

ARTICLE OPEN



A computational signature of self-other mergence in Borderline Personality Disorder

Giles W. Story^{1,2}, Sam Ereira^{2,3}, Stephanie Valle⁴, Samuel R. Chamberlain^{5,6}, Jon E. Grant⁴ and Raymond J. Dolan²

© The Author(s) 2024

A tendency to merge mental representations of self and other is thought to underpin the intense and unstable relationships that feature in Borderline Personality Disorder (BPD). However, clinical theories of BPD do not specify, in computational terms, how the perspectives of self and other might become confused. To address this question, we used a probabilistic false belief task (p-FBT) to examine how individuals with BPD ($N = 38$) and matched controls from the general population ($N = 74$) selectively assigned beliefs to self or other. The p-FBT requires participants to track a gradually changing quantity, whilst also predicting another person's belief about that quantity. We found that BPD participants showed less selectivity in belief assignment compared with controls (Cohen's $d = 0.64$). Behaviourally, participants with BPD tended to predict that others' beliefs resembled their own. Modelling analysis revealed that BPD participants were prone to generalise their own learning signals to others. Furthermore, this generalising tendency correlated with BPD symptomatology across participants, even when controlling for demographic factors and affective psychopathology. Our results support a computational account of self-other mergence, based on a generalisation of learning across agents. Self-other generalisation in learning purports to explain key clinical features of BPD, and suggests a potential transdiagnostic marker of mentalising capability.

Translational Psychiatry (2024)14:473; <https://doi.org/10.1038/s41398-024-03170-w>

INTRODUCTION

In Shakespeare's play, 'Romeo and Juliet', Romeo takes his own life, believing Juliet to be dead. The audience, on the other hand, aware that Juliet has feigned her death, knows that Romeo is tragically mistaken. The dramatic effect rests on the audience's ability to represent Romeo's false belief, distinct from their own knowledge of the true state of affairs (see [1]). More broadly, this capacity to separately represent the mental states of self and other, termed 'mentalising', is fundamental to the effectiveness of our social interactions [2–9].

Limitations in mentalising are thought to underpin a syndrome of unstable self-identity and intense, brittle relationships, known as 'Borderline Personality Disorder' (BPD) [10–23]. This idea gains support in findings that people diagnosed with BPD exhibit reduced performance on false belief tasks [24–26], wherein subjects need to distinguish their own knowledge of a situation from that of a fictive observer, who has only partial information [27–29]. BPD participants also tend to perform less well than healthy controls in describing emotions from another person's perspective [30–32], and in understanding when another person has made a *faux pas* (for instance, when a person reveals knowledge that ought to be hidden) [33–36]. These findings suggest that people with BPD have difficulty representing others' mental states and distinguishing these from their own.

In clinical terms, a difficulty in subjectively separating mental states belonging to self and other putatively leads to a heightened vulnerability to others' responses [14, 21–23]. For example, mentalising-based theories of BPD invoke a mode of 'psychic equivalence', wherein a person equates their own mental processes with external reality [2, 17–19]. Mental states then carry the force of objective facts, such that others are expected to know and share one's thoughts and feelings. In such merged states, disagreement may be experienced, not merely as a difference in opinion or perspective, but as a confusing denial of what was assumed to be a shared reality [12, 20]. Theories across clinical traditions converge on a notion that such disturbance in the capacity to distinguish self from other predisposes to relationships that oscillate between intense closeness and precipitous rupture [12–16].

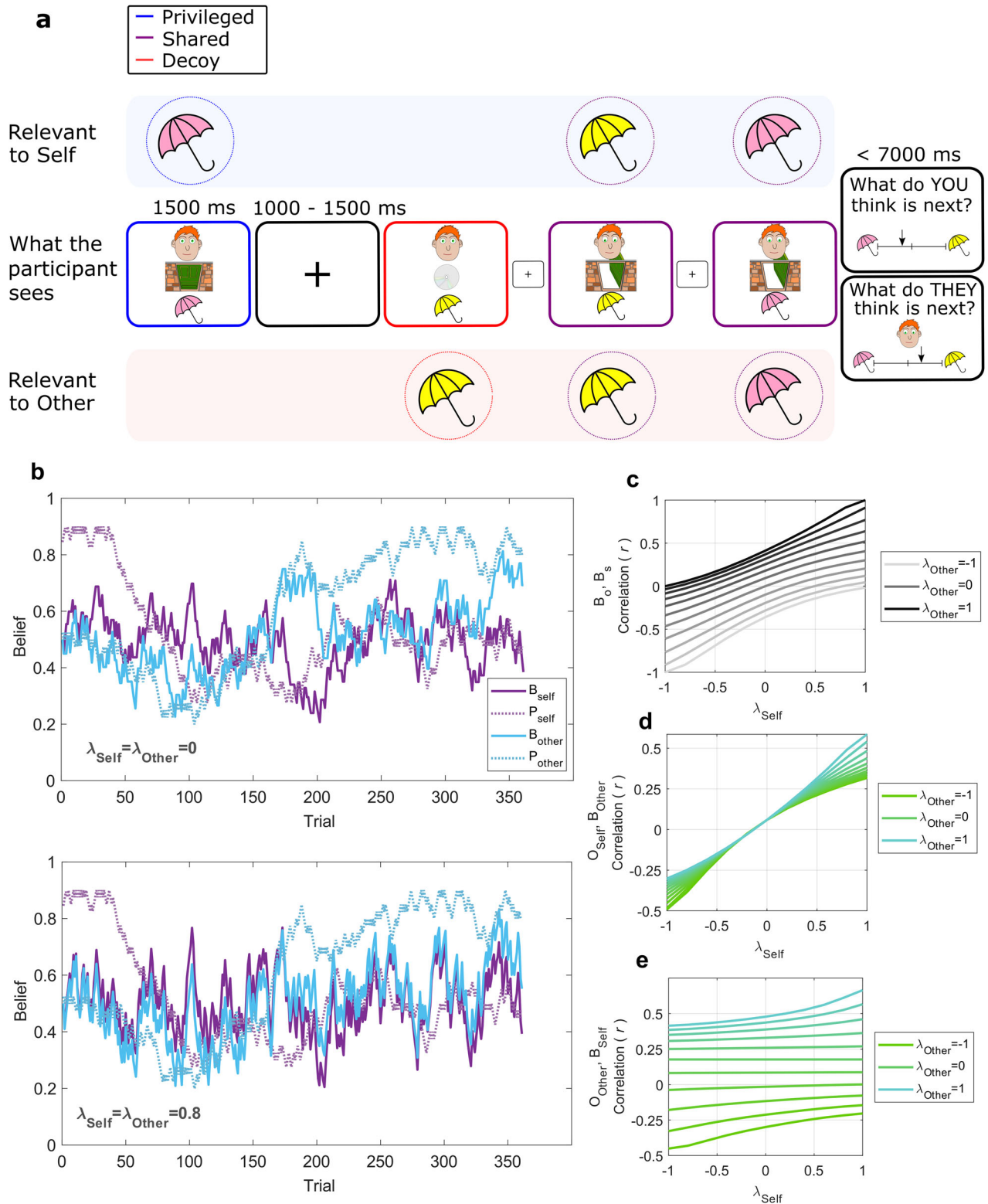
However, a shortcoming of existing theories is that they do not specify, in computational terms, how people selectively ascribe mental states to self or other. Nor do clinical theories explain how this selectivity might go awry in BPD. We address this by measuring belief updating using a novel probabilistic false belief task (p-FBT) (Fig. 1) [37, 38]. Previous false belief tasks classically use one-shot verbal or pictorial narratives wherein the beliefs of self and other are constrained to be either identical or opposite [24–26]. By contrast, the p-FBT requires participants to track a gradually changing quantity, whilst also predicting another

¹Division of Psychiatry, University College London, London, UK. ²Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK. ³Preventative Neurology Unit, Queen Mary University, London, UK. ⁴Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA. ⁵Department of Psychiatry, Faculty of Medicine, University of Southampton, Southampton, UK. ⁶Southern Health NHS Foundation Trust, Southampton, UK.

email: g.story@ucl.ac.uk

Received: 21 August 2023 Revised: 9 October 2024 Accepted: 24 October 2024

Published online: 19 November 2024



person's belief about that quantity. The p-FBT thereby allows beliefs attributed to self and other to develop along trajectories with varying degrees of correlation (Fig. 1b, c), enabling us to apply a model to interrogate the extent to which learning signals are selectively attributed to Self or Other. We administered the p-FBT in participants with diagnoses of BPD ($N=38$) and in control participants recruited from the general population

($N=74$), matched to the former on age (± 4 years), gender and education.

Using computational modelling, we formalise a continuum between *agent-specific* updating, wherein beliefs of Self and Other are maintained separately, and *agent-invariant* updating, wherein Self and Other collapse to a single belief representation. We consider two directions of agent-invariant updating. Firstly,

Fig. 1 Probabilistic false-belief task (p-FBT). **a** Adapted from Ereira et al. (2020) [49]. Participants observed samples (purchases of pink umbrellas or yellow sunshades at a store) from a Bernoulli distribution. An hypothetical other person (the store manager), received partial or misleading information. On 'Privileged' trials, samples were hidden from the other person (shown as the store manager's face behind a closed door). On 'Shared' trials the other person saw the same sample as the participant (shown as the manager's face looking through an open door). On 'Decoy' trials the other person received a false sample (described as the manager's watching out of date security camera footage, shown as the manager's face above an image of a computer disc). Participants were intermittently asked to predict either the next sample, or the other person's belief about the next sample, by placing a cursor on a continuous probability scale. **b** The underlying probability, $P(\text{umbrella})$, varied across time, with differing trajectories for the participant (P_{Self}) and other person (P_{Other}). Simulated belief trajectories are depicted for either fully agent-specific updating ($\lambda_{\text{Self}} = \lambda_{\text{Other}} = 0$, $\alpha = 0.1$, $\tau = 0.04$ and $\delta = 0$), or leakage in updating ($\lambda_{\text{Self}} = \lambda_{\text{Other}} = 0.8$), showing that leakage induces a correlation in beliefs across agents. **c** Pearson correlation in beliefs for Self (B^{Self}) and Other (B^{Other}) derived from model simulation on the task with varying settings of λ_{Self} (horizontal axis) and λ_{Other} (solid lines) ($\alpha = 0.1$, $\tau = 0.04$ and $\delta = 0$). **d** λ_{Self} induces a correlation between a recency-weighted average of outcomes pertinent to Other and B^{Self} . **e** λ_{Other} induces correlation between outcomes pertinent to Self and B^{Other} .

updating can be 'idiocentric', denoting the fact that one's own learning signals generalise to others. Thus, events hidden from another person still influence one's estimate of the other's belief, akin to the mentalising errors seen in conventional false belief tasks [27–29]. On the p-FBT, this induces a correlation between estimates of the Other's belief and outcomes observed by Self (Fig. 1d).

Secondly, updating can be 'allocentric', denoting generalisation of learning signals from Other to Self. Under allocentric updating, seeing another person observe false information nevertheless influences what the subject comes to believe about the world. Within the p-FBT, the subject sometimes sees the other person observe corrupted information. Allocentric updating induces a correlation between the subject's belief and outcomes observed by the Other, even though these may be misleading (Fig. 1e).

Our central prediction was that participants with BPD would exhibit greater agent-invariant updating compared with participants sampled from the general population and that agent-invariant updating should relate to BPD symptom expression, as measured by the Borderline Symptom List [39]. We did not have strong prior hypotheses regarding whether BPD participants would tend to exhibit either idiocentric or allocentric updating.

MATERIALS AND METHODS

BPD participants

BPD participants were recruited to a randomised placebo-controlled trial situated at the University of Chicago, designed to assess the effect of an antipsychotic medication, brexpiprazole, in the treatment of BPD [40]. Recruitment took place between June 2018 and December 2020. Procedures were approved by the University of Chicago Institutional Review Board. The trial was registered with Clinicaltrials.gov (study identifier NCT03418675). Informed consent was obtained from all participants. Participants were compensated \$200 for their participation, which entailed ten study visits.

A total of 80 BPD participants were recruited to the clinical trial, of which 38 completed the p-FBT. Data collection of the p-FBT was interrupted at the onset of the Covid-19 pandemic, when clinical trial follow-up moved online, at which point 29 participants had completed the task. The majority of BPD participants recruited to the trial after this point did not complete the p-FBT; nine BPD participants completed the task online, in a supervised online assessment. The experiment and instructions were identical for all participants regardless of whether performed online or in-person.

To be included in the trial, participants needed to be aged 18–65 years, have a primary diagnosis of BPD, and score at least nine points on the Zanarini rating scale for Borderline Personality Disorder (ZAN-BPD). All participants also met criteria for a BPD diagnosis according to DSM-5 criteria, confirmed by structured clinical interview. Further exclusion criteria, as documented by the authors of the trial [40], are listed in Supporting Material.

Trial participants were assigned to 13 weeks of either brexpiprazole, or placebo treatment. All participants included in our study completed the p-FBT at the baseline visit, before being allocated to either drug or placebo treatment. A small number of participants ($N = 26$) completed a second p-FBT at one of their further study visits, enabling us to assess drug effects,

though this was not our primary interest (see Supporting Material for results).

Control participants from the general population

Control participants were recruited from the general population in the United Kingdom via the online platform Prolific.co. The experiment and written task instructions were identical for the control and BPD participants. Control participants were matched to the BPD sample for age (± 4 years), gender (male, female, non-binary), and educational level (high school diploma/GCSE/A-level, bachelor's degree, post-graduate degree). We did not explicitly screen these participants for diagnoses of BPD. Nor did we stipulate additional inclusion or exclusion criteria besides the aforementioned matching characteristics. Rather, we addressed the possibility that some control participants might have BPD symptoms by correlating model parameters with self-reported symptom scores.

To increase statistical power, we aimed to recruit two matched control participants for each BPD participant, i.e. 76 control participants. We successfully recruited 75 matched controls. After consenting to take part, and prior to completing the task, control participants met with an experimenter through a video link to ensure that they were able to complete the task in a quiet place, free from distractions or interruptions. One control participant was excluded prior to data analysis since the video-call established that they were in a noisy environment (final $N = 74$ controls).

Procedures for control participants were approved by the University College London Research Ethics Committee (project ID number 20399/001). Participants were compensated for their time at a rate of at least £8.91 per hour. Informed consent was obtained from all participants. All participants had study procedures fully explained to them, and had an opportunity to ask any questions before consenting to take part.

Probabilistic false belief task (p-FBT)

In the p-FBT, participants observed binary outcomes sampled from a Bernoulli distribution, with underlying probability, P^1 . Within the cover story of the task, participants observed whether successive customers arriving at a tourist store on a tropical island purchased either pink umbrellas (1) or yellow sunshades (0) [37]. Periodically, participants were asked to predict P , based on recent purchases.

An hypothetical other person, the 'store manager', received partial or misleading information about outcomes (Fig. 1a). The probability (P) of observing a given outcome, therefore, evolved with differing trajectories for the participant (P_{Self}) and the manager (P_{Other}) (Fig. 1b). Specifically, participants were told that the manager was sometimes with them, and able to see purchases. These 'Shared' sampling trials were indicated to the participant as a cartoon face looking through an open door (Fig. 1a). Sometimes, however, the manager was in a back room, and could not see purchases. These 'Privileged' sampling trials were indicated as a cartoon face behind a closed door. Finally, participants were told that the manager sometimes observed purchases on a security camera; however, unbeknown to the manager, the footage was out of date, and thereby uninformative about what customers were really buying. These 'Decoy' sampling trials were indicated as a cartoon face and an icon showing a

¹A Bernoulli distribution is the probability distribution of a random variable which takes the value 1 with probability P and the value 0 with probability $1 - P$.

computer disc. Each *sampling* trial was presented for 1250 ms, followed by a variable intertrial interval with a fixation cross on screen for 400–600 ms.

Pre-generated sequences of 360 outcomes (purchases) were sampled from two uncorrelated random walks, governing P_{Self} and P_{Other} (with step sizes of 0.025, bounded between 0 and 1). ‘Privileged’ samples were drawn from P_{Self} whilst ‘decoy’ samples were drawn from P_{Other} . ‘Shared’ samples were drawn randomly from either walk.

After observing a series of purchases (every 4–9 sampling trials), participants were asked to predict the upcoming purchase, in a *probe* trial. Specifically, participants were asked to report the probability that the next customer would buy a pink umbrella as opposed to a yellow sunshade, by placing a cursor on a continuous probability scale, either from their own perspective (‘Self’ probe trials; P_{Self}) or from the perspective of the manager (‘Other’ probe trials; P_{Other}). In the instructions concerning Other-probe trials, participants were briefed to “pay attention to which sales the manager has and has not seen (including what he has seen on the security footage)”.

Participants were provided with thorough written instructions on the task, including an introductory example, explaining the different trial types and how to use the rating scale. Following this practice phase, participants were asked to rate their understanding of the instructions on a five-point scale, from “Very poor—I do not understand what I am being asked to do” to “Very good—I clearly understand what I am being asked to do”. Participants were asked to repeat the instructions and practice phase if they rated their understanding as poor or very poor. 89% of participants rated their understanding of the instructions as ‘Very good’ or ‘Good’.

Aside from the introductory example, wherein feedback was given, participants received no feedback about their performance. They were briefed that the task was subjectively difficult, and encouraged to do their best. A self-paced break was introduced halfway through the task, which was approximately 15 min in duration.

A laboratory (in-person) version of the task was coded using Matlab (Mathworks, Provo), and visualised using Cogent 2000 (v125) and Cogent Graphics (v1.29). An identical online version was coded using Gorilla.sc.

Computational model of agent-specific belief updating

We fitted participants’ responses with a simple reinforcement-learning model, wherein beliefs (B) about P are updated proportionally to a *prediction error* (PE), given by the difference between new observations (1 or 0) and existing predictions. To model agent-specificity, we instantiated parallel learning processes for Self and Other, as follows:

$$\begin{aligned} B_t^{\text{Self}} &= B_{t-1}^{\text{Self}} + \alpha (PE_t^{\text{Self}} + \lambda^{\text{Other}} \cdot PE_t^{\text{Other}}) + \delta (0.5 - B_{t-1}) \\ B_t^{\text{Other}} &= B_{t-1}^{\text{Other}} + \alpha (PE_t^{\text{Other}} + \lambda^{\text{Self}} \cdot PE_t^{\text{Self}}) + \delta (0.5 - B_{t-1}) \end{aligned} \quad (1)$$

Here, B^{Self} denotes the subject’s belief, while B^{Other} denotes the subject’s representation of the other person’s belief (bounded between 0 and 1). Thus, two belief representations can be separately updated, based on observations available to each agent, and this design feature furnishes agent-specific prediction errors (PE^{Self} and PE^{Other}). $0 < \alpha < 1$ is a learning rate parameter, governing sensitivity to new information, while $0 < \delta < 1$ is a memory decay rate, whereby beliefs drift towards chance level. Importantly, updates for Self and Other may not be fully segregated. We capture this effect by invoking ‘leakage’ parameters, λ^{Self} and λ^{Other} , bounded between -1 and 1, which allow prediction errors pertinent to belief representation of one agent to erroneously update the belief representation of the other agent.

This form of model has previously been found to approximate responses well on the p-FBT [37, 38]. In keeping with previous approaches, PE^{Self} was modelled as 0 on Decoy trials, while PE^{Other} was modelled as 0 on Privileged trials [37, 38]. Distinct from previous applications, we allowed negative values of λ^{Self} and λ^{Other} , which allows for anti-correlation in Self- and Other-attributed beliefs (see below).

The task was designed such that, when $\lambda^{\text{Self}} = 0$ and $\lambda^{\text{Other}} = 0$, corresponding to *agent-specific* updating, belief trajectories of Self and Other are uncorrelated (Fig. 1b, c). However, leakage in updates induces a correlation in beliefs across agents; as either λ^{Self} or λ^{Other} approaches one, corresponding to *agent-invariant* updating, beliefs for Self and Other converge (Fig. 1b, c). The model also allows for anti-correlation in predictions for Self and Other, which occurs when leakage parameters are less than zero (Fig. 1c). Anti-correlated beliefs imply that, if the subject predicts a given outcome (e.g., pink umbrella), they tend to think the other person will predict the opposite (e.g., yellow sunshade). Leakage parameters thus capture a continuum between self-other merge on

the one hand and self-other differentiation on the other. We hypothesized that participants with BPD would exhibit higher degrees of agent-invariance in belief updates compared with general population controls, manifest as more positive λ .

Finally, the model allows asymmetric leakage, which occurs if $\lambda^{\text{Self}} \neq \lambda^{\text{Other}}$. Here, λ^{Self} captures how a prediction error experienced by the self generalises to another agent. We refer to such updating as ‘*idiocentric*’, denoting the fact that, in this instance, one’s own observations generalise to others. Behaviourally, this induces a correlation between estimates of the Other’s belief and outcomes observed by Self (Fig. 1d).

Updating can also be ‘*allocentric*’, denoting generalisation of prediction error from Other to Self, captured by λ^{Other} . Under allocentric updating, this information influences what the subject comes to believe about the world, inducing a correlation between the subject’s belief and outcomes observed by the Other (Fig. 1e).

Model fitting procedure

We tested models with the following configurations of leakage parameters: (i) independent λ^{Self} and λ^{Other} (λ Unrestricted, as shown in Eq. (1)). (ii) $\lambda^{\text{Self}} = \lambda^{\text{Other}}$ (i.e., symmetric leakage), and (iii) a both λ parameters set to zero (i.e., a fully agent-specific model). We tested each of these three models with learning rates for Self and Other which were either (i) constrained to be equal ($\alpha = \alpha^{\text{Self}} = \alpha^{\text{Other}}$, as shown in Eq. (1) or (ii) could differ (α Unrestricted). This yielded a set of six models.

Each model yields a point estimate of B^{Self} and B^{Other} on each trial. These model-derived belief estimates were used to derive the likelihood of a participant’s responses on probe trials. Since participants were asked to estimate P_{Self} and P_{Other} , which are bounded between 0 and 1, we used a Beta distribution for the likelihood, with a mode set to the model-derived point estimate (for details see Supporting Material). For each model, the variance of the Beta distribution was set to a participant-specific parameter, τ , governing the degree of stochasticity in responses. Higher settings of τ entailed more random responding. For each model, choice temperature (τ) and memory decay (δ) parameters were set to be identical for Self- and Other-updates. Parameters were bounded through sigmoid transformation and fitted in inverse sigmoid space. Both α and δ parameters were bounded between 0 and 1; τ between 0.001 and 0.08 (this upper bound on τ produces a near-uniform Beta distribution, corresponding to random responding).

For each participant we sought model parameters with maximum posterior probability, given the data. We used fixed Gaussian priors over parameters and assumed a uniform prior over models (see Supporting Material for details of the model fitting procedure). For each model, we calculated the Bayesian model evidence, which favours models that provide a closer fit to the data, whilst penalising more complex models. We first compared models in a fixed-effects approach, summing model evidences across participants. We also examined the frequency of models across participants, in a random-effects approach. For each model we calculated a ‘protected exceedance probability’, a statistic estimating the probability that a given model amongst a set of models is most frequently the best-fitting [41, 42].

To measure parameter recovery for the best fitting model, we fitted the model to data generated for 560 simulated participants. Simulated participants’ parameters were randomly sampled from the observed maximum a posteriori parameter estimates. Each parameter was sampled independently, with replacement. We computed a Pearson correlation coefficient between the generative parameters and fitted parameter estimates.

Outcome-belief correlations

We derived ‘model-free’ measures of belief updating on the task. Firstly, we measured task performance by correlating a participant’s reported beliefs about P_{Self} and P_{Other} on probe trials with an exponential recency-weighted average, O , of past outcomes for Self and Other respectively (see Supporting Material for details). We used a recency-weighted average, rather than the true generating P , so as to account for sampling error in the observed outcomes. For each participant’s trial sequence, we generated a null distribution for this outcome-belief correlation, by generating 1000 simulated sets of random responses. We expressed performance relative to that expected under this null distribution, such that a score of zero indicates chance-level responding (shown in Fig. 2a). Secondly, we examined performance in each participant group for Self and Other probe trials separately, i.e., *within-agent* outcome-belief correlations.

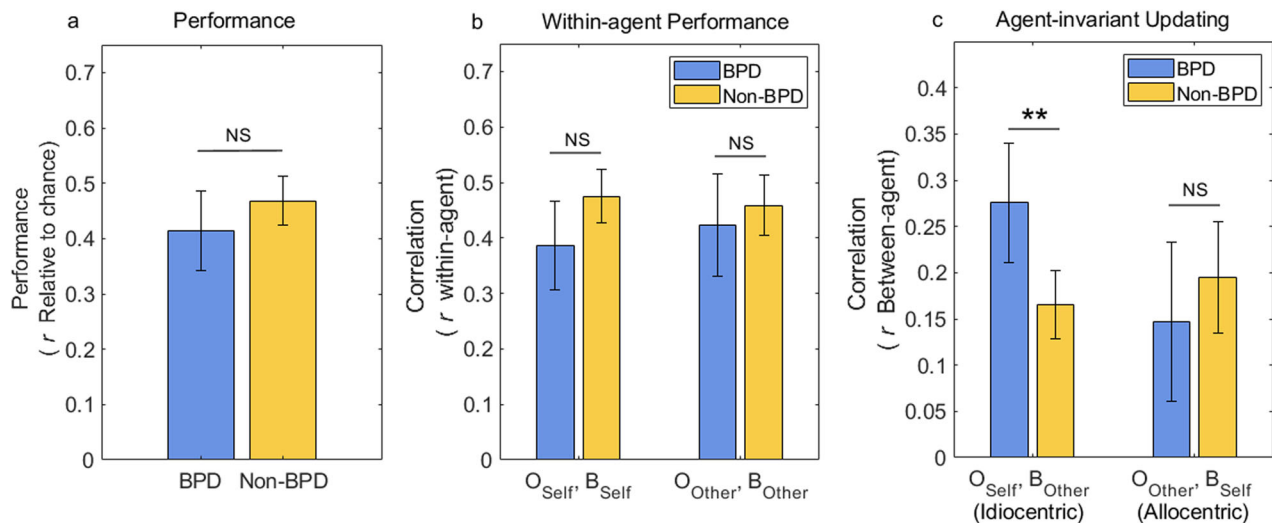


Fig. 2 Task performance and outcome-belief correlations. **a** BPD and Non-BPD groups did not differ significantly in performance, expressed as the correlation between participants' belief estimates and a recency-weighted average of past outcomes ($t(110) = -1.38$, two-tailed $p = 0.170$). Performance is shown relative to that expected under random responding, such that zero indicates chance level. **b** Performance by an agent. Correlations are shown between outcomes pertinent to Self (O^{Self}) and the subject's reported belief (B^{Self}) on Self probe trials, and between outcomes pertinent to Other (O^{Other}) and belief attributed to the Other (B^{Other}) on Other probe trials. Performance for neither agent differed significantly between BPD and Non-BPD groups (Self: $t(110) = -1.06$, $p > 0.25$; Other: $t(110) = -0.03$, $p > 0.25$). **c** Correlation between outcomes observed by Self (O^{Self}) and belief attributed to the Other (B^{Other}), a measure of idiocentric updating, was significantly greater in the BPD group than in the Non-BPD group ($t(110) = 3.12$, $p = 0.002$). Correlation between outcomes observed by the Other (O^{Other}) and the subject's reported belief (B^{Self}), a measure of allocentric updating, did not differ significantly between groups ($t(110) = -0.86$, $p > 0.25$). Error bars show 95% confidence intervals. ** $p < 0.01$, NS: $p > 0.05$.

Finally, as a behavioural measure of agent-invariant updating, we examined *across-agent* outcome-belief correlations using the same method. Correlation between estimates of the Other's belief (B^{Other}) and a weighted-average of outcomes observed by Self (O^{Self}) (on Privileged and Shared trials), measures idiocentric updating, while the correlation between the subject's reported belief (B^{Self}) and outcomes observed by the Other (O^{Other}) (on Decoy and Shared trials) measures allocentric updating.

Symptom questionnaires

All participants completed self-reported questionnaires to assess BPD severity, of which our primary outcome of interest was the Borderline Symptom List [39] (BSL-95). Participants also completed the Zanarini Self-Report scale [43] (Zan-SR). Impulsivity was assessed using the Barratt Impulsiveness Scale (BIS-11) [44]. In addition, we used standardised questionnaires to assess affective symptoms; for BPD participants, clinician-rated measures were administered, namely the Hamilton Depression Rating Scale [45] (HAM-D), and Young Mania Rating Scale [46] (YMRS), while control participants completed self-report scales, namely the Inventory of Depressive Symptomatology [47] (IDS), and Altman Mania Rating Scale [48] (AMRS).

To examine relationships between BPD symptoms and agent-invariant updating, we used univariate simple linear regression, with λ^{Self} (or λ^{Other}) as the dependent variable, and BSL-95 score as the independent variable. To account for possible between-group differences in the effect of symptom scores, we also included a group (BPD vs. Control) by symptom score (BSL-95) interaction term. In a second regression analysis, we also included symptoms of impulsivity, depression and mania, and demographic factors (age, gender, ethnicity, educational level and employment) as covariates. Since depression and mania scales differed between BPD and Control groups, we expressed total scores as percentages of the maximum score on each scale. All statistical tests reported were two-sided.

RESULTS

Group characteristics

Supporting Table S1 shows demographic and symptom scores for the BPD and control groups. Additional characteristics of the BPD sample, such as co-morbid diagnoses, drug and alcohol use, or concomitant medications, are provided in Supporting Table S2.

Outcome-belief correlations

Both BPD and control groups performed above chance on average (BPD: mean = 0.41, 95% CI [0.34 0.49] one-sample $t(37) = 11.6$, $p < 0.001$; Non-BPD: mean = 0.47 95% CI [0.42 0.51], $t(73) = 21.4$, $p < 0.001$), and performance did not differ significantly between the two groups (Fig. 2a; 95% CI of performance difference [-0.13 0.02], two-sample $t(110) = -1.38$, two-tailed $p = 0.170$). Performance was above chance in 107 out of 112 participants. In 95 out of 112 participants, performance was above the 95th percentile of a null distribution generated by random responding. In pairwise comparisons, performance for neither agent differed significantly between BPD and control groups, as shown by *within-agent* outcome-belief correlations. (Fig. 2b; Self: 95% CI of difference [-0.14 0.04], $t(110) = -1.06$, $p > 0.25$; Other: 95% CI [-0.10 0.10], $t(110) = -0.03$, $p > 0.25$). Thus, we found no evidence that participants were systemically ignoring outcomes pertinent to either agent.

As a behavioural test of agent-invariant updating, we next examined *between-agent* outcome-belief correlations. Pairwise tests revealed a significantly greater idiocentric updating (O^{Self} , B^{Other} correlation) in the BPD group compared to the control group (95% CI of difference [0.04 0.18]; $t(110) = 3.12$, $p = 0.002$), with no significant difference in allocentric updating (95% CI of difference in O^{Other} , B^{Self} correlation [-0.15 0.06]; $t(110) = -0.86$, $p > 0.25$). These results indicate that BPD participants tended to use their own observations to estimate the Other's belief to a greater extent than control participants from the general population.

Model comparison

We went on to fit learning models to participants' belief estimates for Self and Other (Fig. 3). In essence, this involved minimising differences between belief estimates derived from each model, and those reported by participants. Models with independent λ^{Self} and λ^{Other} (λ Unrestricted) outperformed those in which leakage was constrained to be symmetric ($\lambda^{\text{Self}} = \lambda^{\text{Other}}$), or no leakage ($\lambda = 0$) (Fig. 3a, b). Also, models in which learning rates for Self

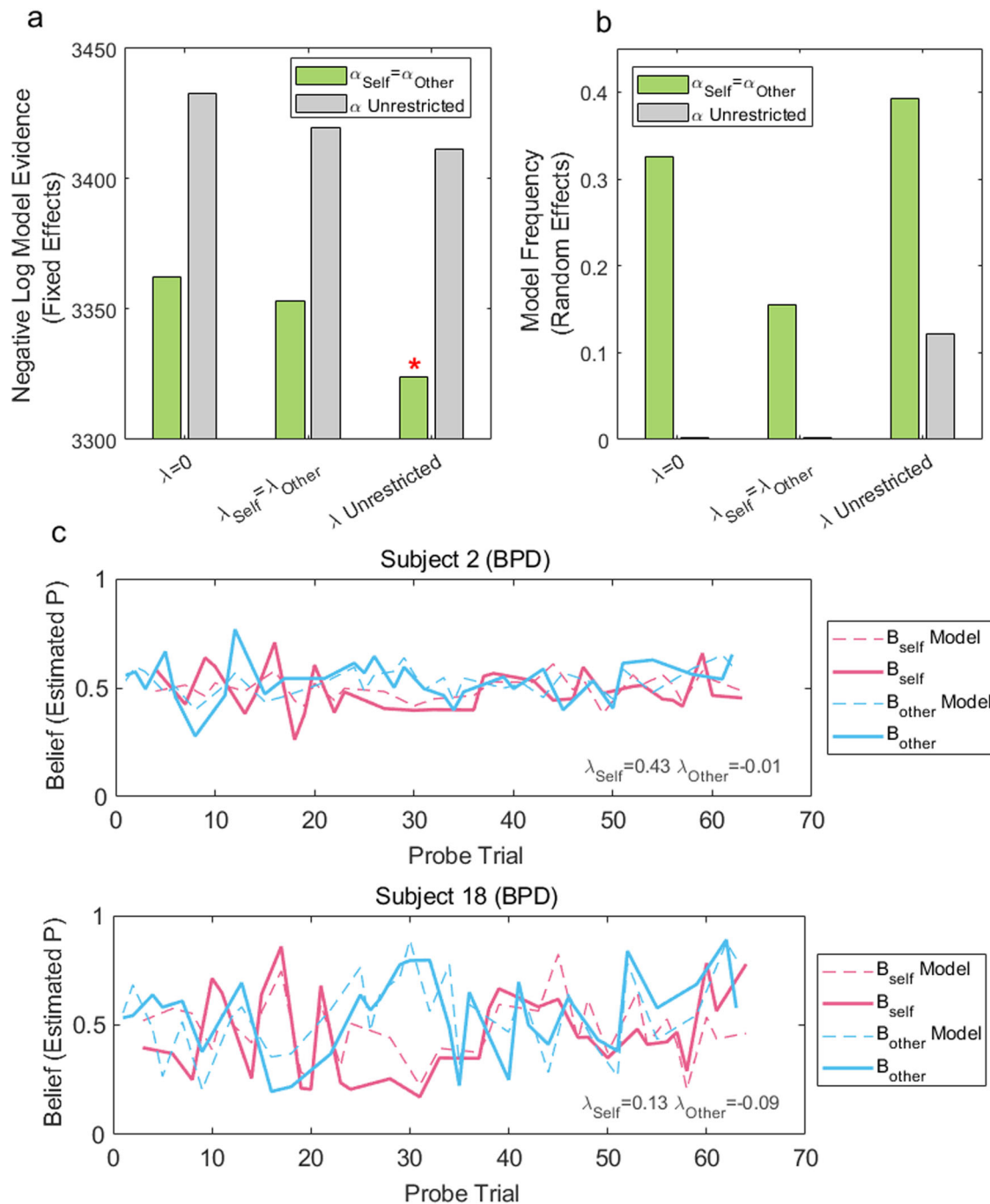


Fig. 3 Model comparison and model fits. **a** Negative log model evidence, summed across all participants ($N=112$), for different model configurations. Lower values indicate a better model fit. $\lambda=0$ entails fully agent-specific learning; $\lambda_{\text{Self}}=\lambda_{\text{Other}}$ denotes that idiocentric and allocentric updating are constrained to be equal; λ Unrestricted denotes that idiocentric (λ_{Self}) and allocentric (λ_{Other}) updating can differ. Learning rates for Self and Other were either constrained to be equal ($\alpha_{\text{Self}}=\alpha_{\text{Other}}$) or could differ (α Unrestricted). A model with unrestricted λ , and equal learning rates (as shown in Eq. (1)) provided the best fit, indicated by an asterisk. **b** Frequency of best-fitting models across participants. A model with unrestricted λ , and equal learning rates was the most frequent model, with a protected exceedance probability of 0.802. **c** Reported beliefs and model predictions from the best fitting model are shown for two sample participants. Self and Other probe trials are interleaved in the order presented to participants.

and Other were constrained to be equal outperformed those in which learning rates could differ by agent.

In fixed-effects model comparison, a model with Unrestricted λ , and equal learning rates for Self and Other (as shown in Eq. (1)) provided the best overall fit (fixed-effect Bayes factor for $[\lambda_{\text{Self}} \neq \lambda_{\text{Other}}, \alpha_{\text{Self}} = \alpha_{\text{Other}}]$ vs. $[\lambda_{\text{Self}} = \lambda_{\text{Other}}, \alpha_{\text{Self}} = \alpha_{\text{Other}}] = 29.5$). This model was also the most frequent in a random-effects model

comparison (best-fitting model in 39% of participants, protected exceedance probability 0.802; Fig. 3b). As shown in Fig. 3c, the model estimated participants' choices well.

Notably, although a model with independent λ_{Self} and λ_{Other} provided the best overall fit, some participants' responses were more parsimoniously captured with either a single leakage parameter or no leakage. Since the models are nested, these

individual differences ought to be reflected in the fitted parameters of the more general model. This was indeed the case. For example, participants for whom a model with no leakage was favoured had λ^{Self} estimates closer to zero (mean $\lambda^{\text{Self}} = 0.0003$) than those for whom models with leakage were favoured (mean $\lambda^{\text{Self}} = 0.07$). When comparing parameters between BPD and control groups, we therefore used the more general model.

Parameter recovery

Correlations amongst the fitted parameters in the participant sample are shown in Supporting Material in Fig. S1; parameter recovery for the best-fitting model is shown in Fig. S2. λ^{Self} and λ^{Other} were negatively correlated across participants (Pearson $r = -0.19$). Notably, however, in parameter recovery these two parameters were separately recoverable, i.e., fitted λ^{Self} was uncorrelated with generative λ^{Other} . Leakage parameters were also recovered separately from memory decay (δ), learning rate (α) or choice noise (τ) parameters (Fig. S1a).

Accuracy for the recovery of λ^{Self} and λ^{Other} was relatively modest ($r = 0.63$ and 0.53 respectively). This was partly attributable to including high settings of choice noise in the recovery analysis. When the analysis was restricted to the lowest quartile of generative τ , recoverability of λ^{Self} and λ^{Other} improved ($r = 0.85$ and 0.82 , respectively; Fig. S2b).

Taken together, these findings indicate that our task design can robustly distinguish directions of leakage. This conclusion, from a model-based analysis, is also consistent with our 'model-free' analyses, which show how estimates of leakage can be derived from first principles, based on between-agent outcome-belief correlations. It appears however that leakage is less reliably estimable at higher settings of choice noise.

Group comparison of agent-invariant updating

We found that λ^{Self} was significantly higher amongst BPD participants than amongst general population controls (95% CI of difference [0.06 0.24]; two-sample, $t(110) = 3.19$, $p = 0.002$; Cohen's $d = 0.64$), with no significant difference in λ^{Other} between the two groups (95% CI [-0.23 0.03]; $t(110) = -1.50$, $p = 0.137$). In essence, we found a significant effect of BPD diagnosis on leakage from Self-relevant learning signals to estimates of the Other's belief, captured by λ^{Self} (Fig. 4a). Mean λ^{Self} was greater than zero in BPD participants (mean 0.14; 95% CI [0.04 0.24], one-sample $t(37) = 2.75$, $p = 0.009$), and not significantly different from zero in control participants (mean -0.01 , 95% CI [-0.06 0.03], $t(73) = -0.55$, $p > 0.25$). Thus, BPD participants generalised their own belief updates to others (positive λ^{Self}), while control participants showed no significant degree of generalisation across agents (see also Fig. 4a). This finding suggests that control participants, relative to BPD participants, were more capable in mentalising differences between Self and Other within the task.

Correlation between agent-invariant updating and BPD symptoms

We next tested for a relationship between agent-invariant updating and BPD symptom severity (measured by the BSL-95) across participants. Notably, although mean BSL-95 scores were significantly higher amongst BPD participants than amongst controls from the general population (mean BPD = 198.1, mean controls = 153.5, $t(112) = 4.5$, $p < 0.001$), substantial variability was observed in both groups (Fig. 4b). Perhaps surprisingly, some of the lowest scores were observed amongst the BPD group. One explanation for this is the possibility this reflects measurement error (variable responding) in the BPD group, since all BPD participants scored at least nine points on the clinician-rated Zanarini Rating Scale, and these clinician-rated scores were positively correlated with BSL-95 scores ($r = 0.45$, $p = 0.005$).

λ^{Self} showed a significant positive relationship with BPD symptoms ($\beta_{\text{BSL95}} = 0.0009$, $t(109) = 2.20$, $p = 0.0300$; Fig. 4b). Thus, idiocentric updating was positively related to BPD symptoms across participant groups. We also observed a significant interaction between BPD symptoms and participant group, with a larger effect of symptoms amongst the BPD group than amongst the control group ($\beta_{\text{BSL95} \times \text{group}} = -0.0007$, $t(109) = -2.58$, $p = 0.011$). A possible explanation for this interaction is that more substantial impairments in mentalising (corresponding to higher settings of λ^{Self}) translate into clinically meaningful symptoms, while subtle mentalising difficulties may remain asymptomatic.

A relationship between λ^{Self} and BPD symptoms remained significant after controlling statistically for symptoms of depression, mania, impulsivity, age, gender, ethnicity, education and employment status, by including these as covariates ($\beta_{\text{BSL-95}} = 0.0014$; $t(82) = 2.73$, $p = 0.0078$; see Supporting Table S3 for full regression results). In this analysis, no other symptom scores were significantly related to λ^{Self} .

In a separate regression, we found no significant relationship between λ^{Other} and BPD symptoms ($\beta_{\text{BSL-95}} = 0.0001$, $t(109) = 0.25$, $p > 0.25$). This negative result is consistent with our finding, in behavioural analyses, of no significant difference between BPD and control groups in allocentric updating, as measured by an O^{Other} , B^{Self} correlation. Fixed effect pairwise correlations between all model parameters and symptom scores are shown in Fig. 4c. We also observed a positive correlation between BPD symptoms and the memory decay parameter, δ . Notably however δ was separately recoverable from λ^{Self} in a parameter recovery analysis (Fig. S1).

Sensitivity analyses

To ensure that the reported effects are not solely attributable to participants whose task performance did not differ from chance, we repeated the key analyses, excluding participants ($N = 17$) for whom we could not reject a null hypothesis of chance responding. Amongst the remaining 95 participants, λ^{Self} remained significantly higher amongst BPD participants than amongst general population controls (95% CI of difference [0.04 0.24]; two-sample, $t(93) = 2.78$, $p = 0.007$; Cohen's $d = 0.55$). Furthermore, a relationship between λ^{Self} and BPD symptoms remained significant, even when controlling statistically for demographic factors and other symptom scores ($\beta_{\text{BSL-95}} = 0.001$, 95% CI [0.0002 0.002]; $t(67) = 2.34$, $p = 0.022$).

DISCUSSION

In this study we formalise self-other mergence in terms of generalisation, or 'leakage', across agent-representations during learning. Under *idiocentric* updating, one's own prediction errors generalise to others, implying that a belief that what is surprising for the self is also surprising for others. By contrast, under *allocentric* updating, others' prediction errors are used to update one's own belief about the world, implying that others' surprise is treated as a self-relevant learning signal. To revisit our introductory example, a theatregoer with extreme idiocentric updating would expect Romeo to share their knowledge that Juliet was feigning her death. By contrast, a theatregoer with allocentric updating, upon seeing Romeo's response, might believe that Juliet was, in fact, dead. Either form of reasoning would render the events of the play highly confusing.

By analysing responses on a probabilistic false belief task (p-FBT) [37, 38], we show that participants with diagnoses of Borderline Personality Disorder ($N = 38$) exhibited higher degrees of idiocentric updating compared with a group of control participants sampled from the general population ($N = 74$). That is, people with BPD tended to generalise their own learning to others. Furthermore, the degree of idiocentric updating correlated with self-reported BPD symptoms across participants, even when controlling statistically for demographic factors, impulsivity, and mood disorder.

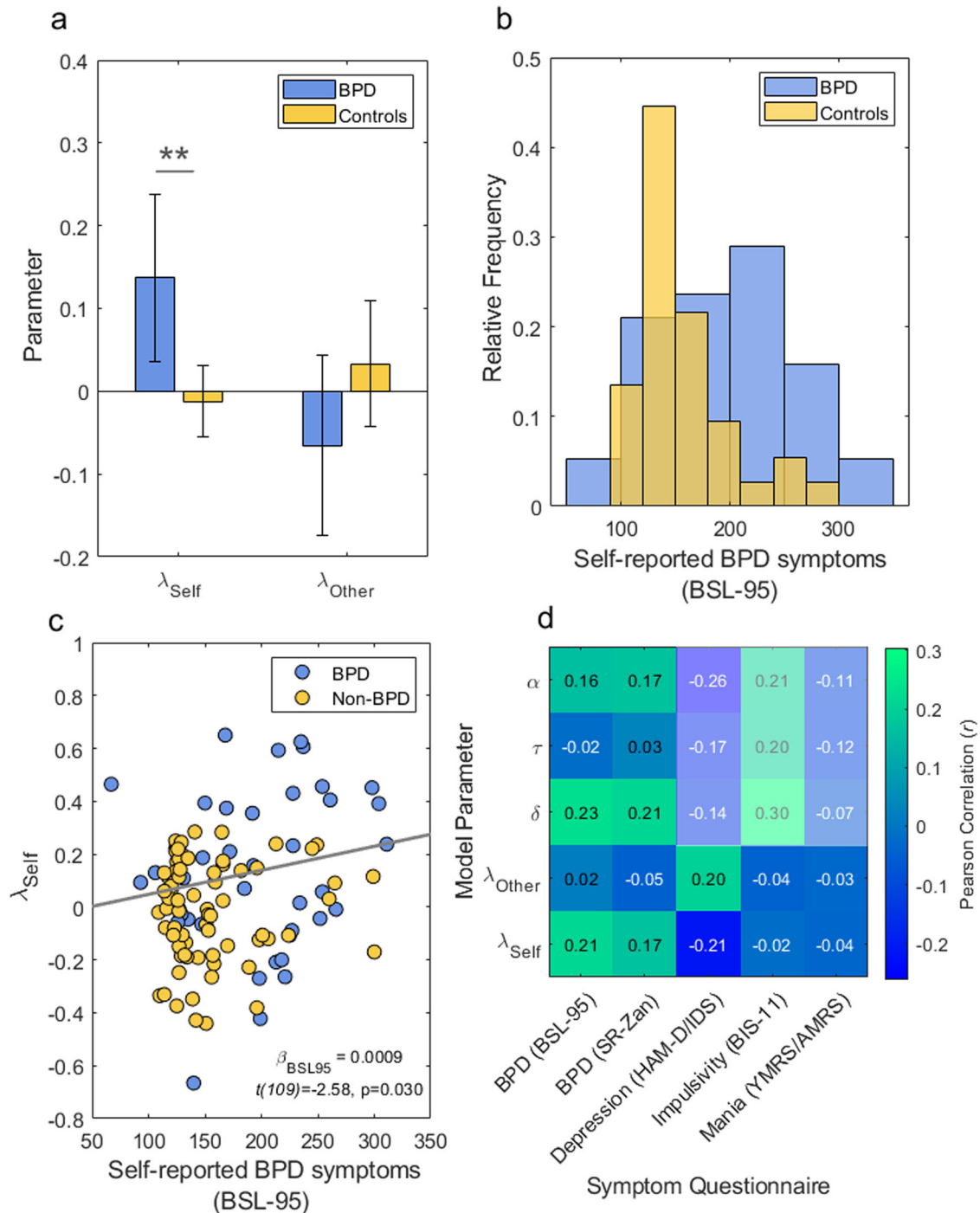


Fig. 4 BPD is associated with greater idiocentric updating. **a** Mean estimates of leakage parameters in BPD and Non-BPD groups. Idiocentric updating, captured by λ_{Self} , was higher amongst BPD participants than amongst Non-BPD controls (two-sample, $t(110) = 3.19, p = 0.002$; Cohen's $d = 0.64$); there was no significant difference in λ_{Other} between the two groups ($t(110) = -1.50, p = 0.137$). **b** Distribution of self-reported BPD symptoms (BSL-95 scores) across BPD and control groups. **c** Idiocentric updating correlated with BPD symptoms across all participants (solid grey line indicates the main effect of BSL-95 scores). **d** Pearson correlations between all model parameters and symptom scores across all participants ($N = 112$). The translucent white box indicates combinations of symptom scores and parameters that were not of prior interest. In multiple regression on λ_{Self} with BPD symptoms (BSL-95), depression, mania, impulsivity, age, gender, ethnicity, education and employment status as predictor variables, BPD symptoms emerged as the only significant predictor ($\beta_{BSL-95} = 0.0014; t(82) = 2.73, p = 0.0078$). BSL-95: 95 item Borderline Symptom List; SR-Zan: Self-report Zanzarini scale; HAM-D: Hamilton Depression Rating Scale; IDS: Inventory of Depressive Symptomatology; BIS-11: Barratt Impulsiveness Scale; YMRS: Young Mania Rating Scale; AMRS: Altman Mania Rating Scale.

Clinical considerations and potential application of the p-FBT Clinically, mentalising in BPD deteriorates under conditions of heightened affect, for example following the withdrawal of an

attachment figure, creating a vicious cycle where emotional and relational instability potentiate each other [49, 50]. A pertinent question therefore concerns whether idiocentric updating, as

measured by the p-FBT, varies dynamically as a function of affective or relational disturbance. Studies using two-person economic games to quantify mentalising support this idea. For instance, computational analyses of such games suggest that representations of the other player's strategy made by BPD participants tend to collapse when cooperation breaks down [51–53].

Agent-specific updating also suggests a novel outcome-measure for relational psychotherapy, the goal of which is to help individuals form more veridical representations of self and other. This finds support in evidence agent-specific neural representations are modifiable through training [38], and that, in non-human primates, neurons organise over time into agent-specific units for simulating the decisions of others [54]. Notably, by contrast to existing measures of theory-of-mind, such as conventional false-belief tasks [24, 29] or the *faux pas* test [55], the p-FBT entails serial belief updates, with changing underlying probabilities. This renders the task suitable for repeated measurement in a clinical context since it is not possible for the subject to simply learn (or 'second-guess') the correct answers.

Mentalising in BPD: a general deficit in meta-representation?

Notably, the p-FBT differs from many existing paradigms used to investigate mentalising, which examine inference about another person's feelings and/or intentions. These studies find that, relative to non-BPD controls, people with BPD tend to describe others' mental states in ways that are either lacking in nuance [56–58] or excessively precise [59–61]. Also, BPD participants have difficulty tracking changes in others' reliability [62, 63], and in re-establishing cooperation following ruptures of trust [51, 52], while also being slow to revise negative appraisals of others [64, 65]. This echoes the developmental histories of people with BPD, where experiences of neglect and abuse are more frequent than in the general population [66, 67].

By contrast to the aforementioned studies, in the p-FBT, the participant receives no information about the other person's actions or expressions. Indeed, the p-FBT as implemented here is 'social' only in its framing, and requires no inference about strictly interpersonal statistics (such as trustworthiness or generosity). Thus, the behavioural effect we observe is unlikely to depend on a participant's inference regarding the other's social characteristics.

Our finding that BPD participants show disrupted mentalising on the p-FBT might therefore be taken to indicate a general difficulty in representing and updating multiple points of view, known as 'meta-representation' [68]. This domain-general impairment may, in turn, give rise to more strictly interpersonal aspects of BPD. Imagine, for example, that an employee of a company is paid proportional to the company's performance; the employee's impression of business performance could differ from that of their manager. Difficulty representing this difference in perspective could encourage false social inference; for instance, a subject who believed the business was performing well might, following a decrease in their wages, conclude that the manager was deliberately exploiting them by cutting their pay. Such disrupted inferential processes may underpin heightened interpersonal sensitivity observed in BPD.

Difficulty adopting alternative perspectives might also entail challenges in imagining *oneself* in future situations. Indeed, computations necessary for understanding others' mental states are known to overlap with those involved in representing one's own [68–70]. Thus, impaired perspective-taking may generate impulsive, as well as interpersonal, features of the BPD syndrome.

A possibility that mentalising deficits observed in BPD are not specific to social cognition begs the question of to what extent the findings reported here are specific to BPD symptomatology. Deficits in theory-of-mind are well documented in other mental disorders, including autistic spectrum disorder (ASD) [71] and schizophrenia [72], whose clinical features partly differ from BPD.

Echoing the considerations above, it has been proposed that theory-of-mind deficits in autism lie downstream of a more general impairment in integrating sensory data to form abstract conceptual representations [73]. Social impairments are thought to ensue because reasoning in social situations relies heavily on prior concepts [74]. It appears possible that a similar deficit in information integration might also underpin BPD. This is consistent with a recent view of BPD as a social communication disorder [75], and with conceptual parallels between the symptoms of BPD and ASD [76]. Such a proposal does not obviate the role of childhood trauma in the aetiology of BPD. Rather, a neurodevelopmental predisposition towards social communication difficulties would be expected to interact with the childhood environment to determine the expression of BPD in adulthood.

Limitations of the current study

Despite the advantages outlined above, the current study has limitations. Firstly, we could reject a null hypothesis of random responding in only 85% of participants. This relatively high rate of random responding does not impact our conclusions, since the key findings remained unchanged whether or not we included random responders in the analysis. Nevertheless, making the task easier (for instance, by increasing stimulus presentation time) would reduce the rates of unusable data. Secondly, a potential limitation of the study itself is that, since data collection was interrupted by the Covid-19 pandemic, a small number of BPD participants (9 out of 38), and all control participants, completed the study online. Mitigating this concern, participants completing the study online met with an experimenter through a video call, to ensure controlled experimental conditions.

Summary

In summary, our findings are in keeping with a clinical notion of 'psychic equivalence' in BPD, wherein mental states are equated with external reality [2, 18, 19]. Our model suggests a candidate computational process underpinning 'psychic equivalence', namely an over-generalisation of self-relevant learning signals. Following existing psychological theories, it appears plausible that the resulting fusion of self- and other-representations contributes to the relational and affective instability observed in individuals with BPD [10–12, 14, 22, 23]. Future work might focus on evaluating the computational model described here as a transdiagnostic measure of mentalising ability in clinical settings.

DATA AVAILABILITY

The patient data that support the findings of this study are available from JEG upon reasonable request. Patient data are not publicly available due to confidentiality issues. Behavioural data from control participants and computer codes necessary to fit the learning models described here are publicly available at <http://github.com/gilestory/FalseBeliefTask>.

REFERENCES

1. Ereira SPA. A neurocomputational account of self-other distinction: from cell to society. Doctoral dissertation, University College London; 2019.
2. Fonagy P, Target M. Playing with reality: I. Theory of mind and the normal development of psychic reality. *Int J Psychanal* 1996;77:217–33.
3. Frith CD, Frith U. Interacting minds—a biological basis. *Science*. 1999;286:1692–5.
4. Fonagy P, Target M. Mentalization and the changing aims of child psychoanalysis. *Psychoanal Dialogues*. 1998;8:87–114.
5. Wellman HM. Making minds: how theory of mind develops. Oxford University Press, Oxford; 2014.
6. Rakoczy H. Foundations of theory of mind and its development in early childhood. *Nat Rev Psychol* 2022;1:223–35.
7. Stojnić G, Gandhi K, Yasuda S, Lake BM, Dillon MR. Commonsense psychology in human infants and machines. *Cognition*. 2023;235:105406.
8. Silston B, Bassett DS, Mobbs D. How dynamic brain networks tune social behavior in real time. *Curr Dir Psychol Sci*. 2018;27:413–21.

9. Schurz M, Radua J, Tholen MG, Maliske L, Margulies DS, Mars RB, et al. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychol Bull* 2021;147:293.
10. Fonagy P, Target M. Playing with reality: III. The persistence of dual psychic reality in borderline patients. *Int J Psychanal*. 2000;81:853–73.
11. Fonagy P, Bateman A. The development of borderline personality disorder—a mentalizing model. *J Pers Disord*. 2008;22:4–21.
12. Fonagy P, Luyten P. A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Dev Psychopathol*. 2009;21:1355–81.
13. De Meulemeester C, Lowyck B, Luyten P. The role of impairments in self–other distinction in borderline personality disorder: a narrative review of recent evidence. *Neurosci Biobehav R*. 2021;127:242–54.
14. Bender DS, Skodol AE. Borderline personality as a self–other representational disturbance. *J Pers Disord*. 2007;21:500–17.
15. Blatt SJ, Auerbach JS, Levy KN. Mental representations in personality development, psychopathology, and the therapeutic process. *Rev Gen Psychol*. 1997;1:351–74.
16. Fuchs T. Fragmented selves: temporality and identity in borderline personality disorder. *Psychopathology*. 2007;40:379–87.
17. Target M, Fonagy P. Playing with reality: II. The development of psychic reality from a theoretical perspective. *Int J Psychanal*. 1996;77:459–79.
18. Bateman AW, Fonagy P. Mentalization-based treatment of BPD. *J Pers Disord*. 2004;18:36–51.
19. Bateman AW. Personality disorders and pathology: integrating clinical assessment and practice in the DSM-5 and ICD-11 era. American Psychological Association, Washington; 2022. pp. 237–58.
20. Allen JG, Fonagy P, Bateman AW. Mentalizing in clinical practice. American Psychiatric Association Publishing, Washington; 2008.
21. Zanarini MC, Frankenburg FR. The essential nature of borderline psychopathology. *J Pers Disord*. 2007;21:518–35.
22. Rey H, Magagna JE. Universals of psychoanalysis in the treatment of psychotic and borderline states: factors of space–time and language. Free Association Books; 1994.
23. Glasser M. Problems in the psychoanalysis of certain narcissistic disorders. *Int J Psychanal*. 1992;73:493.
24. Hillmann K, Neukel C, Krauch M, Spohn A, Schnell K, Herpertz SC, et al. Cognitive and affective theory of mind in female patients with borderline personality disorder. *J Pers Disord*. 2021;35:672–90.
25. Bora E. A meta-analysis of theory of mind and ‘mentalization’ in borderline personality disorder: a true neuro-social-cognitive or meta-social-cognitive impairment? *Psychol Med*. 2021;51:2541–51.
26. Németh N, Matrai P, Hegyi P, Czéh B, Czopf L, Hussain A, et al. Theory of mind disturbances in borderline personality disorder: a meta-analysis. *Psychiat Res*. 2018;270:143–53.
27. Wimmer H, Perner J. Beliefs about beliefs: rRepresentation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*. 1983;13:103–28.
28. Perner J, Leekam SR, Wimmer H. Three year olds’ difficulty with false belief: the case for a conceptual deficit. *Br J Dev Psychol*. 1987;5:125–37.
29. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a “theory of mind”? *Cognition*. 1985;21:37–46.
30. New AS, Rot M, Ripoll LH, Perez-Rodriguez MM, Lazarus S, Zipursky E, et al. Empathy and alexithymia in borderline personality disorder: clinical and laboratory measures. *J Pers Disord*. 2012;26:660–75.
31. Haas BW, Miller JD. Borderline personality traits and brain activity during emotional perspective taking. *Personal Disord*. 2015;6:315.
32. Colle L, Gabbatore I, Riberi E, Borroz E, Bosco FM, Keller R. Mindreading abilities and borderline personality disorder: a comprehensive assessment using the Theory of Mind Assessment Scale. *Psychiatr Res*. 2019;272:609–17.
33. Harari H, Shamay-Tsoory SG, Ravid M, Levkovitz Y. Double dissociation between cognitive and affective empathy in borderline personality disorder. *Psychiatr Res*. 2010;175:277–9.
34. Baez S, Marengo J, Perez A, Huepe D, Giralto Font F, et al. Theory of mind and its relationship with executive functions and emotion recognition in borderline personality disorder. *J Neuropsychol* 2015;9:203–18.
35. Petersen R, Brakoulias V, Langdon R. An experimental investigation of mentalization ability in borderline personality disorder. *Compr Psychol*. 2016;64:12–21.
36. Pluta A, Kulesza M, Grzegorzewski P, Kucharska K. Assessing advanced theory of mind and alexithymia in patients suffering from enduring borderline personality disorder. *Psychiatr Res*. 2018;261:436–41.
37. Ereira S, Dolan RJ, Kurth-Nelson Z. Agent-specific learning signals for self–other distinction during mentalizing. *PLoS Biol*. 2018;16:e2004752.
38. Ereira S, Hauser TU, Moran R, Story GW, Dolan RJ, Kurth-Nelson Z. Social training reconfigures prediction errors to shape Self–Other boundaries. *Nat Commun*. 2020;11:3030.
39. Bohus M, Limberger MF, Frank U, Chapman AL, Kuhler T, Stieglitz RD. Psychometric properties of the borderline symptom list (BSL). *Psychopathology*. 2007;40:126–32.
40. Grant JE, Valle S, Chesivoir E, Ehsan D, Chamberlain SR. A double-blind placebo-controlled study of brexpiprazole for the treatment of borderline personality disorder. *Br J Psychiatr*. 2022;220:58–63.
41. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. *Neuroimage*. 2009;46(4):1004–17.
42. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies—revisited. *Neuroimage*. 2014;84:971–85.
43. Zanarini MC, Weingeroff JL, Frankenburg FR, Fitzmaurice GM. Development of the self-report version of the Zanarini Rating Scale for Borderline Personality Disorder. *Personal Ment Health*. 2015;9:243–9.
44. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol*. 1995;51:768–74.
45. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56.
46. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatr*. 1978;133(5):429–35.
47. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The inventory of depressive symptomatology (IDS): psychometric properties. *Psychol Med*. 1996;26:477–86.
48. Altman EG, Hedeker D, Peterson JL, Davis JM. The Altman self-rating mania scale. *Biol Psychiatry*. 1997;42:948–55.
49. Fonagy P, Bateman AW. Adversity, attachment, and mentalizing. *Compr Psychiatry*. 2016;64:59–66.
50. Nolte T, Bolling DZ, Hudac CM, Fonagy P, Mayes L, Pelphrey KA. Brain mechanisms underlying the impact of attachment-related stress on social cognition. *Front Hum Neurosci*. 2013;7:816.
51. King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR. The rupture and repair of cooperation in borderline personality disorder. *Science*. 2008;321:806–10.
52. Hula A, Vilares I, Lohrenz T, Dayan P, Montague PR. A model of risk and mental state shifts during social interaction. *PLoS Comput Biol*. 2018;14:e1005935.
53. Sripada CS, Angstadt M, Banks S, Nathan PJ, Liberzon I, Phan KL. Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *Neuroreport*. 2009;20:984.
54. Grabenhorst F, Báez-Mendoza R, Genest W, Deco G, Schultz W. Primate amygdala neurons simulate decision processes of social partners. *Cell*. 2019;177:986–98.
55. Baron-Cohen S, O’Riordan M, Jones R, Stone V, Plaisted K. A new test of social sensitivity: detection of faux pas in normal children and children with Asperger syndrome. *J Autism Dev Disord*. 1999;29:407–18.
56. Preißler S, Dziobek I, Ritter K, Heekeren HR, Roepke S. Social cognition in borderline personality disorder: evidence for disturbed recognition of the emotions, thoughts, and intentions of others. *Front Behav Neurosci*. 2010;4:182.
57. Brüne M, Walden S, Edel M-A, Dimaggio G. Mentalization of complex emotions in borderline personality disorder: the impact of parenting and exposure to trauma on the performance in a novel cartoon-based task. *Compr Psychiatry*. 2016;64:29–37.
58. Ortega-Díaz E, García-Campos J, Moya-Martínez A, Ramírez-Cremades C, Rico-Gomis JM, Cuesta-Moreno C, et al. Theory of mind in borderline personality disorder: a possible endophenotypic factor? *Int J Environ Res Public Health* 2021;18:3193.
59. Sharp C, Ha C, Carbone C, Kim S, Perry K, Williams L, et al. Hypermentalizing in adolescent inpatients: treatment effects and association with borderline traits. *J Pers Disord*. 2013;27:3–18.
60. Andreou C, Kelm L, Bierbrodt J, Braun V, Lipp M, Yassari AH, et al. Factors contributing to social cognition impairment in borderline personality disorder and schizophrenia. *Psychiatry Res*. 2015;229:872–9.
61. Cortés-García L, Akça ÖF, Wall K, Sharp C. Exploring mentalizing in adolescents with anorexia nervosa and borderline personality disorder: a comparative study of psychiatric inpatients and healthy controls. *Scand J Child Adol Psychiatry Psychol*. 2021;9:16–26.
62. Henco L, Diaconescu AO, Lahnakoski JM, Brandi ML, Hörmann S, Hennings J, et al. Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS Comput Biol*. 2020;16:e1008162.
63. Fineberg SK, Leavitt J, Stahl DS, Kronemer S, Landry CD, Alexander-Bloch A, et al. Differential valuation and learning from social and nonsocial cues in borderline personality disorder. *Biol Psychiatry*. 2018;84:838–45.

64. Barnow S, Stopsack M, Grabe HJ, Meinke C, Spitzer C, Kronmüller K, et al. Interpersonal evaluation bias in borderline personality disorder. *Behav Res Ther*. 2009;47:359–65.
65. Siegel JZ, Curwell-Parry O, Pearce S, Saunders KEA, Crockett MJ. A computational phenotype of disrupted moral inference in borderline personality disorder. *Biol Psychiatry*. 2020;5:1134–41.
66. Reich RB, Vera SC, Marino MF, Levin A, Yung L, Frankenburg FR. Reported pathological childhood experiences associated with the development of borderline personality disorder. *Am J Psychiatry*. 1997;154:1101–106.
67. Nigg JT, Lohr NE, Westen D, Gold LJ, Silk KR. Malevolent object representations in borderline personality disorder and major depression. *J Abnorm Psychol*. 1992;101:61.
68. Carruthers P. How we know our own minds: the relationship between mind-reading and metacognition. *Behav Brain Sci*. 2009;32(2):121–38.
69. Heyes C. What's social about social learning? *J Comp Psychol*. 2012;126(2):193–202.
70. Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. Associative learning of social value. *Nature*. 2008;456(7219):245–9.
71. Andreou M, Skrimpa V. Theory of mind deficits and neurophysiological operations in autism spectrum disorders: a review. *Brain Sci*. 2020;10(6):393.
72. Thibaut E, Cellard C, Turcotte M, Achim AM. Functional impairments and theory of mind deficits in schizophrenia: a meta-analysis of the associations. *Schizophr Bull*. 2021;47(3):695–711.
73. Pellicano E, Burr D. When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trend Cogn Sci*. 2012;16(10):504–10.
74. Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, De-Wit L, et al. Precise minds in uncertain worlds: predictive coding in autism. *Psychol Rev*. 2014;121(4):649.
75. Luyten P, Campbell C, Fonagy P. Borderline personality disorder, complex trauma, and problems with self and identity: a social-communicative approach. *J Personal*. 2020;88(1):88–105.
76. Dell'Osso L, Cremone IM, Nardi B, Tognini V, Castellani L, Perrone P, et al. Comorbidity and overlaps between autism spectrum and borderline personality disorder: state of the art. *Brain Sci*. 2023;13(6):862.

ACKNOWLEDGEMENTS

The authors would like to thank the participants of this study for their time and contribution. This work was supported by the Wellcome Trust [Ray Dolan Investigator Award 098362/Z/12/Z] and the Max Planck Society. Data from this study were presented as a poster at the Annual Meeting of the Society of Biological Psychiatry, 2023.

AUTHOR CONTRIBUTIONS

SE and RD developed the conceptual basis for the study. SE, GWS, and RD designed the experiments. SE coded the experimental protocols. SV, JEG, SRC, GWS and SE were involved in collecting the data. GWS analysed the data and wrote the manuscript. RD and SE provided feedback and conceptual contributions to data analysis. All authors edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41398-024-03170-w>.

Correspondence and requests for materials should be addressed to Giles W. Story.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024