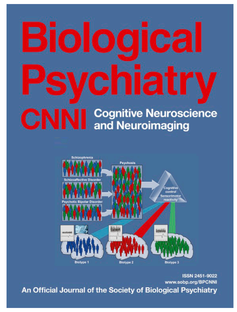
Journal Pre-proof



A computational phenotype of disrupted moral inference in Borderline Personality Disorder

Jenifer Z. Siegel, Ph.D., Owen Curwell-Parry, BMBCh, Steve Pearce, MBBS, Kate E.A. Saunders, BMBCh, Molly J. Crockett, Ph.D

PII: S2451-9022(20)30203-2

DOI: https://doi.org/10.1016/j.bpsc.2020.07.013

Reference: BPSC 639

To appear in: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Received Date: 1 June 2020

Accepted Date: 21 July 2020

Please cite this article as: Siegel J.Z., Curwell-Parry O., Pearce S., Saunders K.E.A. & Crockett M.J., A computational phenotype of disrupted moral inference in Borderline Personality Disorder, *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2020), doi: https://doi.org/10.1016/j.bpsc.2020.07.013.

Title. A computational phenotype of disrupted moral inference in Borderline Personality Disorder

Authors. Jenifer Z. Siegel, Ph.D.[1,2]; Owen Curwell-Parry, BMBCh[3,4]; Steve Pearce, MBBS[3,4]; Kate E.A. Saunders, BMBCh[3,4,*]; & Molly J. Crockett, Ph.D[2,*].

*equal contribution

[1]Department of Experimental Psychology, University of Oxford

[2]Department of Psychology, Yale University, 2 Hillhouse Ave, New Haven, CT 06511

[3]Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, OX3 7JX

[4]Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, OX3 7JX

Corresponding author. Dr. Siegel

Mailing: Jenifer Siegel, Yale University, 10 Hillhouse Ave, New Haven, CT 06511
Phone: 475-655-9612
Email: jenifersiegel@gmail.com

Short running title. Moral inference in Borderline Personality Disorder

Keywords. Borderline Personality Disorder, Computational psychiatry, Bayesian inference, Belief updating, Moral impression formation, Social behavior

Abstract

Background Borderline Personality Disorder (BPD) is a serious mental disorder characterized by marked interpersonal disturbances, including difficulties trusting others and volatile impressions of others' moral character, often resulting in premature relationship termination. We tested a hypothesis that moral character inference is disrupted in BPD and sensitive to Democratic Therapeutic Community (DTC) treatment.

Methods BPD participants (20 treated and 23 DTC-treated) and non-BPD control participants (N=106) completed a moral inference task where they predicted the decisions of two agents with distinct moral preferences: the "bad" agent was more willing to harm others for money than the "good" agent. Periodically, participants rated their subjective impressions of the agent's moral character, and the certainty of those impressions. We fit a hierarchical Bayesian learning model to participants' trial-wise predictions to describe how beliefs about the morality of the agents were updated by new information.

Results The computational mechanisms of moral inference differed for untreated BPD patients relative to matched non-BPD control participants and DTC-treated BPD patients. In BPD patients, beliefs about harmful agents were more certain and less amenable to updating relative to both non-BPD control participants and DTC-treated participants.

Conclusions The findings suggest that DTC may help the maintenance of social relationships in BPD by increasing patients' openness to learning about adverse interaction partners. The results provide mechanistic insights into social deficits in BPD and demonstrate the potential for combining objective behavioral paradigms with computational modelling as a tool for assessing BPD pathology and treatment outcomes.

INTRODUCTION

Borderline Personality Disorder (BPD) is a serious mental disorder affecting up to 5.9% of the general population (1). Marked disturbances in interpersonal relationships constitute one of the core symptom domains of BPD, including difficulties with trust and forgiveness often resulting in premature relationship termination (2–4). Difficulties related to interpersonal relationships contribute to substantial economic and societal costs including high rates of suicide and intensive use of high-cost medical care (5–8). Longitudinal studies indicate that symptoms related to interpersonal relationships are among the hardest to treat; serious social deficits often persist even after years of rigorous and resource exhaustive treatment (9–12). Research identifying the mechanisms of impaired social functioning in BPD is therefore paramount for relieving interpersonal and societal burdens.

Several possible explanations have been proposed for why patients with BPD exhibit a poor ability to maintain interpersonal relationships. For instance, building and maintaining successful social relationships depends on the ability to build accurate representations of others' mental states (e.g., intentions, beliefs, desires), however research suggest that patients with BPD may be limited in their ability to accurately perceive social signals and model the intentions of others (2). Notable, adaptive social functioning also depends on the ability to continuously update representations of others through social learning (13). A growing body of theoretical and empirical work suggests that impaired social learning plays an important role in interpersonal disturbances in BPD, including difficulties trusting others (3,14,15). Here, we consider one aspect of social learning that is especially relevant to forming and maintaining relationships: inferring others' moral character (16,17), i.e., whether they are helpful and trustworthy, or harmful and untrustworthy.

We introduce a novel computational assay of moral inference to investigate how patients with BPD form beliefs about the moral character of others and incorporate new information into existing beliefs. Previous research using these methods indicates that healthy adults hold more uncertain and less rigid beliefs when inferring a “bad” moral character relative to a “good” moral character (17,18). This work implemented a Bayesian inference framework where beliefs are updated in proportion to their uncertainty (19), such that more uncertain beliefs are updated more rapidly. Consequently, more uncertain negative beliefs about others’ morality enables those beliefs to be rapidly updated from new information, which is hypothesized to reflect an adaptive mechanism for sustaining relationships when others sometimes behave badly. Thus, holding negative moral beliefs with some degree of uncertainty may be an important aspect of healthy social functioning. Given that individuals with BPD often hold grudges and have difficulty forgiving others (4,20), we tested a hypothesis that relative to control non-BPD participants, BPD patients have more certain and rigid beliefs about harmful agents and therefore lack this adaptive mechanism for forgiveness that may help sustain relationships.

Understanding the mechanisms underlying interpersonal problems in BPD is essential for developing and assessing effective treatments. Democratic Therapeutic Community (DTC) treatment is one of the most widespread psychosocial treatments for BPD in the UK with a strong focus on developing cooperative strategies to help patients effectively navigate their social environments (21), and has been associated with improvements in social functioning at least 24 months following treatment (22), including more pleasant social relations (23). While DTC aims to help patients learn new strategies for adaptive social functioning, it is unknown how the effects of treatment manifest at the cognitive level. Understanding the cognitive channels through which DTC operates may ultimately help identify which patients may benefit the most from such treatment. To shed light on this question, the present research therefore assessed moral inference in a group of DTC-treated BPD participants compared to a group of untreated BPD participants.

METHODS AND MATERIALS

Participants

Non-BPD group. The online crowdsourcing platform Prolific (www.prolific.ac) enabled us to collect a sample of adult participants precisely matched to our patient population who would not qualify for a diagnosis of BPD. This method has the potential to improve the validity and generalizability of research by enabling efficient and low-cost recruitment of comparison groups for unique samples who may come from specific environments (24). Previous research has established that a diverse set of cognitive tasks (such as the Stroop, Flanker, and category learning) show similar results in the lab and online (25). Subjects recruited through online platforms are at least as attentive (26) and consistent (27) in their task performance as participants recruited through college subject pools. Furthermore, a recent study showed that participants recruited through the platform used in the present research, Prolific, produced data quality that was higher than comparable online crowdsourcing platforms as well as a university subject pool (28). We aimed to recruit five healthy adults who matched each BPD patient in sex, age (+/- 4 years) and education. We ensured that matched participants received the same variant of the moral inference task as their patient counterpart (i.e., same sequence of trials).

Non-BPD participants provided written informed consent after receiving a complete description of the study and were compensated for their time. The Yale University Human Investigation Committee approved the procedures (#2000022385). Participants completed the study on the web application framework, Heroku, and were subsequently directed to a Qualtrics survey to complete additional questionnaires to assess clinically relevant personality traits. Previous work has demonstrated that the moral inference task yields comparable results in lab and online settings (17). Non-BPD participants completed the McLean Screening Inventory for BPD (MSI, see eMethods in the **Supplement**) and were excluded from the analysis if they showed clinically relevant BPD symptoms (MSI score > 6). The final sample of non-BPD participants included 106 adults who scored lower than 7 on the MSI.

BPD group. Participants were treatment-seeking individuals with a primary diagnosis of BPD recruited from an outpatient population. The Structured Clinical Interview for Axis II disorders (SCID-II, see eMethods in **Supplement**) was administered by trained clinicians to establish BPD diagnosis. Inclusion criteria were: diagnosis of BPD, aged between 18 and 65, not currently being treated in group therapy, no current drug or alcohol dependence, and no psychiatric hospital admission in the preceding month. Individuals were excluded if they had a previous or current neurological condition, were unable to provide informed consent, were pregnant or breastfeeding, or met criteria for an Axis I illness (e.g., anxiety, mood, eating disorders). Nine participants were taking antidepressant or antipsychotic medication or both at the time of participation. The final sample included 20 participants with BPD.

DTC group. Participants with a primary diagnosis of BPD who completed DTC treatment (22) within three years prior to recruitment were recruited from the Oxfordshire and Buckinghamshire Complex Needs Service database. As part of the program anyone who is finding DTC unhelpful or is not deemed to be progressing their therapy would leave the program by mutual consent. Eligible participants were contacted by post and sent a copy of the information sheet along with an invitation to participate in the study. The SCID-II was administered to interested individuals by trained clinicians to establish BPD diagnosis. Inclusion criteria were: diagnosis of BPD, aged between 18 and 65, completed DTC at the Oxfordshire and Buckinghamshire Complex Needs Service (22) within the past three years, and no current drug or alcohol dependence. Individuals were excluded on the same basis as participants in the untreated BPD group. Eleven participants were taking antidepressant or antipsychotic medication or both at the time of participation. The final sample included 23 participants with BPD who had completed DTC treatment.

Behavioral testing of BPD participants (untreated BPD and DTC-treated groups) took place at the University of Oxford Department of Psychiatry. We used the Borderline Evaluation of Severity over Time (BEST) scale to assess the severity of BPD symptomology in participants with BPD at the time of participation (eMethods in **Supplement**). Participants provided written informed consent after receiving a complete description of the study and were compensated for their time. The study was approved by the local National Health Service ethics committee in Oxford, ethics number 14/SC/1430.

Moral Inference Task

In the moral inference task (17), participants predicted and observed the choices of two agents (called “Decider A” and “Decider B”) who repeatedly decided whether to inflict painful electric shocks on a victim in exchange for various amounts of money (**Figure 1a**). The two agents differed substantially in their moral preferences: the “good” agent required more compensation to inflict pain on others than the “bad” agent (**Figure 1b**). Periodically, participants rated their subjective impressions of the agent’s morality (from 0 = *nasty* to 100 = *nice*), and the certainty of those impressions (from 0 = *very uncertain* to 1 = *very certain*). Before observing any of the agent’s choices, participants additionally indicated how nasty or nice they *expected* the agent would be and how certain they were. This provided an indication of participants' prior expectations about people's moral character in general and their confidence in those prior expectations. We confirmed that the groups were equally motivated to learn about the agents and predict their decisions (see eResults in **Supplement**).

Figure 1 Moral Inference Task. (A) *Schematic representation of the moral inference task. Participants predicted sequences of choices for two agents (Decider A and Decider B). On each trial the agent chose between two options: more shocks inflicted on another person in exchange for more money, or fewer shocks in exchange for less money. After making each prediction, participants observed the agent's actual choice and received feedback indicating whether their prediction was correct or incorrect. Every third trial participants rated their subjective impression about the agent's moral character (ranging from nasty to nice) and how certain they were about their impression. (**B**) Heatmaps summarize the good and bad agents' probabilities of choosing the more profitable and harmful option as a function of the amount of money gained and number of shocks inflicted.*

Computational modelling

We fit a generative Bayesian reinforcement learning model (17–19,29) to participants' trial-by-trial predictions. The model identified participant-specific parameters to describe how each participant updated their beliefs about the morality of the agents, as described in (17). In the model, beliefs about an agent's moral preference (i.e., their exchange rate between money and shocks) are updated from new information with dynamic learning rates. Learning rates capture the weight participants place on new information over prior beliefs when updating beliefs on the current trial. When prior beliefs are less precise, learning rates are higher, such that less precise beliefs are more heavily updated from new information. Random-effects Bayesian Model Selection indicated our model with a dynamic learning rate was preferred over: (a) a model where beliefs were updated by new information with a fixed learning rate, and (b) a model where beliefs were updated by new information with separate fixed learning rates for positive (helpful) and negative (harmful) information (see eResults in the **Supplement**). Additionally, the proportion of participants whose data was best explained by our model with a dynamic learning rate did not significantly differ across BPD, non-BPD, and DTC groups ($\chi^2 = 3.044$, $p = 0.218$; see eResults in the **Supplement**).

Analysis

We used robust linear regression models with bisquare weighting functions to analyze standardized learning rates, subjective character impression ratings, and certainty ratings (using the RobustOpts setting in the fitlm function in Matlab, Mathworks). Certainty ratings were reverse scored such that higher values indicated greater uncertainty in subjective impressions of the agents' moral character. Because learning rates and subjective ratings evolve over time, we initially considered whether groups differed as a function of time dynamics (i.e., trial number) and found no evidence to support this prediction. Consequently, regression models included the effects of agent (bad, good), group (BPD, non-BPD, DTC), and their interaction, controlling for trial number. Further analyses used two-sided nonparametric statistical tests that do not make any assumptions about the underlying distributions of variables (e.g., Wilcoxon rank-sum test).

RESULTS

An omnibus test for group X agent interactions, where group was coded as a dummy variable (with untreated BPD as the reference group), found significant differences in the effect of agent between groups on uncertainty ratings (non-BPD, $\beta = 0.264 \pm 0.080$, $t = 3.310$, $p < .001$; DTC, $\beta = 0.266 \pm 0.100$, $t = 2.665$, $p = .008$) and learning rates (non-BPD, $\beta = 0.113 \pm 0.025$, $t = 4.607$, $p < .001$; DTC, $\beta = 0.319 \pm 0.031$, $t = 10.355$, $p < .001$; see eResults in Supplement for full analyses). For clarity, here we first present comparisons between untreated BPD participants and non-BPD participants, followed by comparisons between untreated BPD and DTC-treated groups.

Moral inference in BPD

We analyzed data in the moral inference task for untreated BPD and non-BPD participants who were matched for sex, age, education, and self-report psychopathy, but significantly differed in levels of clinically relevant personality traits (**Error! Reference source not found.**).

We first inspected participants' subjective impressions of the agents' moral character, and their uncertainty about those impressions. While there were no differences between BPD and non-BPD participants in average character impressions (see eResults in **Supplement**), group differences emerged for the uncertainty ratings. Consistent with prior findings (17), participants overall held more uncertain impressions of the bad agent than the good agent (main effect of agent: β = 0.418±0.032, t = 13.099, p < .001), however this effect was substantially reduced in BPD participants (interaction between agent and group, β = -0.263±0.080, t = -3.284, p = .001; Figure 2a). Relative to non-BPD participants, BPD participants held less uncertain impressions of the bad agent (β = -0.162 ± 0.058, t=-2.805, p = .005), but were similarly uncertain about their impressions of the good agent (β = 0.098±0.055, t = 1.761, p = .078).

Figure 2 Negative beliefs are more certain and slower to update in untreated BPD participants relative to non-BPD control participants. *(**A**) Relative to non-BPD participants, BPD participants held less uncertain impressions of the bad agent. (**B**) BPD participants were slower to update beliefs about the bad agent following new information. Error bars represent 95% confidence intervals. a.u. = arbitrary units. **P<0.01, *P<0.05, n.s.t. = non-significant trend (P<0.1), where significance refer to the interaction between group and agent in our regression models.*

Learning rate data were consistent with the uncertainty rating data. Overall, participants updated beliefs faster for the bad agent than the good agent (main effect of agent, β =

0.323±0.017, t = -18.601, p < .001), however this effect was substantially smaller in BPD participants (interaction between agent and BPD group, β = -0.167±0.044, t = -3.827, p < .001; **Figure 2b**). Specifically, BPD participants were slower to update beliefs about the bad agent (β = -0.109±0.034, t = -3.222, p = .001) and faster to update beliefs about the good agent (β = 0.062±0.027, t = 2.287, p = .022) relative to non-BPD participants. The findings suggest that BPD is associated with more confident and less flexible beliefs about harmful agents, but less confident and more flexible beliefs about helpful agents. A supplementary analysis (using data across all BPD patient groups) revealed that BPD symptom severity moderated the observed effects, such that participants with more severe BPD symptoms expressed less uncertain impressions of the bad agent and more uncertain impressions of the good agent (see eResults in Supplement).

Participants with BPD indicated more pessimistic expectations before observing any of the agents' choices than non-BPD participants (Z = -2.491, p = .013), though BPD and non-BPD participants were similarly certain about their expectations (Z = -0.327, p = .743). Thus, a plausible explanation for the observed pattern of results is that the good agent violated BPD participants' expectations to a greater degree than the bad agent. Given our particular model, this could make beliefs about the good agent more amenable to Bayesian updating in BPD, by which belief updates are optimized to minimize surprise (19). Previous research indicates that healthy adults are able to override externally generated prior expectations and rapidly adjust their learning as a function of moral character information (17), prioritizing belief updating for putatively "bad" agents. We replicated this finding in the non-BPD participants (see eResults in **Supplement**). However, analyses suggested that unlike healthy adults, learning may be especially sensitive to prior expectations in BPD (see eResults in **Supplement**).

Moral inference in DTC-treated BPD participants

Next, we compared performance on the moral inference task for DTC-treated and untreated BPD participants who were matched for sex, age, education, self-report psychopathy, and clinically relevant personality traits (**Error! Reference source not found.**). We confirmed that the severity of BPD symptomology in DTC treated participants was significantly lower than untreated BPD participants (BEST, $Z = 3.690$, $p < .001$).

DTC-treated participants expressed more favorable impressions in general than untreated BPD participants (main effect of group, $\beta = 0.146 \pm 0.046$, $t = 3.197$, $p = .001$). This group difference appeared to be primarily driven by impressions of the good agent (interaction between agent and group, $\beta = -0.236 \pm 0.064$, $t = -3.668$, $p < .001$), such that the DTC-treated participants expressed more favorable impressions of the good agent, relative to untreated participants ($\beta = 0.151 \pm 0.043$, $t = 3.507$, $p < .001$). Group differences in impressions of the bad agent did not reach significance ($\beta = -0.090 \pm 0.048$, $t = -1.869$, $p = .062$).

Turning to the uncertainty of impressions and learning rates, we found that DTC-treated participants, relative to untreated participants, showed more uncertain impressions of the bad agent ($\beta = 0.188 \pm 0.067$, t = 2.802, $p = .005$; **Figure 3a**) and faster learning rates for the bad agent ($\beta = 0.543 \pm 0.040$, $t = 13.698$, $p < .001$; **Figure 3b**), as indicated by significant interactions between agent and group for both measures (uncertainty ratings: $\beta = 0.277 \pm 0.095$, $t = 2.904$, $p = .003$; learning rates: $\beta = 0.589 \pm 0.052$, $t = 11.588$, $p < .001$; see eResults in **Supplement** for full regression analyses). No group differences were observed on impression uncertainty or learning rates for the good agent (uncertainty: $\beta = -0.081 \pm 0.068$, t = -1.196, p = .232; learning rates: $\beta = -0.030 \pm 0.030$, $t = -0.989$, $p = .323$). Thus, DTC treatment was associated with increased uncertainty and more flexible beliefs about the bad agent, specifically.

Figure 3 Negative beliefs are more uncertain and faster to update in DTC-treated participants than untreated BPD participants. *(A) Relative to untreated BPD participants, DTC treatment was associated with more uncertain impressions of the bad agent. (B) DTC-treated participants were faster to update beliefs about the bad agent from new information than untreated BPD participants. Error bars represent 95% confidence intervals. A.u. = arbitrary units. ***P<0.001, **P<0.01, n.s. = not significant (P>0.1),* where significance refers to the interaction between group and agent in our regression model.

DTC-treated and untreated BPD participants had similar expectations about the agents' morality ($Z = 0.585$, $p = .559$) and were similarly certain about their expectations ($Z = 0.585$, $p = .559$). Negative expectations therefore do not account for the observed group differences in moral inference. For completeness, we investigated whether prior expectations covaried with the interaction between group and agent and report the results in the online Supplement. Overall, we found that even though DTC-treated and untreated BPD participants had similar moral expectations, the groups differed in how expectations subsequently shaped learning.

In the present study many participants were taking psychotropic medication at the time of participation. It is possible that group differences in pharmacological treatments drove increased flexibility and belief updating for the bad agent, rather than DTC treatment. However, we observed a similar interaction between agent and group on uncertainty ratings and learning rates when controlling for medication use (uncertainty: $\beta = 0.277 \pm 0.095$, $t = 2.898$, $p = .004$; learning rates: $\beta = 0.577 \pm 0.050$, $t = 11.441$, $p < .001$; see eResults in **Supplement** for full regression analyses).

DISCUSSION

Here we identify a computational phenotype that may characterize some aspects of BPD pathology and is sensitive to a common treatment. Unlike healthy adults, who maintain flexibility in their beliefs about potentially harmful social partners, BPD participants hold more certain negative beliefs about others and are slower to update those beliefs. DTC treatment was associated with more uncertain, flexible beliefs about putatively harmful social partners, suggesting that DTC may improve social interactions in BPD by increasing participants' openness to learning about partners who exhibited potentially threatening social interactions.

Cumulatively, our results could provide a computational framework for understanding seemingly paradoxical findings of both volatility and rigidity of social beliefs in BPD. Our observation of more rigid negative beliefs in BPD is consistent with past reports that BPD patients show slower learning rates in a task that requires learning about the probability of social and nonsocial cues, less conciliatory social behavior following a rupture of trust (2), and difficulty forgiving others(4). We also found some evidence that BPD participants hold less certain positive beliefs about others and are faster to update those beliefs. This finding is consistent with the ease patients have in terminating relationships as well as clinical observations that the patient can shift rapidly from a period of admiration to dislike in response to even minor slights (30).

In contrast to past work, by modelling social learning within a Bayesian framework we are able to consider another important aspect of healthy social cognition. In optimal Bayesian inference learning is intrinsically tied to prior expectations. Observations that are consistent with prior expectations help reinforce them, while those that are inconsistent may be used to update expectations. However, moral inference departs from Bayes optimality in an important way: healthy adults maintain more uncertain beliefs about the moral character of putatively bad agents even when observations are consistent with prior expectations (17). We hypothesize that humans have evolved to rapidly discount prior expectations to adapt learning according to moral information. This feature of healthy social cognition provides the flexibility to promptly update beliefs about bad agents when those beliefs turn out to be wrong, preserving social relationships in the wake of accidental harms.

One possibility is that BPD impacts cognitive processes important for the ability to adapt learning as a function of moral information. In turn, patients may rely heavily on pessimistic prior expectations born from adversity and volatility in their social environment (31–33). While the ability to rapidly discount externally generated prior expectations in moral inference may be advantageous in environments where social partners are consistently trustworthy, it can be costly when partners behave unpredictably. By shutting down the gateway for learning when behavior misaligns with antisocial expectations, rigidity then provides a protective mechanism that prevents responding to unreliable social cues. We found evidence consistent with the hypothesis that untreated BPD participants may be especially reliant on pessimistic expectations in moral inference (outlined in eResults in **Supplement**). However, more work is needed to assess whether abnormal moral inference in BPD can be explained by an increased tendency to rely on pessimistic prior expectations.

DTC offers a safe environment for patients with BPD to learn the skills necessary for successful social functioning and has shown promise in ameliorating social difficulties (22). Our findings suggest that DTC may positively impact social interactions by increasing patients' openness to learning about potentially threatening social interaction partners, allowing information to be integrated over longer timescales before establishing a negative evaluation. On the other hand, whether DTC impacts learning about positive social interaction partners, and the development of stable positive beliefs, remains uncertain. If mentalization-based therapies have an impact on epistemic trust, as recent models are proposing (14,34), it may be especially effective in addressing difficulties in establishing stable positive social beliefs in BPD. By applying and comparing this measure in alternative treatment groups we can better understand the mechanisms through which they impact moral inference and social functioning. Additionally, the research methods presented here can help future studies determine whether the impact of DTC on moral inference can be attributed to the specific therapeutic environment, or a more general result of recovery from BPD symptoms that may arise from any treatment modality.

A major limitation of this study is that we chose to investigate moral inference in individuals with a primary diagnosis of BPD, rather than considering symptom clusters associated with a primary diagnosis of BPD. However, it is likely that these disruptions to moral inference are not specific to BPD as a category, but rather relate to aspects of cognition that are predictive of a variety of disorders. This initial study provides a proof of concept that we have identified a dimension of cognition that distinguished between BPD patients and a sample of healthy controls. Future work should apply this measure to larger and more diverse samples to characterize how moral inference relates to a variety of other cognitive and affective dimensions that are relevant for psychiatric symptoms. Additionally, data collection in the present study relied on the availability of a small population of BPD participants who had completed DTC treatment, and a matched set of treatment-seeking BPD participants. Given that our sample size was determined by participant availability, further studies are needed to replicate the present findings and assess their generalizability to the larger population of individuals diagnosed with BPD.

A final limitation is that a number of DTC-treated and untreated BPD patients were receiving psychotropic medication. Preliminary analyses (outlined in eResults in **Supplement**) suggest that our main findings remain significant after accounting for medication use. Nonetheless, future work should investigate moral inference in a sample of BPD patients free from psychotropic medication and evaluate whether, in a larger sample, psychotropic medications influence the BPD computational phenotype that we describe.

Our moral inference paradigm captures some of the richness of BPD pathology and may have significant utility. As is the case for all disorders, clinical diagnosis of BPD relies largely on informal observation and subjective self-report. The categorical diagnostic system that relies on these data yields heterogeneous groupings that correspond poorly to disease mechanisms (35). This problem is especially serious for personality disorder, with most patients meeting criteria for multiple diagnoses (36–38). Indeed, the most common diagnosis for personality disorder patients is “not otherwise specified”, which is provided when a clinician decides a personality disorder is in fact present but the patient is not well described by existing diagnostic categories (37). This highlights the pressing need for better diagnostic tools. The paradigm described here, which can be delivered online and at scale, has potential to identify the mechanisms by which current treatments act and thus improve them. For instance, the specificity of DTC on learning about adverse social interaction partners raises the possibility that different treatments may improve different aspects of social beliefs in BPD. Using the tools presented here, we may be better equipped to identify individual differences in aberrant moral inference and match patients with treatments best suited for them. Computational modeling of moral inference dynamics may therefore prove a useful tool for investigating longitudinally how aspects of learning and impression updating might predict the course of treatment.

Translating advances in theoretical models of BPD into quantifiable benefits for patients is both conceptually and operationally challenging given the richness of BPD pathology. Tackling this problem requires precise techniques to objectively measure latent cognitive mechanisms that generate observed behavior. Here, we combine a generative model for inferring the morality of others with a moral inference task to provide mechanistic insights into social deficits in BPD. We show that BPD is associated with a specific computational phenotype of moral inference, characterized by rigid negative beliefs about other's morality. This may impact patients' ability to forgive others for their misdeeds and impact the maintenance of healthy relationships. DTC may shape social interactions in BPD by decreasing the rigidity of negative beliefs, subsequently increasing patients' openness to learning about potentially adverse others. Together, the findings demonstrate the potential for combining objective behavioral paradigms with computational modelling as a tool for assessing BPD pathology and treatment outcomes.

ACKNOWLEDGEMENTS

The authors thank Philip R. Corlett, Robb Rutledge, Hanna Pickard and Sarah Fineberg for helpful feedback. Dr. Siegel was supported by a Clarendon and Wellcome Trust Society and Ethics award (104980/Z/14/Z). Dr Saunders was supported by the Oxford Health NIHR Biomedical Research Centre. This work was supported by the Academy of Medical Sciences through grant SBF001\1008 and the NARSAD. The views expressed are not those of the NHS or the NIHR.

DISCLOSURES

The authors report no biomedical financial interests or potential conflicts of interest.

REFERENCES

1. Grant BF, Chou SP, Goldstein RB, Huang B, Stinson FS, Saha TD, *et al.* (2008): Prevalence, Correlates, Disability, and Comorbidity of DSM-IV Borderline Personality Disorder: Results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *J Clin Psychiatry* 69: 533–545.
2. King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR (2008): The Rupture and Repair of Cooperation in Borderline Personality Disorder. *Science* 321: 806–810.
3. Fineberg SK, Leavitt J, Stahl DS, Kronemer S, Landry CD, Alexander-Bloch A, *et al.* (2018): Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. *Biological Psychiatry* 84: 838–845.
4. Thielmann I, Hilbig BE, Niedtfeld I (2014): Willing to Give but Not to Forgive: Borderline Personality Features and Cooperative Behavior. *Journal of Personality Disorders* 28: 778–795.
5. Berk MS, Jeglic E, Brown GK, Henriques GR, Beck AT (2007): Characteristics of Recent Suicide Attempters with and without Borderline Personality Disorder. *Archives of Suicide Research* 11: 91–104.
6. American Psychiatric Association (2001): *Practice Guideline for the Treatment of Patients with Borderline Personality Disorder*. American Psychiatric Pub.
7. Kjær JN, Biskin R, Vestergaard CH, Munk-Jørgensen P (2015): A Nationwide Study of Mortality in Patients with Borderline Personality Disorder. *European Psychiatry* 30: 202.
8. van Asselt ADI, Dirksen CD, Arntz A, Severens JL (2007): The cost of borderline personality disorder: societal cost of illness in BPD-patients. *European Psychiatry* 22: 354–361.

9. Bateman A, Fonagy P (2009): Randomized Controlled Trial of Outpatient Mentalization-Based Treatment Versus Structured Clinical Management for Borderline Personality Disorder. *AJP* 166: 1355–1364.

10. Giesen-Bloo J, Dyck R van, Spinhoven P, Tilburg W van, Dirksen C, Asselt T van, *et al.* (2006): Outpatient Psychotherapy for Borderline Personality Disorder: Randomized Trial of Schema-Focused Therapy vs Transference-Focused Psychotherapy. *Arch Gen Psychiatry* 63: 649–658.

11. Zanarini MC, Frankenburg FR, Reich DB, Fitzmaurice G (2010): Time to Attainment of Recovery From Borderline Personality Disorder and Stability of Recovery: A 10-year Prospective Follow-Up Study. *AJP* 167: 663–667.

12. Gunderson JG, Stout RL, McGlashan TH, Shea MT, Morey LC, Grilo CM, *et al.* (2011): Ten-Year Course of Borderline Personality Disorder: Psychopathology and Function From the Collaborative Longitudinal Personality Disorders Study. *Arch Gen Psychiatry* 68: 827–837.

13. Frith CD, Frith U (2012): Mechanisms of Social Cognintion. *Annu Rev Psychol* 63: 287–313.

14. Fonagy P, Luyten P, Allison E, Campbell C (2017): What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personality Disorder and Emotion Dysregulation* 4: 9.

15. Dixon-Gordon KL, Tull MT, Hackel LM, Gratz KL (2018): The Influence of Emotional State on Learning From Reward and Punishment in Borderline Personality Disorder. *J Pers Disord* 32: 433–446.

16. Burnette JL, McCullough ME, Tongeren DRV, Davis DE (2012): Forgiveness Results From Integrating Information About Relationship Value and Exploitation Risk. *Pers Soc Psychol Bull* 38: 345–356.

17. Siegel JZ, Mathys C, Rutledge RB, Crockett MJ (2018): Beliefs about bad people are volatile. *Nature Human Behaviour* 2: 750.

18. Siegel JZ, Estrada S, Crockett MJ, Baskin-Sommers A (2019): Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature Communications* 10: 1942.

19. Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011): A Bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci* 5: 39.

20. Sansone RA, Kelley AR, Forbis JS (2013): The Relationship Between Forgiveness and Borderline Personality Symptomatology. *J Relig Health* 52: 974–980.

21. Whiteley S (2004): The Evolution of the Therapeutic Community. *Psychiatr Q* 75: 233–248.

22. Pearce S, Scott L, Attwood G, Saunders K, Dean M, Ridder RD, *et al.* (2017): Democratic therapeutic community treatment for personality disorder: Randomised controlled trial. *The British Journal of Psychiatry* 210: 149–156.

23. Debaere V, Vanheule S, Van Roy K, Meganck R, Inslegers R, Mol M (2016): Changing encounters with the other: A focus group study on the process of change in a therapeutic community. *Psychoanalytic Psychology* 33: 406–419.

24. Azzam T, Jacobson MR (2013): Finding a Comparison Group: Is Online Crowdsourcing a Viable Option? *American Journal of Evaluation* 34: 372–384.

25. Crump MJC, McDonnell JV, Gureckis TM (2013): Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS One* 8. https://doi.org/10.1371/journal.pone.0057410

26. Hauser DJ, Schwarz N (2016): Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res* 48: 400–407.

27. Behrend TS, Sharek DJ, Meade AW, Wiebe EN (2011): The viability of crowdsourcing for survey research. *Behav Res* 43: 800.

28. Peer E, Brandimarte L, Samat S, Acquisti A (2017): Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70: 153–163.

29. Mathys C, Lomakina E, Daunizeau J, Iglesias S, Brodersen K, Friston K, Stephan KE (2014): Uncertainty in perception and the Hierarchical Gaussian Filter. *Front Hum Neurosci* 8: 825.

30. Bender DS, Skodol AE (2007): Borderline Personality as a Self-Other Representational Disturbance. *Journal of Personality Disorders* 21: 500–517.

31. Barnow S, Stopsack M, Grabe HJ, Meinke C, Spitzer C, Kronmüller K, Sieswerda S (2009): Interpersonal evaluation bias in borderline personality disorder. *Behaviour Research and Therapy* 47: 359–365.

32. Critchfield KL, Levy KN, Clarkin JF, Kernberg OF (2008): The relational context of aggression in borderline personality disorder: using adult attachment style to predict forms of hostility. *Journal of Clinical Psychology* 64: 67–82.

33. Unoka Z, Seres I, Áspán N, Bódi N, Kéri S (2009): Trust Game Reveals Restricted Interpersonal Transactions in Patients With Borderline Personality Disorder. *Journal of Personality Disorders* 23: 399–409.

34. Fonagy P, Luyten P, Allison E (2015): Epistemic Petrification and the Restoration of Epistemic Trust: A New Conceptualization of Borderline Personality Disorder and Its Psychosocial Treatment. *Journal of Personality Disorders* 29: 575–609.

35. Kapur S, Phillips AG, Insel TR (2012): Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry* 17: 1174–1179.

36. Lenzenweger MF, Lane MC, Loranger AW, Kessler RC (2007): DSM-IV Personality Disorders in the National Comorbidity Survey Replication. *Biological Psychiatry* 62: 553–564.

37. Verheul R, Widiger TA (2004): A Meta-Analysis of the Prevalence and Usage of the Personality Disorder Not Otherwise Specified (PDNOS) Diagnosis. *Journal of Personality Disorders* 18: 309–319.

38. Tyrer P, Reed GM, Crawford MJ (2015): Classification, assessment, prevalence, and effect of personality disorder. *The Lancet* 385: 717–726.

Table 1

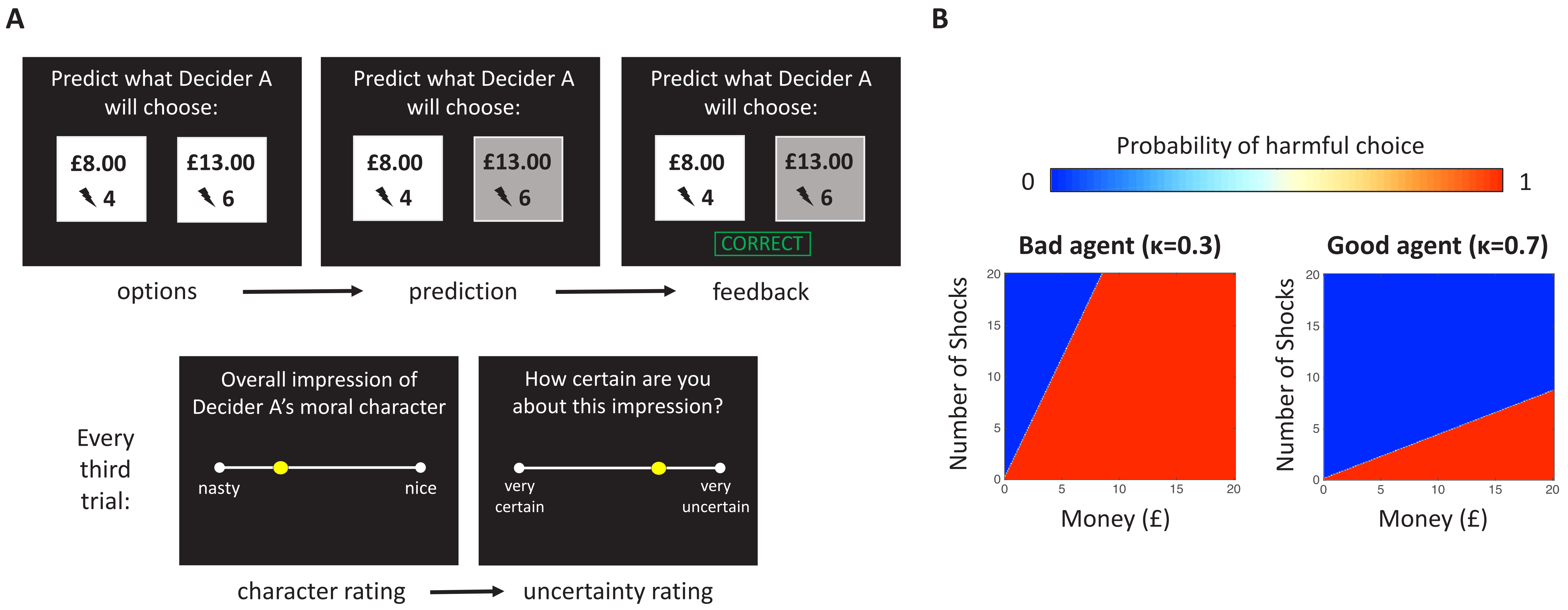
Participant demographic information, BPD vs. non-BPD. SEM = Standard error of the mean.

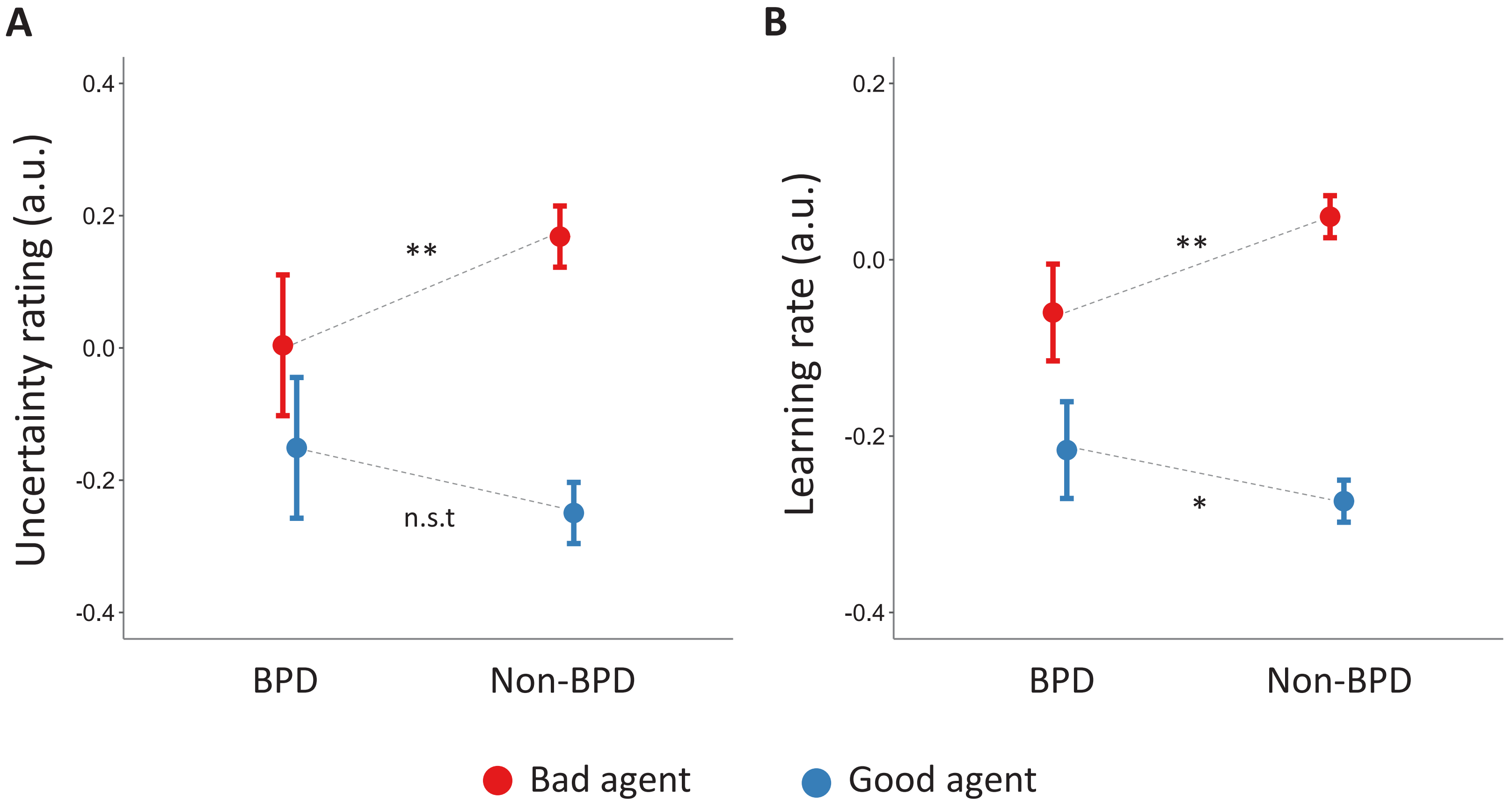
	Untreated BPD (N=20)		Non-BPD control (N=106)			
	Mean	**SEM**	**Mean**	**SEM**	***Z*-stat**	***p*-value**
Age on date of participation	39.500	2.561	40.957	1.140	-0.612	0.540
Highest level of education	2.412	0.195	2.587	0.094	-0.861	0.389
Psychopathy	42.053	2.024	38.387	0.795	1.437	0.151
Personality inventory for DSM-V	39.950	3.042	18.740	1.202	5.269	<0.001

Table 2

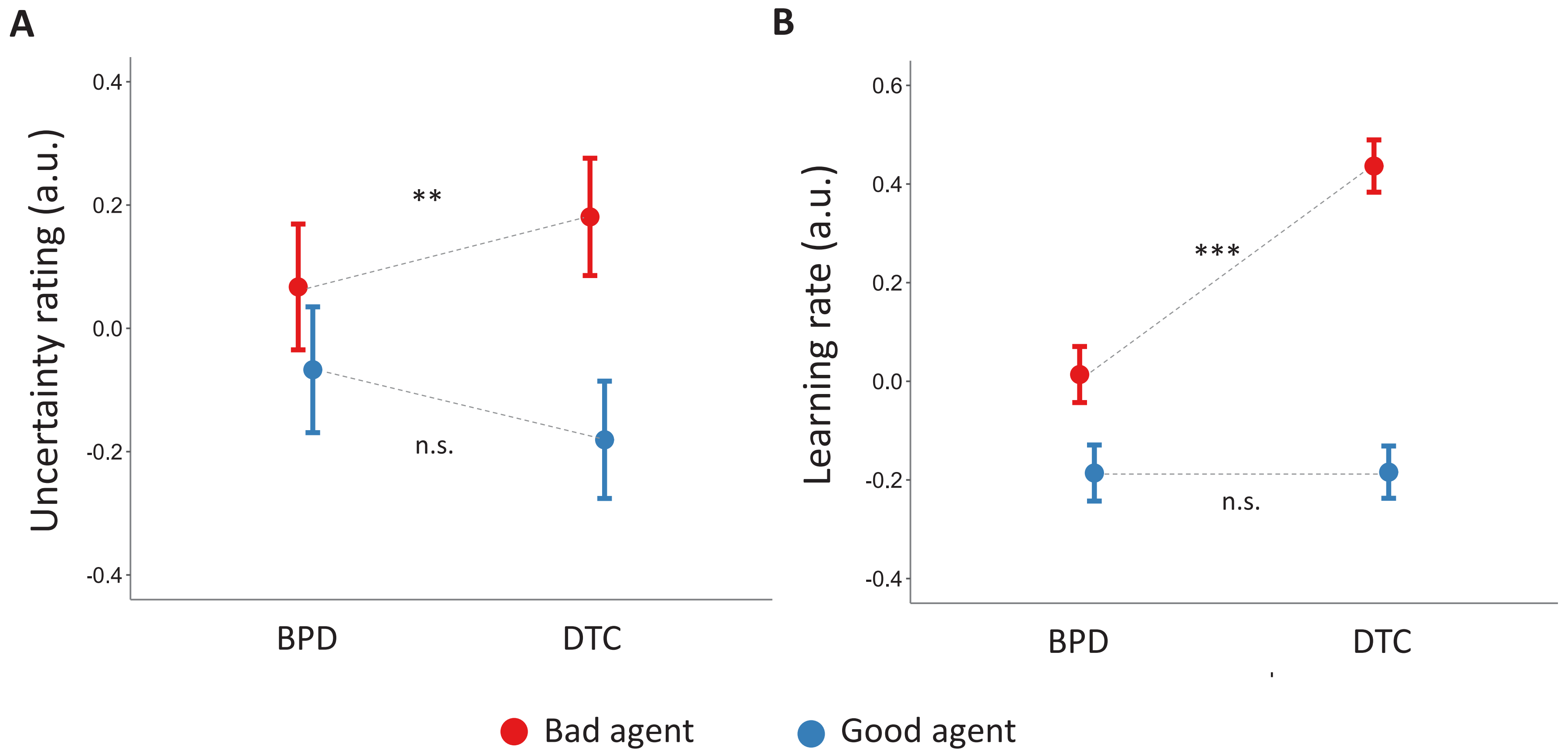
Participant demographic information, untreated vs. DTC-treated BPD. SEM = Standard error of the mean.

	Untreated BPD (N=20)		**DTC-treated (N=23)**			
	Mean	**SEM**	**Mean**	**SEM**	***Z*-stat**	**p-value**
Age on date of participation	39.500	2.561	41.609	2.205	-0.573	0.567
Highest level of education	2.412	0.195	2.632	0.211	-0.748	0.455
Psychopathy	42.053	2.024	40.217	2.628	0.999	0.318
Personality inventory for DSM-V	39.950	3.042	33.478	3.029	1.572	0.116
Borderline evaluation of severity over time (BEST)	41.444	1.975	26.867	1.956	3.690	<0.001


A
Predict what Decider A will choose:
£8.00
4
£13.00
6
CORRECT
options
prediction
feedback
Every third trial:
Overall impression of Decider A's moral character
nasty
nice
How certain are you about this impression?
very certain
very uncertain
character rating
uncertainty rating
B
Probability of harmful choice
0
1
Bad agent (κ=0.3)
Good agent (κ=0.7)
Number of Shocks
Money (£)
0
5
10
15
20



A
B
Uncertainty rating (a.u.)
0.4
0.2
0.0
-0.2
-0.4
**
n.s.t
BPD
Non-BPD
Learning rate (a.u.)
0.2
0.0
-0.2
-0.4
**
*
BPD
Non-BPD
Bad agent
Good agent


A
B
Uncertainty rating (a.u.)
0.4
0.2
0.0
-0.2
-0.4
**
n.s.
BPD
DTC
Learning rate (a.u.)
0.6
0.4
0.2
0.0
-0.2
-0.4
***
n.s.
BPD
DTC
Bad agent
Good agent

A Computational Phenotype of Disrupted Moral Inference in Borderline Personality Disorder

Supplemental Information

SUPPLEMENTAL METHODS

Moral Inference Task

In the Moral Learning Task, participants predicted a series of 50 decisions, made by each of two agents. For each decision, agents chose whether to increase their own profit at the expense of a greater amount of harm, in the form of electric shocks, to an anonymous stranger (**Figure 1a**). Thus, each choice involved choosing between a more harmful option (more money and more shocks) and a less harmful option (less money and less shocks). We simulated the agents to have significantly different preferences towards harming the stranger: one agent was more harmful, accepting less money to increase shocks to the victim ('bad' agent; $0.43 per shock), and the other was less harmful and required more money to increase shocks ('good' agent; $2.40 per shock; **Figure 1b**). After predicting each choice, participants received feedback about their accuracy. Participants did not receive any information about the agents' harm preference prior to the task. Thus, to optimally predict the agents' decisions participants must gather information across trials and learn about the agents' harm preference (i.e., the agent's exchange rate between money and shocks). For complete details about the task and how the agent's choices were simulated, see Siegel et al. 2018 (1).

On every third trial participants indicated their general impression of the agent's moral character (from 0 = *nasty* to 100 = *nice*) and how *certain* they were about their impression (from 0 = *very uncertain* to 100 = *very certain*). This provided us, for each subject and agent, a trajectory of trial-wise *subjective impression ratings* and *uncertainty ratings*. Before observing any of the agent's choices, participants additionally indicated how nasty or nice they *expected* the agents would be and how certain they were. This provided an indication of participants' prior expectations about people's moral character in general and their confidence in those prior expectations.

Hierarchical Gaussian Filter (HGF)

The HGF (2,3) draws on the idea that the brain has evolved to process information in a manner that approximates statistical optimality given individually varying priors about the nature of the process being predicted; effectively maintaining and updating a generative model of its inputs to infer on hierarchically organized hidden states. A basic feature of the model is the division into perceptual and response models, which describes both how participants update their beliefs about hidden states from inputs (perceptual model) and how they are used to make predictions (response model).

Perceptual model. Our model comprises only two hidden states x_1^i and x_2^i, where i signifies the trial index. The first state, x_1, is time-varying and denotes the agent's upcoming choice. x_1 is binary because there are only two options that the agent can choose: the more harmful option (greater profit for the self and more shocks for the victim) or the less harmful option (less profit for the self and fewer shocks for the victim). The probability that an agent will choose the more harmful option ($x_1^i = 1$) versus the less harmful option ($x_1^i = 0$) is governed by the next state in the hierarchy, x_2. x_2 is a continuous state evolving over time as a Gaussian random walk, and signifies the belief about the agent's exchange rate between money and pain. The hierarchical coupling between x_1^i and x_2^i explains that a participant's prediction about an agent's choice on trial i is dependent on their current belief about that agent's exchange rate between money and pain, defined as a probability density.

The conditional probability of x_1 given x_2 is described in **Equation 1**.

Equation 1

$$p(x_1|x_2) = s(x_2)^{x_1}\big(1 - s(x_2)\big)^{1-x_1} = \text{Bernoulli}(x_1; s(x_2))$$

Where $s(\cdot)$ is a logistic sigmoid (softmax) function:

Equation 2

$$s(x) \stackrel{\text{def}}{=} \frac{1}{1 + \exp(-x)}$$

The temporal evolution of x_2 is governed by a participant-specific parameter ω, which allows for inter-individual differences in belief updating. Thus, ω captures inter-individual variability in the rate at which beliefs evolve over time, and consequently how rapidly people update their beliefs about the agent's harm aversion across all trials. As ω approaches ∞ beliefs become increasingly unstable and new information is favored over historical information. Conversely, as ω approaches -∞ beliefs become increasingly stable, so greater weight is instead placed on historical information. Given ω and the previous value (with time index $i - 1$) of x_2, we now have the generative model for the current values (with time index i) of x_1 and x_2 in **Equation 3** (for details see (2)).

Equation 3

$$p\big(x_1^i, x_2^i, |\omega, x_2^{i-1}\big) = p\big(x_1^i\big|x_2^i\big)p\big(x_2^i\big|x_2^{i-1}, \omega\big)$$

with

Equation 4

$$p\big(x_2^i\big|x_2^{i-1}, \omega\big) = \mathcal{N}\big(x_2^i; x_2^{i-1}, \exp(\omega)\big)$$

Model inversion was used to optimize the posterior densities over hidden states, x_1 and x_2, and parameter ω. Participants' posterior beliefs were represented by probability distributions with mean μ and variance σ. Variational Bayesian inversion yields a simple update equation under a mean-field approximation, where beliefs are updated as a function of precision-weighted prediction errors. For the present study we focus on the update at level 2 of the hierarchy (2).

Equation 5

$$\Delta\mu \propto \sigma_2 \delta_1^i$$

with

Equation 6

$$\delta_1^i = \mu_1^i - \hat{\mu}_1^i$$

and

Equation 7

$$\sigma_2 = \frac{\hat{\pi}_1^i}{\hat{\pi}_2^i \hat{\pi}_1^i + 1}$$

Where π is the precision (i.e., the inverse variance) in participants' posterior belief $\frac{1}{\sigma}$, and δ_1^i is the prediction error on the trial outcome. Caret symbols (^) are used to denote predictions *prior* to observing the outcome at trial *i*. Thus, $\hat{\pi}_1^i$is the precision of the prediction at the first hierarchical level and $\hat{\pi}_2^i$ is the precision of the prediction of the posterior belief. It can be shown from **Equation 7** that prediction errors are given a larger weight when the precision of the prediction of the agent's choice is high, or when the precision of the belief about the agent's preference (i.e., exchange rate between money and pain) is low. In summary, these equations describe trial-wise updating of beliefs about an agent's preference towards harming the victim, which approximates Bayes optimality (in an individualized sense given differences in ω) and determines the participant's estimate of the probability that an agent will harm. Crucially, our model provides a trial-by-trial estimate of the subject's uncertainty about the agent's preference towards harming the victim as measured by the variance of beliefs, σ. The variance weights predictions errors on a trial-by-trial basis and thus represents a *dynamic* learning rate because it accounts for the precision of the belief at any given time.

Decision model. The decision model describes how the participant's posterior belief about the agent's preference maps onto their predictions of the agent's decisions (*y*). In the HGF, this belief $\hat{\mu}_1^i$corresponds to the logistic sigmoid transformation of the predicted preference $\mu_2^{i-1}$of the agent towards harming the victim.

Equation 8

$$\hat{\mu}_1^i = s(\mu_2^{i-1})$$

For the present study, we assumed that participants would predict others' decisions using a similar rationale to how they make decisions themselves. In other words, we assumed that people's preferences are described by a utility model, and that people think others' preferences are described by the same model. Consequently, we applied a decision model that accurately describes human choices in the same choice setting (4–6).

Equation 9

$$V_{\text{harm}}^i = \left(1 - \hat{\mu}_1^i\right)\Delta \mathrm{m}^i - \hat{\mu}_1^i \Delta \mathrm{s}^i$$

This applied the predicted belief about the agent's preference derived from the perceptual model $\hat{\mu}_1^i$ to compute the value that the agent will choose the more harmful option on trial *i,* given the difference in money (Δm) and shocks (Δs) between the two options. The probability that the participant predicts the more harmful option ($y = 1$) as opposed to the more helpful option ($y = 0$) is described by the softmax function in **Equation 10**.

Equation 10

$$P_{\text{harm}}^i = s(\beta V_{\text{harm}}^i)$$

Where β is a free parameter (individually estimated like ω) that describes how sensitive predictions are to the relative utility of different outcomes, or the prediction noise.

Estimation of model parameters. A crucial aspect of Bayesian inference is the specification of a prior distribution for the belief (listed in **Supplementary Table** S**1**). We defined the priors based on previous research using the same experimental design. Specifically, in keeping with our experimental design, which did not give participants any basis for assumptions about the agent's tendency to harm, we chose to initialize the prior mean over μ_2 and σ_2. such that it amounted to a neutral prior belief about κ which was equidistant from the true value of the agents' preferences. For the free parameters ω and β, we chose a prior mean that was relatively uninformative (with large variance) to allow for substantial individual differences in learning both between participants and within participants (i.e. between agents).

Supplementary Table S1

Prior mean and variance of the perceptual and response model parameters.

Parameter	Notes		mean	variance
ω	Constant component of the tonic volatility at the second level. Represents the temporal evolution of x_2. *Estimated in native space.*		-4	1
Predictions (x_1)	Predictions are a sigmoid transformation of x_2, and so do not have prior values.	μ_1:	none	none
		σ_1:	none	none
Probabilities (x_2)	The prior mean on x_2 (prior belief about agent's harm-aversion, κ) was fixed to a neutral point that was equidistant from the true κ value of both agents. Estimated in logit space.	μ_2:	0.5	0
	The prior variance on x_2 was fixed to ensure that any differences in learning about good and bad agents derived from the model could not result from differences in the prior estimates. Estimated in log-space.	σ_2:	0.35	0
β	Constant component that describes how sensitive prior beliefs are to the relative utility of different outcomes, or the prediction noise. Estimated in log-space.		1	1

The perceptual model parameter ω and decision model parameter β were estimated from the trial-wise predictions using the Broyden Fletcher Goldfarb Shanno optimization algorithm as implemented in the HGF Toolbox (https://tnu.ethz.ch/tapas). This allowed us to obtain the maximum-a-posteriori estimates of the model parameters and provided us with state trajectories and parameters representing an ideal Bayesian observer given the individually estimated parameter ω.

We fit the model separately for participant's predictions of the bad and good agent. This produced for each agent a sequence of trial-wise beliefs about the agent's preference ($\hat{\mu}_1^i$), as well as the precision of each belief (σ^i), and two participant-specific parameters, ω and β. to the temporal emphasis of belief stability in BPD, we focus out analysis on variance of beliefs σ, which reflects a dynamic learning rate dictating trial-by-trial belief updating as a function of the precision (i.e., inverse uncertainty) of beliefs about the agent's moral preference.

Additional Measures

Borderline evaluation of severity over time (BEST). We used the BEST (7) to assess the severity of BPD symptomology in participants with BPD at the time of participation. The BEST is a 15-item questionnaire which measures thoughts, emotions, and behaviors (positive and negative) typical of BPD. Positive behaviors were not measured in this study, and thus participants responded to only 12 of the 15 items. Each item asks participants to rate their experience with each of the items since their last clinical session; the lowest score of 1 means that it caused little or no problems, and the highest score of 5 means that it caused extreme distress, severe difficulties with relationships, and/or kept them from completing tasks. The scores from the 12 items were added together to yield a score between 12 and 60, where higher scores indicated greater BPD severity.

Personality Inventory for DSM-5, brief form (PID-5-BF). We used the PID-5-BF (8), a 25-item self-report questionnaire, to assess clinically relevant personality traits that do not necessarily constitute a personality disorder. The PID-5-BF constitutes five personality trait domains: negative affect, detachment, antagonism, disinhibition, and psychoticism. Each item on the questionnaire asks participants to rate how well the item describes him or her generally on a scale from 0 (*very false or often false*) to 3 (*very true or often true*). The scores from all items were added together to produce a score between 0 and 75, with higher scores indicating greater general overall personality dysfunction.

McLean Screening Instrument for BPD (MSI). The MSI (9) was used as a screening measure for the presence of clinically relevant BPD in the control group. The validated instrument consists of ten true-false self-report questions to assess the occurrence of symptoms typically found in BPD, such as "*Have you deliberately hurt yourself physically (e.g. punched yourself, cut yourself, burned yourself)*". The screen is regarded as positive when seven or more of the symptoms are true.

Self Report Psychopathy - Revised, short form (SRP-R-SF). We used the SRP-R-SF (10), a 29-item self-report questionnaire, to assess psychopathic personality traits across BPD participants and non-BPD control participants. The instrument constitutes four factors of psychopathy: affective callousness, interpersonal manipulation, antisociality, and erratic lifestyle. Each item on the questionnaire asks participants to rate the extent to which they thought the item reflected their own beliefs using a 5-point likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The scores from all items were added together to produce a total psychopathy score, with higher scores indicating greater general overall psychopathic personality traits.

Structured Clinical Interview for axis II disorders (SCID-II). The SCID-II is a semi-structured clinical interview administered by trained clinical and designed to asses a clinical diagnosis of axis II disorders consistent with the DSM-IV. The SCID-II was used to establish a clinical diagnosis of BPD in untreated BPD and DTC-treated participants.

SUPPLEMENTAL RESULTS

Motivation to accurately predict the agents' choices. Because non-BPD and BPD participants completed the task under very different experimental settings (non-BPD participants: conducted online, BPD participants: conducted in the clinic), we wanted to verify that the groups were equally motivated to learn about the agents and predict their decisions. Consequently, after predicting all the choices for a given agent, we explicitly asked participants to indicate on a continuous scale from 0 (*very unmotivated*) to 100 (*very motivated*) "How motivated to be accurate did you feel during the task?". We additionally calculated the percent of choices accurately predicted by each participant and compared between groups. We confirmed that BPD and non-BPD participants were similarly accurate (% accuracy: bad: $Z = -1.103$, $p = 0.270$; good: $Z = 0.295$, $p = 0.768$) and motivated in their predictions (motivation rating: bad: $Z = -\ 0.879$, $p = 0.\ 379$; good: $Z = -1.704$, $p = 0.088$).

Model validation. Three computational models were compared to describe how participants learned the agents' preferences and predicted their choices. We fit the HGF (2,3), which identified participant-specific parameters to describe each individual participant's learning process. Beliefs about an agent's harm preference were updated using a Bayesian reinforcement learning algorithm, with precision-weighted prediction errors driving belief updating at the different levels of the hierarchical model. Second, we fit a Rescorla Wagner model, in which beliefs were updated by prediction errors with a fixed learning rate. Third, we fit a modified Rescorla Wagner model, in which beliefs were updated by prediction errors with separate fixed learning rates for helpful and harmful outcomes. For details about the alternative models, see **Supplementary Table 2**.

Supplementary Table S2

Details of alternative models for model comparison

Model	Notes	Estimated parameters
Rescorla Wagner with one learning rate	Beliefs are symmetrically updated, with a single learning rate for each participant.	α = Learning rate β = Prediction noise
Rescorla Wagner with two learning rates	Beliefs are asymmetrically updated, with separate learning rates for positive versus negative outcomes, for each participant.	α_{pos} = Learning rate positive outcomes α_{pos} = Learning rate negative outcomes β = Prediction noise
HGF	A two level model, with one estimated parameter governing the volatility of beliefs at the second level, and a second estimated parameter governing the prediction noise.	ω = Tonic volatility β = Prediction noise

The log-model evidence (LME) indicated that the HGF model (sum LME = -7149) outperforms both a simple single learning rate RW model (sum LME = -7444) and a RW model with separate learning rates for positive and negative outcomes (sum LME = -7192). We validated these findings using formal Bayesian Model Selection. To this end, we used LME data to compare between the HGF and our two RW models. This analysis yielded a protected exceedance probability indistinguishable from 1 for the HGF model for both agents, indicating effectively a 100% probability that the HGF model better explains the data than the other models included in the comparison.

Subjective uncertainty ratings in BPD versus non-BPD and DTC-treated participants. For completeness, we performed an omnibus robust linear regression analysis on subjective uncertainty ratings that included all three groups (BPD, non-BPD, and DTC) in a single model, where group was dummy coded with BPD as the reference group. Tests of group effects were conducted using Bonferroni adjusted alpha levels of .025 to account for multiple comparisons. The analysis yielded a significant main effect of agent (β = -0.155±0.073, t = 2.126, p = .034), indicating that participants held more uncertain impressions of the bad agent than the good agent. Overall, BPD participants uncertainty ratings did not significantly differ from non-BPD participants (β = -0.098±0.056, t = -1.739, p = .082), or DTC-treated participants (β = -0.089±0.071, t = -1.258, p = .209). The effect of agent was significantly smaller in BPD participants, relative to both non-BPD participants (β = 0.264±0.080, t = 3.310, p < .001) and DTC-treated participants (β = 0.266 ±0.100, t = 2.665, p = .008), as indicated by significant interactions between agent and group.

Learning rates in BPD versus non-BPD and DTC-Treated participants. We performed an omnibus robust linear regression analysis on learning rates that included all three groups (BPD, non-BPD, and DTC) in a single model, where group was dummy coded with BPD as the reference group. Again, this analysis yielded a significant main effect of agent (β = -0.831±0.023, t = 36.888, p < .001), indicating that participants updated beliefs about the bad agent at a faster rate than the good agent. Overall, learning rates for the untreated BPD participants did not differ from non-BPD participants (β = -0.001±0.017, t = -0.060, p = .953), or DTC participants (β = -0.024±0.022, t = -1.103, p = .270). However, relative to untreated BPD participants, the effect of agent on learning rates was significantly larger relative to both non-BPD participants (β = 0.113±0.025, t = 4.607, p < .001) and DTC-treated participants (β = 0.319 ±0.031, t = 10.355, p < .001), as indicated by significant interactions between agent and group.

Subjective moral impressions in BPD versus non-BPD participants. Examining subjective impression ratings revealed that participants formed more negative impressions about the 'bad' agent than the 'good' agent (mean±SEM, β = -1.178 ± 0.027, t = -44.299, p < .001). The main effect of group (β = -0.041 ± 0.047, t = -.872, p = .383) and the interaction between agent and group were not significant (β = -0.441 ± 0.067, t = -1.706, p = .088). Thus, the valence of moral impressions did not vary as a function of BPD diagnosis.

Subjective uncertainty ratings in BPD versus DTC-treated participants. Examining subjective uncertainty ratings yielded a significant main effect of agent (β = 0.156±0.070, $t = 2.240$, $p = .025$), indicating that participants held more uncertain impressions of the bad agent than the good agent. DTC-treated and untreated BPD participants were similarly uncertain about their impressions overall (β = -0.085 ± 0.067, $t = -1.265$, $p = .206$). However, we found that DTC-treated BPD participants, relative to untreated BPD participants, showed more uncertain impressions of the bad agent (β = 0.188±0.067, t = 2.802, $p = .005$; **Figure 3a**) as indicated by significant interactions between agent and group (β = 0.277±0.095, $t = 2.904$, $p = .003$).

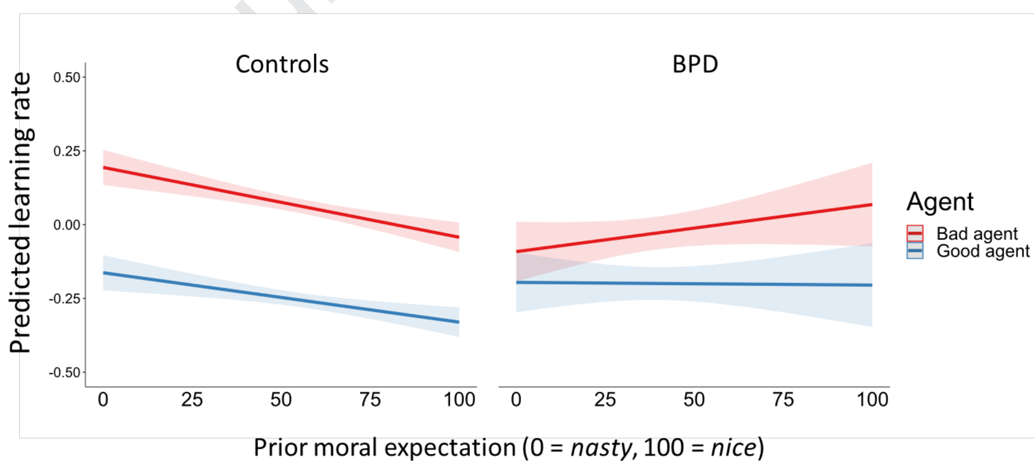
Learning rates in BPD versus DTC-Treated participants. Again, we observed a significant main effect of agent on learning rates (β = 0.153±0.037, $t = 4.115$, $p < .001$), indicating that BPD participants updated beliefs about the bad agent at a faster rate than the good agent. Overall, learning rates for the DTC-treated and untreated BPD participants did not significantly differ (β = -0.031 ± 0.036, $t = -0.870$, $p = .384$). However, we found that DTC-treated BPD participants, relative to untreated BPD participants, showed faster learning rates for the bad agent (β = 0.543±0.040, $t = 13.698$, $p < .001$; **Figure 3b**), as indicated by significant interactions between agent and group (β = 0.589±0.052, $t = 11.588$, $p < .001$).

Effect of individual differences in the severity of BPD symptomology on subjective uncertainty ratings and learning rates. We used a robust linear regression model that included the effects of agent (bad, good), and Borderline Evaluation of Severity over Time (BEST) scores, and their interaction (controlling for trial number) to investigate their effects on subjective uncertainty ratings and learning rates. Consistent with prior findings, participants overall held more uncertain impressions of the bad agent than the good agent (main effect of agent: β = 0.904±0.171, $t = 5.272$, $p < .001$) and faster learning rates for the bad agent than the good agent (β = 1.308±0.052, $t = 25.193$, $p < .001$). However this effect decreased with increasing BPD symptomology (interaction between agent and BEST: *uncertainty rating,* β = -0.018±0.005, t = -3.784, p < .001; *learning rate*, β = -0.004±0.001, t = -2.821, p = .005). Specifically, higher BEST scores were associated with less uncertain impressions of the bad agent (β = -0.012 ± 0.003, t=-3.262, $p = .001$), though the effect on learning rates did not reach significance (β = -0.003 ± 0.002, t=-1.514, $p = .130$). Higher BEST scores were associated with *more* uncertain impressions of the good agent (β = 0.007 ± 0.003, t=2.078, $p = .038$), and faster belief updating (β = 0.003 ± 0.001, t=6.118, $p < .001$).

Prior expectations in moral inference. BPD participants expressed more pessimistic expectations about the agents' moral behavior than non-BPD participants. Thus, a plausible explanation for more certain beliefs about bad agents and less certain beliefs about good agents is that the good agent violated BPD participants' expectations to a greater degree than the bad agent. In other words, the bad agent's behavior would be more consistent with patient's prior expectation (and therefore increase confidence and rigidity of posterior beliefs) while the good agent's behavior would be less consistent with patient's prior expectations (thus, decrease confidence and rigidity of posterior beliefs).

Previous work suggests that prior moral expectations are unlikely to impact the ability to adapt learning as a function of moral information in healthy adults (1). Human may have evolved to adapt learning according to moral information to aid survival. In turn, this adaptive mechanism may enable healthy adults to discount expectations to build richer models of agents when harmful outcomes are expected (i.e., in response to negative moral expectations). One possibility is that patients with BPD lack the mechanism for adapting learning according to moral information. That is, while healthy adults may be able to override prior expectations and rapidly adjust their learning for putatively bad agents, this adaptive mechanism may be absent in BPD. As a result, learning may be more sensitive to prior expectations in BPD. If this is the case, we would expect learning in BPD to be more strongly influenced by prior moral expectations than learning in non-BPD participants.

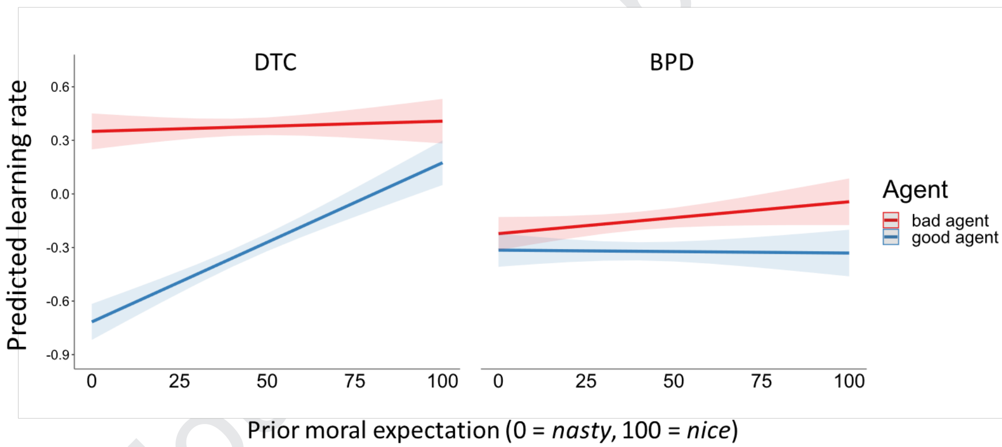
In line with this prediction, we found a significant three-way interaction between prior expectations, BPD diagnosis, and agent ($\beta = 0.004 \pm 0.002$, $t = 2.214$, $p = .027$). To unpack the interaction, we performed a similar regression splitting the data as a function of BPD diagnosis. Consistent with previous findings (1), prior expectations were not associated with differences in learning rates between the good and bad agent in non-BPD control participants ($\beta = -0.001 \pm 0.001$, $t = -1.454$, $p = .146$; **Supplementary Figure S1**). Conversely, prior expectations predicted asymmetric learning rates for good and bad agents in BPD participants: more pessimistic expectations were associated with a smaller learning asymmetry ($\beta = 0.003 \pm 0.001$, $t = 2.250$, $p = .025$; **Supplementary Figure** S1). The findings provide preliminary evidence to suggest that the mechanisms underlying the ability to rapidly adapt learning towards moral information in healthy adults may be absent in BPD.



Supplementary Figure S1. Prior moral expectations moderate belief updating in BPD. Effect of prior moral expectations on estimated learning rates for the control (i.e., non-BPD) group (left) and BPD group (right). Prior moral expectations were measured on a continuous scale before

observing any of the agent's choices. The scale asked participants to indicate how nasty or nice they expected the agents would be in the task. Error bands represent 95% confidence intervals.

Prior expectations did not significantly differ between DTC-treated and untreated BPD participants. Nonetheless, we performed a similar regression analysis to explore the three-way interaction and observed a significant interaction between prior expectations, agent, and group on learning rates (β = -0.010±0.002, t = -4.752, p < .001; **Supplementary Figure S2**). To unpack the interaction, we fit the regression model separately for untreated BPD and DTC treated participants. Again, we found that worse expectations were associated with smaller asymmetric updating between agents in the BPD group (β = 0.003±0.001, t = 2.250, p = .025). However, the opposite pattern was observed for the DTC treated group: worse expectations were associated with larger asymmetric updating between agents (β = -0.007±0.002, t = -4.615, p < .001). These findings suggest that even though DTC-treated and untreated BPD groups had similar moral expectations, the groups differed in how expectations subsequently shaped learning.



Supplementary Figure S2. Prior moral expectations moderate belief updating. Effect of prior moral expectations on estimated learning rates for the DTC group (left) and BPD group (right). Prior moral expectations were measured on a continuous scale before observing any of the agent's choices. The scale asked participants to indicate how nasty or nice they expected the agents would be in the task. Error bands represent 95% confidence intervals.

BPD, medication use, and moral inference. A supplementary analysis investigated the interaction between group (DTC vs. BPD) and agent (bad vs. good) on subjective uncertainty and learning rates, controlling for medication use. Medication use was entered into the regression as a dummy variable and indicated whether the participants were receiving psychotropic or antidepressant medication during the time of participation. Medication use did not significantly predict subjective uncertainty ratings (β = 0.027±0.048, t = 0.551, p = .582) nor did patient group (β = -0.082±0.068, t = -1.201, p = .230). Overall, participants were more uncertain about their impressions of the bad agent relative to the good agent (β = 0.156±0.070, t = 2.240, p = .025). The interaction between group and agent on subjective uncertainty remained significant after controlling for medication use (uncertainty: β = 0.277±0.095, *t* = 2.898, *p* = .004; learning rates: β = 0.577±0.050, *t* = 11.441, *p* < .001). Relative to untreated BPD participants, DTC-treated participants were more uncertain about their impressions of the bad agent (β = 0.183±0.068, *t* = 2.691, *p* = .007) but did not significantly differ in their uncertainty about their impressions of the good agent (β = -0.068±0.069, *t* = -0.998, *p* = .318).

Patient group did not significantly predict overall learning rates (β = -0.051±0.036, *t* = -1.427, *p* = .154). Medication use was associated with slower learning rates overall (β = -0.180±0.026, *t* = -7.050, *p* < .001) and participants had higher learning rates for the bad agent relative to the good agent (β = 0.169±0.037, *t* = 4.587, *p* < .001). Notably, the interaction between group and agent on learning rates remained significantly after controlling for medication use (β = 0.577±0.050, *t* = 11.441, *p* < .001). Relative to untreated BPD participants, DTC-treated participants had higher learning rates for bad agent (β = 0.516±0.040, *t* = 12.853, *p* < .001) but marginally lower learning rates for the good agent (β = -0.058±0.030, *t* = -1.922, *p* = .055).

Supplemental References

1. Siegel JZ, Mathys C, Rutledge RB, Crockett MJ (2018): Beliefs about bad people are volatile. *Nature Human Behaviour* 2: 750.

2. Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011): A Bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci* 5: 39.

3. Mathys C, Lomakina E, Daunizeau J, Iglesias S, Brodersen K, Friston K, Stephan KE (2014): Uncertainty in perception and the Hierarchical Gaussian Filter. *Front Hum Neurosci* 8: 825.

4. Crockett MJ, Siegel JZ, Kurth-Nelson Z, Ousdal OT, Story G, Frieband C, *et al.* (2015): Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making. *Curr Biol* 25: 1852–1859.

5. Crockett MJ, Kurth-Nelson Z, Siegel JZ, Dayan P, Dolan RJ (2014): Harm to others outweighs harm to self in moral decision making. *PNAS* 111: 17320–17325.

6. Crockett MJ, Siegel JZ, Kurth-Nelson Z, Dayan P, Dolan RJ (2017): Moral transgressions corrupt neural representations of value. *Nat Neurosci* 20: 879–885.

7. Pfohl B, Blum N, St. John D, McCormick B, Allen J, Black DW (2009): Reliability and validity of the borderline evaluation of severity over time (BEST): a self-rated scale to measure severity and change in persons with borderline personality disorder. *J Pers Disord* 23: 281–293.

8. Krueger RF, Derringer J, Markon KE, Watson D, Skodol AE (2012): Initial Construction of a Maladaptive Personality Trait Model and Inventory for DSM-5. *Psychol Med* 42: 1879–1890.

9. Zanarini MC, Vujanovic AA, Parachini EA, Boulanger JL, Frankenburg FR, Hennen J (2003): A screening measure for BPD: the McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD). *J Pers Disord* 17: 568–573.

10. Neumann C, Pardini D (2012): Factor Structure and Construct Validity of the Self-Report Psychopathy (SRP) Scale and the Youth Psychopathic Traits Inventory (YPI) in Young Men. *Journal of Personality Disorders* 28: 419–433.