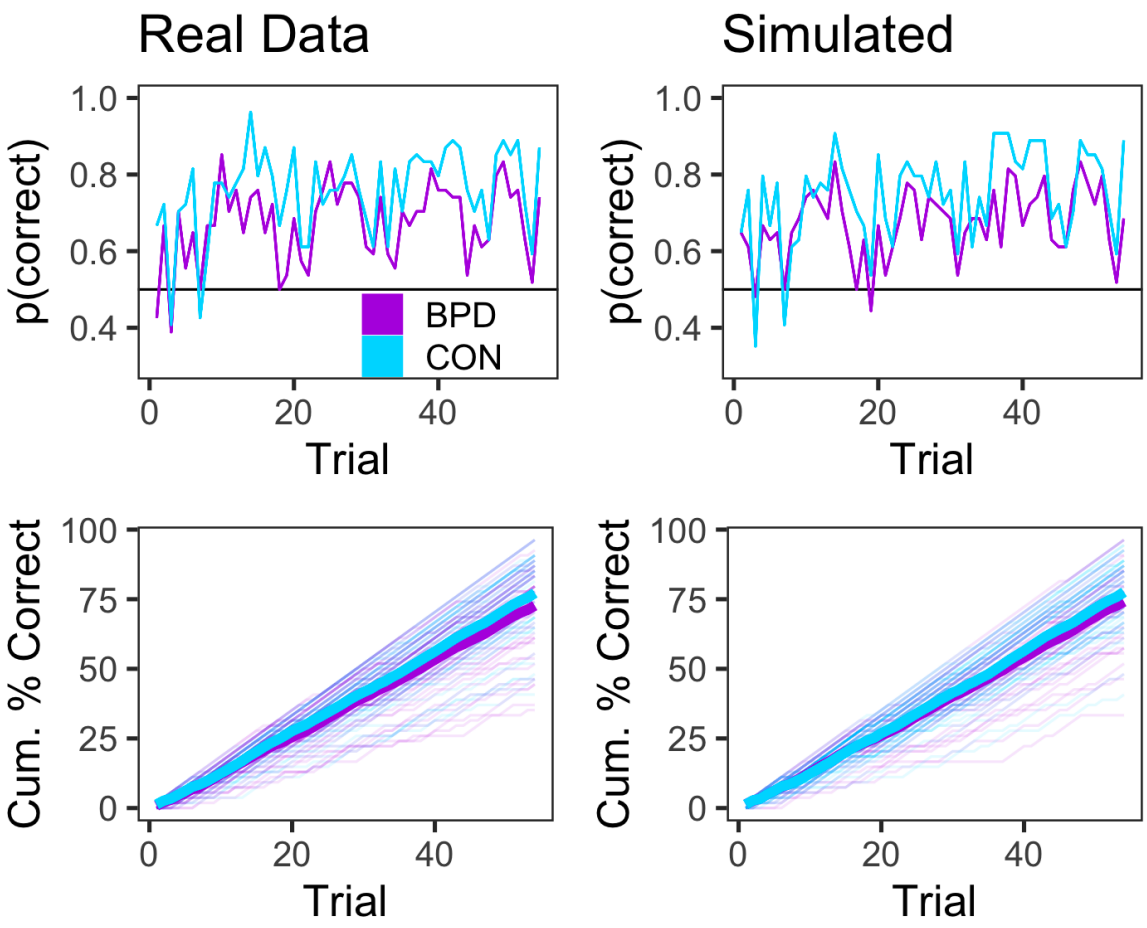
Supplementary Figure 2. Individual Level Parameter Distributions Per Group. BPD (purple) participants were explained by M4 which has two extra free parameters (alpha_par) and (beta_par) than CON participants (blue) who were best explained by M1.



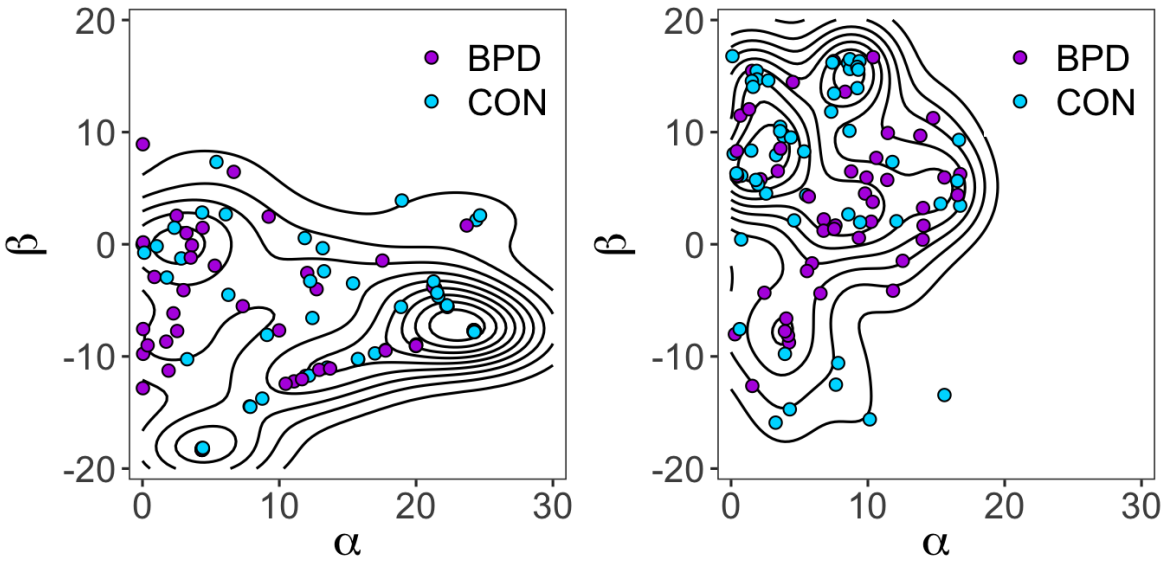
Supplementary Figure 3. Simulation of Phase 2 priors that may be drawn from a memory of an aversive other vs from the self alone. We can imagine a scenario where a prosocial participant (typical of BPD and CON) has a strong impression of an other from memory who is particularly aversive (competitive). Using a mixture of the median belief of the self (β^m_{ppt};classified in phase 1) and a mixture of the belief about how this notional competitive other (β^m_{np}) would act we can create a causal model of how priors in phase 2 about an anonymous partner might draw on different sources. Here, the median of the prior over the partner in phase 2 is a mixture of median belief of self and 'notional' other [$\bar{\beta}^m_{par} = \beta^m_{ppt}\omega + \beta^m_{np}(1-\omega)$]. An equal mixture of self and other belief would equally explain the naïve prior BPD participants hold over their partner in phase 2. However, as mentioned, given that BPD participants hold a naïve prior even when they are themselves competitive goes against this hypothesis. It is worth testing.



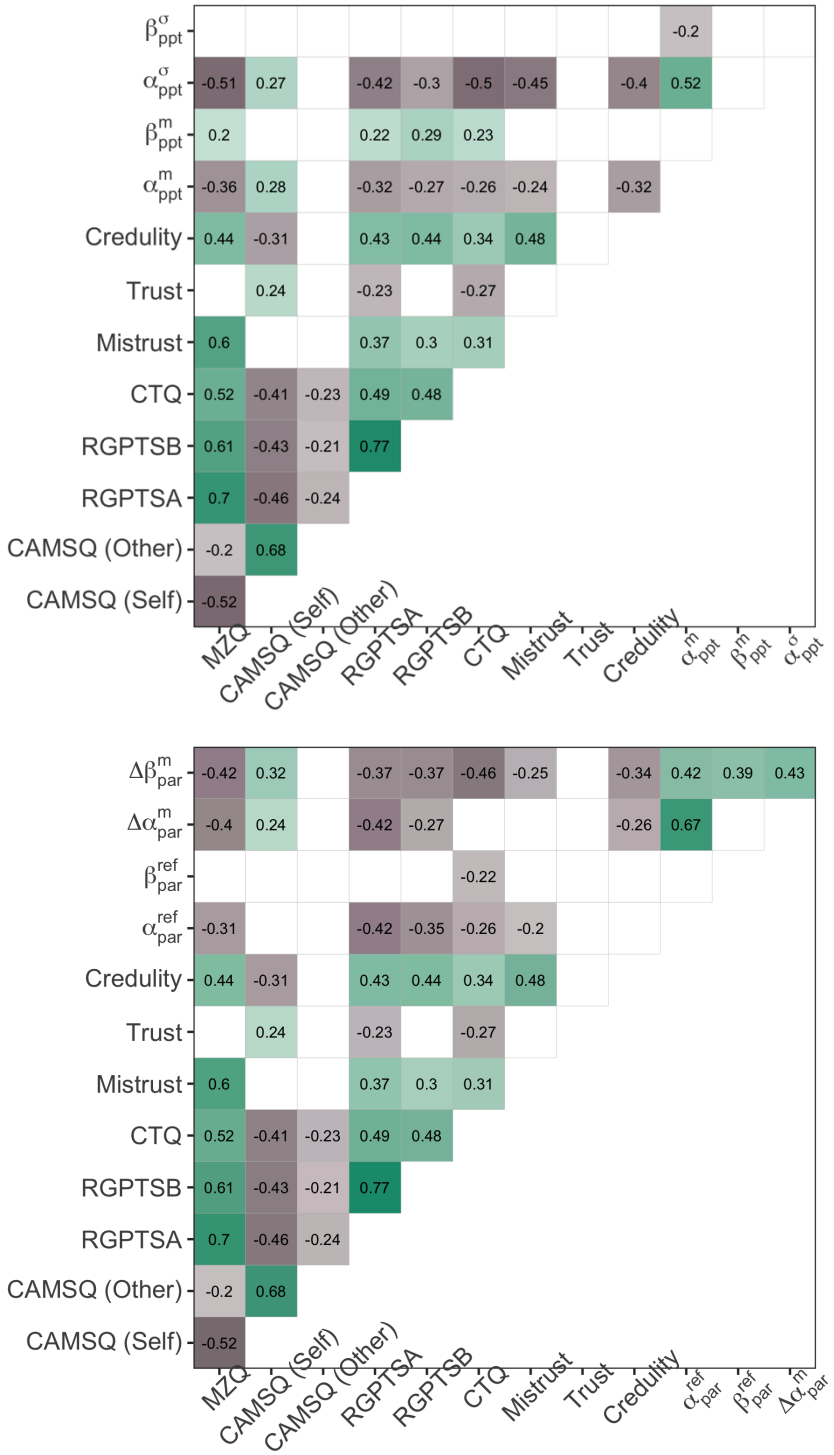
Supplementary Figure 4. Exemplar distribution from an individual with a diagnosis of BPD who was competitive in phase 1 and matched with a partner who was prosocial in phase 2. We note that irrespective of the valence of BPD participants' preferences, there was still a neutral prior generated that was not integrated into the model of self.



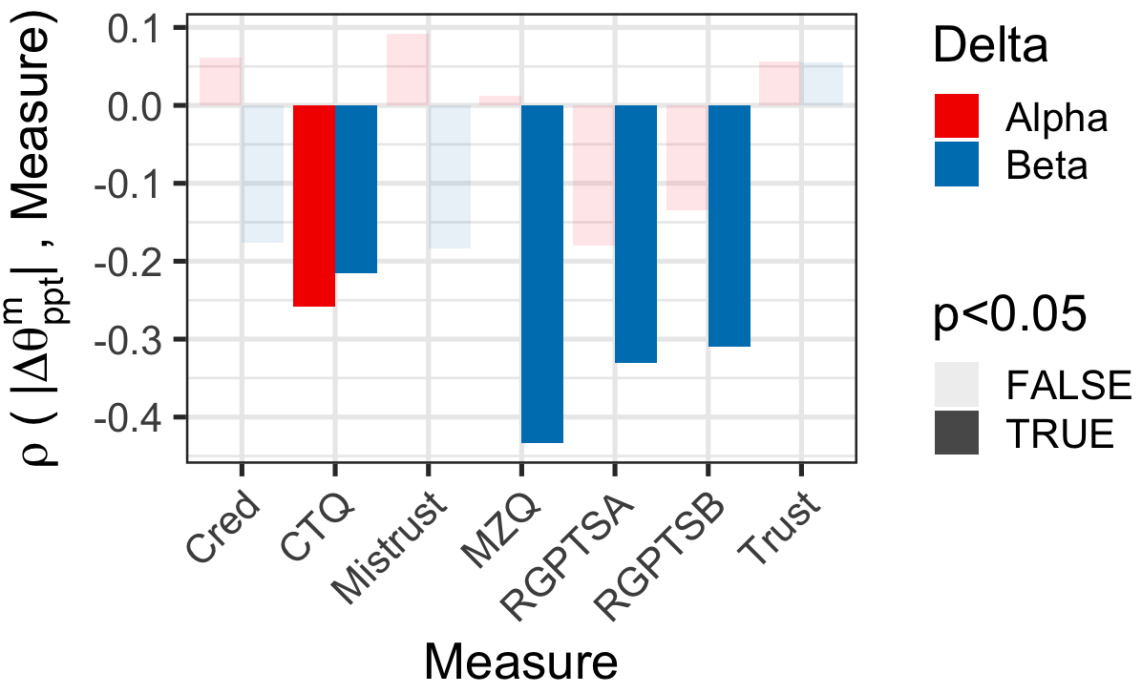
Supplementary Figure 5. (top panels) Raw trial-wise probability of correct responses from real and model-simulated observations for each group. Probabilities were approximated by grouping by trial across each group, summing the total correct responses and dividing by 54. (bottom panel) Cumulative percentage of correct predictions in phase 2 for each group are shown as thick solid lines. Individual cumulative scores are depicted as thin translucent lines.



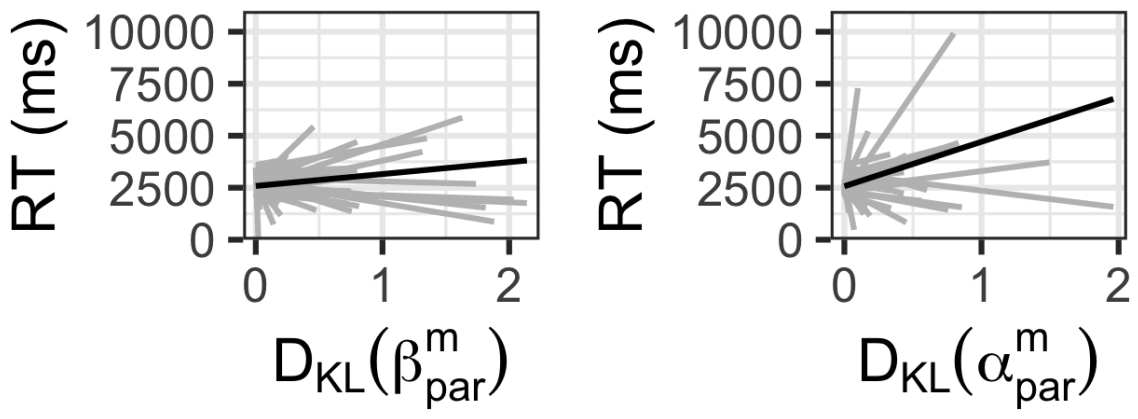
Supplementary Figure 6. 2D Distribution of participant and partner parameters estimated through Bayesian inference at the AWS server backend during the participant-partner matching protocol. As a sanity check we also assessed the degree to which server-derived participant parameters ($\hat{\alpha}^m_{ppt}, \hat{\beta}^m_{ppt}$) matched model-fitting derived model parameters; any discrepancy may have inappropriately matched partners to participants on the server-side. We observed excellent correlations between server-derived participants (not used for analysis; only for partner matching in game) and model-derived phase 1 parameters [$r(\alpha^m_{ppt}, \hat{\alpha}^m_{ppt}) = 0.85$, $p<0.001$; $r(\beta^m_{ppt}, \hat{\beta}^m_{ppt}) = 0.83$, $p<0.001$].



Supplementary Figure 7. Spearman Correlations Between Psychometric Scores at Baseline and Self/Other Parameters. (Top) Psychometric correlations with parameters for self. **(Bottom)** Psychometric correlations with parameters for other. All correlations with p-values > 0.05 are omitted.



Supplementary Figure 8. Spearman's ρ between psychometric measures and change absolute change in self-preferences from phase 1 to 3. All beliefs metrics are extracted from M3 which assumes all participants engage in social contagion. Cred = Credulity. Delta = whether the shift in belief was along preferences for absolute (alpha) or relative (beta) reward.



Supplementary Figure 9. Linear random effects relationship between reaction time (ms) and belief updating. Grey lines are individual participants. Black line is the average linear effect. Reaction time is capped at 10000ms for visual illustration, but linear models do not apply an upper limit.

Supplementary Table 1. Option pair rewards for each phase and their corresponding 'type'. Within phase order of trials were randomised. P=Prosocial, I=Individualistic, C=Competitive. S1 = reward to self for option 1. S2 = reward to self for option 2. O1 = reward to other for option 1. O2 = reward to other for option 2.

S1	O1	S2	O2	Self-Disparity	Other Disparity	Type
10	6	8	2	2	4	I-C
7	7	10	7	3	0	P-I
7	1	8	5	1	4	C-I
10	5	10	10	0	5	C-P
12	9	9	9	3	0	I-P
10	5	8	1	2	4	I-C
6	2	8	6	2	4	C-I
8	2	9	6	1	4	C-I
5	5	5	1	0	4	P-C
7	7	7	2	0	5	P-C
12	8	8	8	4	0	I-P
8	8	8	2	0	6	P-C
9	5	7	1	2	4	I-C
6	6	8	6	2	0	P-I
6	1	7	5	1	4	C-I
12	6	10	2	2	4	I-C
7	7	7	1	0	6	P-C
10	6	6	6	4	0	I-P
4	4	8	4	4	0	P-I
11	6	9	2	2	4	I-C
5	1	7	5	2	4	C-I
8	5	8	8	0	3	C-P
6	6	10	6	4	0	P-I
6	3	6	6	0	3	C-P
5	5	8	5	3	0	P-I
11	5	9	1	2	4	I-C
9	5	9	9	0	4	C-P
7	3	7	7	0	4	C-P
8	3	8	8	0	5	C-P
12	10	10	10	2	0	I-P
10	8	8	8	2	0	I-P
6	6	6	1	0	5	P-C
6	6	6	2	0	4	P-C
7	2	8	6	1	4	C-I
10	7	7	7	3	0	I-P
8	8	10	8	2	0	P-I