UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF

Institut für Systemische Neurowissenschaften

Prof. Dr. med. Christian Büchel

Learning, cooperating, sharing: A multi-stage account of social decision-making in borderline personality disorder

Dissertation

zur Erlangung des akademischen Grades Dr. rer. biol. hum.
an der Medizinischen Fakultät der Universität Hamburg.

vorgelegt von
Lisa Marie Doppelhofer
aus Wien, Österreich

Hamburg 2026

Gutachtende:

Gutachter: Prof. Dr. Christoph Korn

Gutachterin: Prof. Dr. Anja Riesel

Mietglieder der Prüfungskommission:

Vorsitz der Prüfungskommission: Prof. Dr. Christoph Korn

Mitglied der Prüfungskommission: Prof. Dr. Anja Riesel

Mitglied der Prüfungskommission: PD Dr. Sarah Biedermann

Datum der Disputation: 28.08.2026

Für Mama und Papa

Table of Contents

Summary 6
Zusammenfassung 8
List of publications 10
1 General introduction 11
1.1 Cooperation as a multi-stage process 11
1.2 Borderline Personality Disorder and interpersonal dysfunction 12
1.3 Methods toolbox for isolating social decision processes 13
1.3.1 Economic games and controlled interdependence paradigms 14
1.3.2 Computational models: reinforcement-learning and utility-based approaches 15
1.4 Stage 1: Learning about others 16
1.5 Stage 2: Cooperation, expectations, and prefrontal systems 17
1.6 Stage 3: Sharing the gains: allocations, fairness, and social preferences 18
1.7 Thesis approach and contributions 21
1.8 Open science contributions 22
2 Aims 23
3 Study I: Estimating and learning personality traits 24
4 Study II: Social decision-making and expectations 45
5 Study III: Fairness preferences and allocation of payoffs 112
6 Study IV: Shift in prosocial motivation but not in social value orientation 132
7 General discussion 142
7.1 What is altered, what is preserved? 142
7.1.1 Altered social priors with preserved learning mechanisms 142
7.1.2 Preserved cooperative motivation with altered expectations 143
7.1.3 Preserved allocation behavior in non-interactive settings 144
7.1.4 Preserved prosocial behavior and altered underlying motives 145
7.2 An integrative account: negative priors and preserved cooperative capacity 146
7.3 Convergence and divergence across paradigms 147
7.4 Neural implications 148
7.5 Clinical implications: targets and opportunities 150
7.6 Limitations and strengths 151
7.7 Future directions 152
8 Conclusion 153
References 154
Erklärung des Eigenanteils an den Publikationen 166

Acknowledgments 167
Curriculum Vitae 168
Eidesstattliche Versicherung 171
Supplementary material for all studies 172

Summary

Human social life relies heavily on cooperation. Cooperative behavior supports social connection, relationship stability, and the pursuit of shared goals, and it requires individuals to evaluate others, anticipate their behavior, and coordinate actions over time. When these processes break down, interpersonal relationships can become unstable and difficult to sustain. Borderline Personality Disorder (BPD) is characterized by pervasive interpersonal instability, making it a clinically relevant framework for investigating how disruptions in cooperation processes can contribute to unstable social interaction. The present thesis addresses this problem by conceptualizing cooperation not as a unitary behavior, but as a multi-stage social decision-making process that unfolds over time. Specifically, cooperation is decomposed into three interdependent components: (i) evaluating and learning about other people, (ii) making cooperative choices under interdependence and uncertainty, and (iii) distributing gains according to social motives. Across four projects, this thesis examines which of these components are altered in BPD, which are preserved, and under which contextual conditions interpersonal dysfunction emerges.

Projects 1-3 draw on a well-characterized clinical sample of women with BPD and age- and intelligence-matched controls, while Project 4 adds a mechanistic comparison using a rare case of selective bilateral amygdala lesions. Project 1 investigates how individuals with and without BPD evaluate personality traits and update beliefs about others over time using a personality-based social learning task combined with computational reinforcement-learning models. Project 2 examines cooperative decision-making and partner expectations using a novel repeated interaction paradigm that systematically varies partner dependence (whether the partner's outcome depends on the participant's choice) and personal risk (whether helping entails personal costs), combined with functional magnetic resonance imaging (fMRI). Project 3 assesses allocation behavior and social preferences using a battery of baseline, one-shot economic games and utility-based modeling. Project 4 extends this allocation framework by testing whether selective amygdala lesions are associated with shifts in prosocial motive composition.

Across projects, the central finding is not a generalized unwillingness to cooperate in BPD. Instead, interpersonal dysfunction is best characterized by a set of selective, context-dependent biases that map onto distinct stages of the thesis framework. First, at the stage of evaluating and learning about others, individuals with BPD showed markedly more negative self-evaluations and more negative initial evaluations of others' personality traits, while relying on largely intact learning mechanisms. Computational modeling indicated preserved belief updating, suggesting that group differences primarily reflect negatively biased starting points rather than impaired learning per se. Second, at the stage of cooperative choice, cooperative behavior was preserved when a partner's outcome depended on the participant's choice. However, individuals with BPD had more pessimistic expectations about others' willingness to help when they themselves depended on a partner. Behavioral and neural group differences in cooperative engagement emerged most reliably under weak external constraints – when

personal costs were low and/or the partner did not depend on help – indicating altered action engagement rather than a broad deficit in cooperative capacity. These effects were accompanied by altered recruitment of lateral prefrontal and orbitofrontal brain regions during decision phases, implicating systems involved in regulatory engagement and integrative valuation rather than core social decision-making or mentalizing circuitry. Third, at the stage of allocating gains, allocation behavior and fairness sensitivity were largely preserved under anonymous, one-shot conditions. Both active (i.e., the willingness to give) and reactive cooperation (i.e., responding to unfairness), as well as social motives (e.g., inequality aversion) and preferences were comparable between women with BPD and controls. Extending this stage mechanistically, preserved prosocial behavior was shown to arise from qualitatively different underlying motives. Selective bilateral amygdala lesions were associated with a shift from inequality aversion toward joint-gain maximization despite intact prosocial orientation, suggesting that the amygdala contributes to the weighting of prosocial motives rather than to prosociality per se.

Together, this thesis proposes a mechanistic account in which interpersonal dysfunction in BPD reflects negatively biased priors and expectation-driven dynamics across stages of cooperation, rather than a global deficit in prosocial motivation. By identifying both vulnerabilities and preserved capacities, the findings offer a principled explanation for heterogeneous results in the literature and highlight clinically relevant targets and opportunities for intervention.

Zusammenfassung

Kooperation ist eine zentrale Grundlage menschlichen Zusammenlebens. Kooperatives Verhalten stärkt soziale Verbundenheit, stabilisiert Beziehungen und ermöglicht das Erreichen gemeinsamer Ziele. Dafür müssen Personen andere einschätzen, deren Verhalten antizipieren und ihre Handlungen über längere Zeiträume koordinieren. Wenn diese Prozesse gestört sind, werden zwischenmenschliche Beziehungen häufig instabil und können schwer aufrechterhalten werden. Die Borderline-Persönlichkeitsstörung (BPS) ist durch anhaltende und tiefgreifende Beeinträchtigungen zwischenmenschlicher Beziehungen gekennzeichnet und stellt daher einen klinisch relevanten Rahmen dar, um zu untersuchen, wie Störungen kooperationsbezogener Prozesse zu instabilen sozialen Interaktionen beitragen.

Diese Dissertation greift dieses Problem auf, indem sie Kooperation nicht als einheitliches Verhalten, sondern als mehrstufigen sozialen Entscheidungsprozess versteht, der sich über zeitliche Etappen entwickelt. Kooperation wird dabei in drei eng miteinander verknüpfte Komponenten gegliedert: (i) Bewertung anderer Personen und soziales Lernen, (ii) kooperative Entscheidungen unter Interdependenz und Unsicherheit sowie (iii) Verteilung von Gewinnen gemäß sozialen Motiven. In vier Projekten wird untersucht, welche dieser Komponenten bei Personen mit BPS verändert sind, welche weitgehend unverändert sind und unter welchen Kontextbedingungen dysfunktionale zwischenmenschliche Beziehungen besonders wahrscheinlich auftreten.

Die Projekte 1-3 basieren auf einer gut charakterisierten klinischen Stichprobe von Frauen mit BPS sowie einer nach Alter und Intelligenz parallelisierten Kontrollgruppe. Projekt 4 ergänzt diesen Ansatz um einen mechanistischen Vergleich anhand eines seltenen Falls einer Person mit selektiven bilateralen Amygdala-Läsionen. Im Einzelnen untersucht Projekt 1 mittels einer persönlichkeitsbasierten sozialen Lernaufgabe und mithilfe computationaler Verstärkungslernmodelle (*Reinforcement-Learning-Modelle),* wie Personen mit und ohne BPS Persönlichkeitsmerkmale anderer einschätzen und ihre Überzeugungen im Zeitverlauf anpassen. Projekt 2 analysiert kooperatives Entscheidungsverhalten und Erwartungen an Interaktionspartnerinnen in einem neu entwickelten Paradigma, das bei wiederholten Interaktionen systematisch variiert, ob das Ergebnis der Partnerperson von der Entscheidung der Versuchsperson abhängt und ob Helfen mit persönlichen Kosten verbunden ist. Dieses Projekt untersucht dabei auch die neuronalen Korrelate mittels funktioneller Magnetresonanztomographie (fMRT). Projekt 3 erfasst Verteilungsverhalten und soziale Präferenzen anhand einer Batterie von basalen Einmal-Entscheidungsspielen (*one-shot games*) der experimentellen Ökonomie sowie anhand ökonomischer Modelle zur Nutzenmaximierung. Projekt 4 erweitert dieses Rahmenmodell, indem geprüft wird, ob selektive Amygdala-Läsionen mit Veränderungen in der Zusammensetzung verschiedener prosozialer Motive einhergehen.

Über alle Projekte hinweg zeigt sich als zentrales Ergebnis, dass die BPS *nicht* durch eine generelle Unwilligkeit zur Kooperation gekennzeichnet ist. Vielmehr lässt sich die zwischenmenschliche

Dysfunktion am besten als Muster selektiver, kontextabhängiger Verzerrungen beschreiben, die jeweils unterschiedlichen Stufen des hier vorgeschlagenen Rahmenmodells zuzuordnen sind.

Erstens: Auf der Stufe der Bewertung und des sozialen Lernens zeigten Personen mit einer BPS deutlich negativere Selbstbewertungen und negativere Einschätzungen der Persönlichkeitsmerkmale anderer, bei zugleich weitgehend intakten Lernmechanismen. Die computationalen Modelle sprachen für eine erhaltene Aktualisierung von Überzeugungen (*belief updating*), sodass Gruppenunterschiede vor allem negativ verzerrten Ausgangsannahmen widerspiegeln und weniger ein Lerndefizit im engeren Sinne.

Zweitens: Auf der Stufe der kooperativen Entscheidungen war kooperatives Verhalten unverändert, wenn das Ergebnis der Partnerperson von der Entscheidung der Versuchsperson abhing. Personen mit BPS erwarteten jedoch stärker als Kontrollpersonen, dass andere weniger bereit wären zu helfen, wenn sie selbst auf die Partnerperson angewiesen waren. Unterschiede im kooperativen Engagement auf Verhaltens- und neuronaler Ebene traten besonders unter Bedingungen geringer externer Einschränkungen auf, also dann, wenn die eigenen Kosten gering waren und/oder die Partnerperson nicht auf Hilfe angewiesen war. Dies spricht eher für eine veränderte Handlungsbereitschaft als für ein globales Defizit kooperativer Fähigkeiten. Begleitend zeigte sich während der Entscheidungsphase eine veränderte Rekrutierung lateraler präfrontaler und orbitofrontaler Hirnregionen, was eher für Veränderungen in der regulatorischen Beteiligung und der integrativen Berechnung des subjektiven Wertes (*subjective value*) hindeutet als auf Störungen zentraler Prozesse sozialer Kognition (z.B. Mentalisierung).

Drittens: Auf der Stufe der Gewinnverteilung blieben Verteilungsverhalten und Fairnesssensitivität unter anonymen Einmal-Entscheidungsspielen weitgehend unverändert. Sowohl aktive Kooperation (die Bereitschaft zu geben) als auch reaktive Kooperation (Reaktionen auf unfaires Verhalten) sowie soziale Motive (z.B. Ungleichheitsaversion) und soziale Präferenzen waren zwischen Frauen mit BPS und Kontrollpersonen vergleichbar. Eine mechanistische Erweiterung dieser Stufe zeigt jedoch, dass vergleichbares prosoziales Verhalten auf qualitativ unterschiedlichen Motiven beruhen kann: Bei insgesamt erhaltener prosozialer Orientierung, gingen selektive bilaterale Amygdala-Läsionen mit einer Verschiebung von Motiven der Ungleichheitsaversion hin zu Motiven der Maximierung gemeinsamer Gewinne einher. Dies legt nahe, dass die Amygdala weniger Prosozialität an sich vermittelt, sondern vielmehr die Gewichtung verschiedener prosozialer Motive beeinflusst.

Insgesamt entwickelt diese Dissertation ein mechanistisches Rahmenmodell, nach dem zwischenmenschliche Dysfunktionen bei der BPS vor allem durch negativ verzerrte Vorannahmen (*priors*) und erwartungsgeleitete Dynamiken über mehrere Stufen kooperativen Entscheidens hinweg geprägt sind und nicht durch ein globales Defizit prosozialer Motivation. Indem sowohl Vulnerabilitäten als auch erhaltene Kompetenzen systematisch herausgearbeitet werden, liefern die Befunde eine schlüssige Erklärung für heterogene Ergebnisse in der Literatur und weisen auf klinisch relevante Ansatzpunkte sowie Chancen für Interventionen hin.

List of publications

This publication-based thesis consists of three papers published in peer-reviewed journals and one manuscript representing ongoing work. I am the first author on all four manuscripts.

The following papers have been published in peer-reviewed journals and are included in this thesis:

I. **Doppelhofer, L. M**., Hurlemann, R., Bach, D. R., & Korn, C. W. (2021). Social motives in a patient with bilateral selective amygdala lesions: Shift in prosocial motivation but not in social value orientation. *Neuropsychologia*, *162*, 108016. https://doi.org/10.1016/j.neuropsychologia.2021.108016

II. **Doppelhofer, L. M.**, Löloff, J., Neukel, C., Herpertz, S. C., & Korn, C. W. (2025). Cooperative decision-making in borderline personality disorder: insights from a preregistered study using a comprehensive economic task battery. *Borderline Personality Disorder and Emotion Dysregulation*, *12*(1), 24. https://doi.org/10.1186/s40479-025-00295-2

III. **Doppelhofer, L. M.**, Perla, R., Frolichs, K. M. M., Rosenblau, G., Herpertz, S. C., & Korn, C. W. (2026). Estimating and learning personality traits of and from women with borderline personality disorder. *Borderline Personality Disorder and Emotion Dysregulation*. *13*(1), 5. https://doi.org/10.1186/s40479-025-00320-4

The manuscripts are labeled I-III according to publication date; in terms of the thesis' conceptual progression, they correspond to Projects 4, 3, and 1, respectively.

In addition, I am the second author on the following publication, which was completed during my PhD but is not included in this thesis:

IV. Kuper-Smith, B. J., **Doppelhofer, L. M.**, Oganian, Y., Rosenblau, G., & Korn, C. W. (2021). Risk perception and optimism during the early stages of the COVID-19 pandemic. *Royal Society Open Science*, *8*(11), 210904. https://doi.org/10.1098/rsos.210904

1 General introduction

1.1 Cooperation as a multi-stage process

Imagine a PhD student at the start of a new collaboration. Before any concrete work begins, they form an initial impression of a potential collaborator. They draw on past experiences with colleagues and on available cues about the other person's reliability and competence and revise this evaluation as new interactions either confirm or challenge their expectations. On the basis of these evolving impressions, the student then decides whether and how to cooperate. They may be considering a project that is exceptionally creative and innovative, but also risky because it is early-stage and potentially time-intensive. Committing fully could lead to a breakthrough paper, but it could also mean investing months or years of work that ultimately does not pay off. Under these conditions, decisions about sharing ideas, dividing tasks, and sustaining effort become tightly interdependent. Finally, once the work has produced results, the partners must decide how to share the benefits, (re-)negotiating how contributions translate into authorship order, recognition, and future opportunities. Ultimately, such cooperation can result in a PhD thesis.

This example illustrates a broader point: cooperation is not a single act but a sequence of interdependent, recursively interconnected processes that unfold over time. The projects in this thesis are structured accordingly, moving from how partners are evaluated and learned about, to cooperation under interdependence and uncertainty, and finally to how rewards are divided. In this thesis, cooperation refers to interdependent choices that shape both one's own and another person's outcomes. Whether such choices are cooperative or self-serving in practice depends on the incentive structure of the situation, but also on how individuals represent interaction partners, update expectations from experience, and value outcomes for self and other. Cooperative behavior is therefore inherently context-dependent and can vary systematically with risk, uncertainty, and the degree of interdependence between partners. Disruptions of any of the above-mentioned processes can impair social functioning and contribute to persistent interpersonal difficulties. Borderline Personality Disorder (BPD) is marked by pervasive interpersonal difficulties (American Psychiatric Association, 2013; Gunderson, 2007; Gunderson et al., 2018) and therefore provides a clinically relevant framework for investigating how these component processes contribute to unstable social interactions.

Accordingly, cooperation is conceptualized here as a multi-stage social process involving (i) constructing and updating evaluations of others' personalities, (ii) cooperating under interdependence and uncertainty, and (iii) deciding how to distribute rewards. These components are addressed across four projects. Projects 1-3 draw on the same well-characterized sample of women with BPD and age- and intelligence-matched controls, with project-specific inclusion and exclusion criteria. **Project 1** focuses on how women with and without BPD evaluate personality traits and learn trait information over time, and on how controls learn about BPD-like versus control-like personality profiles. **Project 2** examines cooperation and partner expectations in women with and without BPD in a novel repeated

interaction paradigm and links choice phases to prefrontal recruitment using functional magnetic resonance imaging (fMRI). **Project 3** uses a battery of economic games to assess allocation behavior and social preferences in women with and without BPD under baseline, one-shot conditions. **Project 4** adds a mechanistic comparison by testing whether selective bilateral amygdala lesions in an Urbach-Wiethe patient are associated with shifts in social motives during payoff allocations. In addition, a large dataset of nonclinical participants was included in Project 4 to characterize social motives in a wider population.

Across projects, differences in BPD are treated as potentially selective and context-dependent, rather than reflecting uniform impairments in social decision-making.

1.2 Borderline Personality Disorder and interpersonal dysfunction

BPD is characterized by an unstable self-image, impulsive behavior, emotional instability, and pervasive difficulties in interpersonal relationships (American Psychiatric Association, 2013; Gunderson, 2007). BPD is also a common and clinically burdensome condition, with short-term prevalence estimates around 0.7-2% in community samples and substantially higher lifetime prevalence estimates of around 6% (Grant et al., 2008; Gunderson et al., 2018; Johnson et al., 2008; Torgersen et al., 2001). In psychiatric treatment settings, BPD is markedly overrepresented, comprising roughly 15-23% of patients in clinics or hospitals (Korzekwa et al., 2008; Widiger & Weissman, 1991). Comorbidity is the norm rather than the exception in BPD, with frequent co-occurrence of mood, anxiety, stressor-related, and substance use disorders, as well as other personality disorders and neurodevelopmental disorders like attention-deficit/hyperactivity disorder (Asherson et al., 2014; Pagura et al., 2010; Tomko et al., 2014; Zimmerman & Mattia, 1999). Although historically viewed as persistent and difficult to treat, longitudinal studies indicate that many patients achieve symptomatic remission over time, and evidence-based psychotherapies can reduce symptoms (Bateman & Fonagy, 2009; Clarkin et al., 2007; Linehan et al., 2015). However, symptomatic improvement does not necessarily imply full recovery, and patients with BPD remain more socially dysfunctional than patients with other personality disorders (Gunderson, 2011).

The etiology of BPD is best conceptualized as multifactorial, reflecting interactions between genetic and environmental influences across development (Gunderson et al., 2018). Twin data suggest substantial heritability (reported around 0.70), while family-wide environmental effects appear comparatively small (Torgersen et al., 2000). Adverse childhood experiences – especially maltreatment – show a strong association with BPD and are considered the most significant environmental risk factor, though not a necessary precondition (Ball & Links, 2009; Bozzatello et al., 2021; Hengartner et al., 2013). A notable epidemiological pattern is that prevalence estimates in community-based samples are relatively similar in males and females, whereas clinical settings show an approximately 3:1 female:male ratio (Grant et al., 2008; Korzekwa et al., 2008; Lenzenweger et al., 2007; Tomko et al., 2014). Accordingly, Projects

1-3 focus on women with BPD, reflecting the composition of many treatment-seeking cohorts and supporting comparability with much of the existing clinical literature, while recognizing that this choice constrains generalizability.

The emphasis here is on interpersonal functioning and social decision-making mechanisms in BPD, rather than providing a comprehensive account of the full diagnostic spectrum. In everyday social life, these difficulties often present as fewer social connections that are experienced as more unstable and ambivalent (Bakker et al., 2016; Lazarus et al., 2014; Liebke et al., 2017; Stepp et al., 2009). To better characterize these interpersonal difficulties, empirical studies have investigated whether individuals with BPD differ in how they perceive others and form expectations about them. Work on social perception and social cognition in BPD points to a broad negatively biased processing style. In emotion recognition paradigms, individuals with BPD appear less likely to detect or utilize positive social cues and show heightened sensitivity to negative or ambiguous information (Bertsch et al., 2017; Domes et al., 2009; Thome et al., 2016). Consistently, they tend to form more unfavorable impressions of others, evaluating others more negatively (Barnow et al., 2009; Hepp, Kieslich, Schmitz, et al., 2021; Hepp, Kieslich, Wycoff, et al., 2021; Korn et al., 2016; Scheunemann et al., 2023) and reporting more pessimistic partner expectations, including more selfishness (Lévay et al., 2021) and less trustworthiness (for a review, see Masland et al., 2020). These patterns co-occur with elevated rejection sensitivity and feelings of exclusion even in contexts of explicit inclusion (Cavicchioli & Maffei, 2020; De Panfilis et al., 2015; Domsalla et al., 2014; Liebke et al., 2018; Poggi et al., 2019; Renneberg et al., 2012; Staebler et al., 2011).

From a mechanistic standpoint, interpersonal problems in BPD may arise from differences in (i) evaluating others and updating beliefs about partners over time, (ii) forming expectations and behaving (non-)cooperatively under interdependence, and (iii) weighing fairness, self-interest, and prosocial motives when allocating outcomes. Accordingly, the projects described here target partner learning (Project 1), cooperation and partner expectations (Project 2), and allocation decisions (Projects 3-4) as separable components of interpersonal functioning, using controlled tasks alongside computational and neurobiological frameworks.

1.3 Methods toolbox for isolating social decision processes

This section outlines the methods used across projects and the rationale for why these approaches are informative for studying cooperation in BPD. The section is organized by method type rather than by project order. It first introduces economic games as standardized experimental settings that isolate core features of cooperative behavior. It then describes computational modeling as a complementary approach for interpreting behavior. Within this framework, reinforcement-learning (RL) models are presented as a formal language for describing how partner evaluations and beliefs are updated from experience, whereas utility-based models quantify valuation and different social motives in payoff

allocations. By linking these approaches to the existing literature, the section explains their value for studying mechanisms relevant to social behavior in BPD.

1.3.1 Economic games and controlled interdependence paradigms

To examine the processes underpinning cooperation more closely, it is useful to recreate key elements of everyday cooperative situations within simplified, controlled settings, such as economic games. Economic games are a class of experimental paradigms that emerged from experimental and behavioral economics and are now widely used in psychology and related behavioral sciences to study cooperation (Camerer & Fehr, 2004; Pisor et al., 2020; Thielmann et al., 2021). By abstracting from real-life complexity, they isolate core features of social preferences – such as fairness, altruism, self-interest, reciprocity, and inequality aversion – in interdependent choice situations in which decisions have direct implications for one's own and others' outcomes (Camerer & Fehr, 2004). Well-established examples include the Prisoner's Dilemma, the Trust Game, the Ultimatum Game, the Dictator Game, and public-goods and punishment paradigms, each emphasizing distinct components of cooperation and social preferences. In this thesis, these standard economic games as well as a newly designed interactive and interdependent paradigm are used in Projects 2-4 to probe cooperation and allocation behavior under well-defined incentive structures.

Several studies using economic games suggest that individuals with BPD have difficulty maintaining cooperation or restoring it after disruptions (Hula et al., 2018; King-Casas et al., 2008; Saunders et al., 2015; Unoka et al., 2009). At the same time, findings across tasks are inconsistent. Some studies describe more self-interested or economically "rational" responding in BPD (for a review, see Jeung et al., 2016), whereas other studies report little or no evidence for group differences in prosocial preferences or overall cooperation (Graumann et al., 2025; Lévay et al., 2021; Preuss et al., 2016; Thielmann et al., 2014).

One reason for this heterogeneity is that different games emphasize different components of cooperation. A key distinction separates active cooperation (e.g., the willingness to give or share) from reactive cooperation (e.g., responding to unfairness, refraining from retaliation, re-engaging after norm-violations) (Hepp et al., 2014; Hilbig et al., 2013; Niedtfeld et al., 2020; Thielmann et al., 2014). Borderline features have been linked more consistently to difficulties in reactive cooperation than to reduced active cooperation (Hepp et al., 2014; Thielmann et al., 2014). Accordingly, individuals with BPD often behave similarly to controls in tasks that index active sharing, such as the Social Value Orientation Measurement or the Dictator Game, in which participants actively allocate money between themselves and another person (Lévay et al., 2021; Wischniewski & Brüne, 2013). By contrast, differences are more frequently reported in paradigms such as the Ultimatum Game or the Trust Game, which place greater demands on reactive responding (King-Casas et al., 2008; Seres et al., 2009; Unoka et al., 2009). Moreover, cooperative responses in BPD appear especially sensitive to situational cues and social context. For example, negative emotional primes can reduce appraisals of trustworthiness in

individuals with elevated BPD symptoms (Masland & Hooley, 2020), and disruptions in social behavior may be particularly evident after unexpectedly positive interactions or feedback (Abramov et al., 2020; De Panfilis et al., 2019; Liebke et al., 2018; Niedtfeld & Kroneisen, 2020). Importantly, atypical patterns – particularly in reactive cooperation – are often evident in repeated or dynamic interaction settings, where trust and reciprocity can change across phases of exchange (Abramov et al., 2020; Hula et al., 2018; King-Casas et al., 2008; Seres et al., 2009; Unoka et al., 2009).

1.3.2 Computational models: reinforcement-learning and utility-based approaches

Beyond behavioral descriptions, computational approaches complement experimental paradigms by forcing theories to be specified as explicit mechanisms and by generating testable predictions about how behavior should unfold over time. In this thesis, computational modeling is used in combination with two method families that serve different inferential goals: RL models for belief updating about partners (Project 1) and utility-based social preference models for valuation and motives in payoff allocations (Projects 3-4).

In the RL framework, learning is captured as trial-by-trial updating from prediction errors (PEs) – the discrepancy between an initial expectation and subsequent feedback – such that PEs incrementally adjust future predictions. RL theory offers a general framework that has been applied to a broad range of cognitive and social behaviors (Charpentier & O'Doherty, 2018; Dayan & Niv, 2008; Lockwood & Klein-Flügge, 2020; Niv & Langdon, 2016; Ruff & Fehr, 2014; Sutton & Barto, 2014; Zhang et al., 2020), and it has increasingly been used to investigate mechanisms relevant to psychiatric disorders (Czekalla et al., 2024; Liebenow et al., 2022; Maia & Frank, 2011; Pike & Robinson, 2022; Reiter et al., 2017; Schuermann et al., 2011; Waite et al., 2024). A major advantage of computational models is that they make mechanistic assumptions explicit and yield interpretable parameters that can be related to latent psychological processes, rather than merely summarizing group-level outcomes (Guest & Martin, 2021; Suppes, 1966; van Rooij & Baggio, 2021). In Project 1, RL-based modeling is used to quantify how beliefs about other people are formed and updated over time, by comparing multiple candidate models that implement different generative assumptions about social learning.

In contrast, Projects 3-4 apply utility-based economic models to formalize how payoff allocations are valued and which motives shape that valuation. Here, computational modeling is used to test how strongly different motive components (i.e., own gain, joint gain, inequality aversion) contribute to evaluations of payoff allocations. Concretely, candidate models are implemented as linear regressions predicting each participant's ratings (i.e., the desirability of payoff pairs) from motive-related predictors, following prior model-based approaches in this paradigm (Haruno & Frith, 2010; Liu et al., 2019). Additionally, the design of the conditions in the novel social decision-making task in Project 2 draws on components from both RL and utility-based economic models (but for the sake of brevity these are not presented in details in this thesis).

Across projects, controlled tasks provide standardized measures of cooperation and learning, and computational models provide a formal language for inferring latent mechanisms.

1.4 Stage 1: Learning about others

People vary substantially in their personality traits, making it essential to infer and learn trait information about others in order to navigate social interactions and relationships. Personality traits can be conceptualized as relatively stable patterns in thoughts, emotions, and behavior that generalize across time and situations (Kandler & Bleidorn, 2015). Trait impressions are often shaped through repeated contact and observation, but they can also form quickly – even on the basis of sparse information or without direct interaction (Borkenau & Liebler, 1992; Bott et al., 2024; Jaksic & Schlegel, 2020; Little & Perrett, 2007; Naumann et al., 2009). Project 1 addresses these mechanisms in a personality-based social learning task in which participants repeatedly predict and receive feedback about others' traits, combined with computational modeling of belief updating.

Evaluating and learning about others is closely tied to how individuals view themselves (Albright et al., 1988; Cho & Knowles, 2013; Thielmann et al., 2020; Wood et al., 2010). Negative self-views and maladaptive social evaluation and learning have been proposed as contributors to interpersonal dysfunction across psychiatric disorders, including BPD (Kopala-Sibley et al., 2012; Korn et al., 2016; Niedtfeld et al., 2020; Rosenblau et al., 2023; Van Schie et al., 2024; Werner et al., 2019). Whereas nonclinical individuals tend to maintain positive self-perceptions and show self-related positivity biases (Elder et al., 2023; Korn et al., 2012; Leary, 2007), BPD is associated with self-views that are less positive, more negative, and more inconsistent (Korn et al., 2016; Van Schie et al., 2020; Vater et al., 2015; Winter et al., 2017). In self-report personality trait frameworks, individuals with BPD show higher Neuroticism and lower Agreeableness and Conscientiousness relative to controls (Samuel & Widiger, 2008; Saulsman & Page, 2004). They also score higher across maladaptive personality trait domains compared to nonclinical controls (D'Adda et al., 2025) and compared to clinical samples with other personality disorders (Calvo et al., 2016).

Learning about others requires flexible social representations that can be adjusted as new information becomes available (Rosenblau et al., 2023). Prior work suggests that learning about others' personality traits can depend not only on RL-driven PEs but also on reference points and the granularity of social representations (Frolichs et al., 2022). Reference points reflect aggregated prior knowledge or assumptions about other people and provide a baseline for interpreting new information. Granularity refers to the level of detail that people apply when updating their assumptions: while coarse-grained representations imply broad generalizations across traits (e.g., based on broad Big Five personality factors), fine-grained representations indicate that the distinct relationships between traits (e.g., based on similarity between specific traits) are considered when making new predictions. Using computational models that incorporate reference points and granularity, prior work showed that nonclinical adults

adjust these components flexibly to support context-specific learning (Frolichs et al., 2022). Such models have been proposed as tools to detect social-learning impairments in clinical groups, potentially reflecting inadequate reference points, generalization patterns, or updating (Rosenblau et al., 2021, 2023).

Project 1 applies this modeling framework to test whether individuals with BPD show constrained or negatively biased evaluations and social learning – potentially drawing more strongly on negative self-views as reference points. Learning about a partner's traits can be understood as building a working model of the other person, which guides expectations about cooperation and reciprocity. In interdependent contexts, these expectations shape the perceived meaning and value of helping and may therefore influence cooperative responding even when incentives are held constant. This motivates investigating both trait learning and repeated cooperation paradigms that make expectation-dependent dynamics observable.

1.5 Stage 2: Cooperation, expectations, and prefrontal systems

Cooperative choice involves weighing personal costs against the potential benefits of helping, anticipating how others will behave, and navigating uncertainty about outcomes. Project 2 addresses these mechanisms in a novel repeated, helping-based paradigm that manipulates social context (dependent versus independent partner) and personal risk (high versus low), combined with fMRI and a separate expectation task assessing beliefs about the partner's cooperative behavior. On each trial, participants choose between a Wait option associated with a small but sure loss and a Play option that can yield gains but also potential losses, which are larger than the sure loss of the Wait option. The success probability of the Play option depends on both players' concomitant choices, such that joint choice of the Play option increases the probability of a positive outcome for both players and therefore involves a mixture of self- and other-regarding motives. Critically, the social meaning of Play varies by social context. In dependent-partner trials, choosing Play increased the partner's chance of survival within the task and thus constitutes a helping choice (i.e., more other-regarding). In the independent-partner trials, the partner's survival was not directly contingent on the participant's decision, turning Play into a more mixed-motive option (both self- and other-regarding) without an immediate helping imperative. Throughout the thesis, this manipulation is referred to as social context, defined by whether the partner's survival is contingent on the participant's choice or not.

These task demands draw on processes that have been widely studied in the neuroimaging literature on social decision-making. In nonclinical samples, neuroimaging research consistently identifies the medial and lateral prefrontal systems as core nodes for social choice (Frith & Singer, 2008; Hiser & Koenigs, 2018; Lee & Harris, 2013) and for strategic decisions in economic game settings (Chang et al., 2011; Cutler & Campbell-Meiklejohn, 2019; Decety et al., 2004; McCabe et al., 2001; Schurz et al., 2021). The medial prefrontal cortex (PFC) is a particularly relevant region because it supports both mentalizing

– representing and predicting others' beliefs, intentions, and actions (Amodio & Frith, 2006; Frith & Frith, 2012; Saxe & Powell, 2006) – and the subjective valuation of diverse and abstract rewards (Bartra et al., 2013; Haber & Knutson, 2010; Levy & Glimcher, 2012). Social decisions further recruit broader reward-related systems and regions involved in processing of social and affective signals (Apps et al., 2013; Bhanji & Delgado, 2014; Cutler & Campbell-Meiklejohn, 2019; Lamm et al., 2011).

Although studies combining economic games and fMRI in BPD remain relatively scarce, evidence from social-cognitive and emotion-processing paradigms suggests consistent alterations in neural systems that support social functioning. In particular, findings point to dysregulation in fronto-limbic circuits involved in socio-emotional processing (Mier et al., 2013; Schurz et al., 2024; Tebartz Van Elst et al., 2003). A recurrent pattern is increased limbic reactivity – including amygdala and insula responses – paired with impaired habituation and heightened sensitivity to negative and neutral social stimuli (Donegan et al., 2003; Herpertz et al., 2014; Koenigsberg et al., 2009; Lischke et al., 2017; Minzenberg et al., 2007; Schulze et al., 2019). This limbic hyperreactivity often co-occurs with reduced recruitment of prefrontal and lateral prefrontal regions, consistent with altered top-down regulation (Koenigsberg et al., 2009; Mauchnik & Schmahl, 2010; Minzenberg et al., 2007). Taken together, these findings imply that interpersonal difficulties in BPD likely reflect more than affective dysregulation alone and may also involve alterations in how networks centered on the medial PFC integrate social inference and valuation during interdependent decisions.

In sum, Project 2 uses a repeated and interactive cooperation paradigm to test how partner dependence and personal risk shape cooperative choices and explicit partner expectations in BPD and whether these context-dependent decisions are accompanied by altered recruitment of (medial) prefrontal systems. Cooperation shapes the (joint) outcomes of interaction partners and may generate gains and therefore also raises the questions of how these gains should be distributed. Conversely, stable social preferences provide a valuation framework that can shape willingness to cooperate in the first place. This motivates a transition from studying cooperative action to studying allocation decisions that formalize how payoffs are shared and valued.

1.6 Stage 3: Sharing the gains: allocations, fairness, and social preferences

In the context of this thesis, "sharing the gains" refers to allocation decisions that determine how payoffs are distributed between self and other and to the related social preferences. Project 3 addresses these mechanisms in a battery of baseline, one-shot economic games and fairness-related choices, combined with utility-model-based analyses.

A key construct in allocation decisions is fairness perception, which is in turn closely tied to inequality aversion, defined as a preference against unequal outcomes (Fehr & Schmidt, 1999). Based on theoretical accounts emphasizing fairness and justice concerns in BPD (Lis et al., 2018), heightened

inequality aversion is a plausible hypothesis. However, inequality aversion has not been directly tested in BPD in a systematic way. At the same time, several studies report null differences between BPD and controls in fairness perception (Franzen et al., 2011; Jeung et al., 2020), which underscores the need for carefully controlled baseline task variants and mechanistic modeling.

Beyond clinical comparisons, understanding prosocial allocation also requires attention to the motivational structure of prosocial preferences. Social groups can function because individuals both cooperate and share outcomes. People still differ in social motives, and these motives shape how they evaluate and allocate payoffs (MacCrimmon & Messick, 1976). Economic and psychological models describe allocation preferences as depending on self-gain as well as on social motives such as maximizing joint gains and minimizing inequality (Fehr & Schmidt, 1999; Murphy et al., 2011). Critically, similar allocation choices can arise from qualitatively different underlying motives, such that apparently intact fairness or prosocial behavior at the surface level may conceal differences in how outcomes for self and other are valued. Neuroimaging studies link individual differences in amygdala responses to social motives during allocation evaluation (Haruno & Frith, 2010; Liu et al., 2019).

Project 4 addresses the neural basis of social motives in allocation decisions by examining how selective bilateral amygdala lesions shape Social Value Orientation and allocation preferences, using a rare Urbach-Wiethe case combined with matched controls and broader population-level reference samples. Whereas prior neuroimaging work relates amygdala activity to social motives in a correlational manner (Haruno et al., 2014; Haruno & Frith, 2010; Liu et al., 2019), lesion-based evidence provides a complementary test of whether the amygdala is causally involved in weighting different prosocial motives during allocation. The Urbach-Wiethe case provides a mechanistic addition to the clinical comparison by targeting a neural component repeatedly implicated in social motive representation during allocation evaluation. Rather than focusing on diagnostic differences, this comparison isolates whether selective bilateral amygdala lesions are associated with shifts in social motivation. Together with additional nonclinical samples, this approach strengthens the thesis-wide emphasis on decomposing cooperation into separable components and situates individual and clinical findings within broader population-level variability.

Together, Projects 3 and 4 implement a set of complementary allocation and evaluation tasks that formalize how payoffs are valued and shared. Project 3 primarily targets allocation behavior and social preferences, using baseline variants of standard economic games to assess how payoffs are distributed, while delineating active sharing (e.g., active cooperation) from reactive responses to allocations (e.g., reactive cooperation) under minimal contextual demands. In addition, Social Value Orientation and Joint Payoff Evaluation tasks are used to decompose allocation behavior into distinct motivational components using utility-based models. Project 4 builds on this task framework by examining the same core constructs in a mechanistic lesion model, combining the same Social Value Orientation and Joint Payoff Evaluation tasks used in Project 3 with a rare case of selective bilateral amygdala damage and

broader population reference samples. By integrating clinical comparisons with lesion-based evidence, this stage situates allocation preferences and their motivational structure within a neurobiological context. Whereas the preceding stages focus on learning about partners and selecting cooperative actions under interdependence and uncertainty, this final stage addresses how payoffs are evaluated and distributed within a theoretically defined, multi-stage framework, and highlights how differences in interpersonal functioning may arise not only from whether individuals cooperate, but from how social value is computed when outcomes must be shared.

1.7 Thesis approach and contributions

This thesis combines established measures from personality and social decision-making research with novel, interactive paradigms and rigorous analytic approaches to delineate candidate mechanisms of social functioning in clinical populations. Across projects, I integrated controlled experimental tasks with advanced statistical approaches to yield convergent evidence of partner learning, cooperative choice, and allocation preferences. A further cornerstone of the thesis is a strong commitment to open science, aimed at strengthening transparency, reproducibility, and cumulative inference in clinical decision-making.

Key strengths of the thesis include:

1. Well-characterized sample: A comparably large, diagnostically confirmed, and clinically well-characterized sample of women with BPD, alongside age- and intelligence-matched controls
2. Multi-method convergence: Integration of validated questionnaires and established economic games with a novel, complex interactive cooperation task
3. Multi-method evidence: Combination of self-report, behavioral experiments, and fMRI to link computational and psychological constructs to neural implementations
4. Internal replication and robustness: Use of baseline variants of economic tasks to reduce contextual confounds, complemented by a mini meta-analysis to contextualize findings within prior evidence
5. Analytic rigor: Consistent use of statistical approaches, including multilevel (hierarchical) models, Bayesian equivalence testing, Bayesian model comparison, and computational modeling
6. Rare mechanistic comparison: Inclusion of a rare Urbach-Wiethe patient with selective bilateral amygdala lesions to test specificity of prosocial motive components
7. Open science and transparency: Preregistered hypotheses and methods for all BPD projects, and public sharing of data and code for all published projects, with remaining datasets to be released upon publication

1.8 Open science contributions

Preregistrations and open materials are available at the following repositories:

Project 1 (Estimating and learning personality traits of and from women with borderline personality disorder)
Preregistration: https://osf.io/jwky9
Data and code: https://github.com/dnhi-lab/PersonalityLearningBPD_2025

Project 2 (Social decision-making and expectations in women with borderline personality disorder)
Preregistration: https://osf.io/sd2tm
Data and code: to be shared upon publication

Project 3 (Cooperative decision-making in borderline personality disorder)
Preregistration: https://osf.io/jkvh8
Data and code: https://github.com/dnhi-lab/CooperativeDecisionsBPD_2025
Archived release: https://zenodo.org/records/15322897

Project 4 (Social motives in a patient with bilateral selective amygdala lesions)
Data and code: https://github.com/LisaDoppelhofer/SocialMotives_2021

2 Aims

Across four projects, I examined cooperation as a multi-stage social process: (1) forming and updating models of other people's personalities, (2) making cooperative choices under interdependence and uncertainty, and (3) distributing gains according to social motives.

Specifically, I aimed to investigate whether:

I) women with BPD show negatively biased evaluations of their own and others' personality traits and altered learning about others' personalities (Project 1).

II) cooperation in BPD is selectively impaired as a function of social context and personal risk, and whether this is associated with altered prefrontal mechanisms as well as expectations about the partner's cooperative behavior (Project 2).

III) women with BPD differ from controls in economic exchange decisions and social preferences in anonymous, one-shot economic games (Project 3).

IV) selective bilateral amygdala lesions shift social preferences and underlying motives (Project 4).

3 Study I: Estimating and learning personality traits

Estimating and learning personality traits of and from women with borderline personality disorder. *Borderline Personality Disorder and Emotion Dysregulation, 13*(1),5 (2026)

Lisa M. Doppelhofer, Raphael Perla, Koen M. M. Frolichs, Gabriela Rosenblau, Sabine C. Herpertz, & Christoph W. Korn

Doppelhofer *et al.*
Borderline Personality Disorder and Emotion Dysregulation (2026) 13:5
https://doi.org/10.1186/s40479-025-00320-4

Borderline Personality Disorder and Emotion Dysregulation

RESEARCH Open Access

Estimating and learning personality traits of and from women with borderline personality disorder

Lisa M. Doppelhofer[1,2,3*], Raphael Perla[1,2], Koen M. M. Frolichs[1,2], Gabriela Rosenblau[4,5], Sabine C. Herpertz[1,2] and Christoph W. Korn[1,2]

Abstract

Background Accurately perceiving and learning about others' personalities is crucial for successful social relationships. Borderline Personality Disorder (BPD) is marked by unstable interpersonal dynamics, an unstable and negative self-concept, and a tendency to evaluate others unfavorably. Individuals with BPD are also more likely to be viewed negatively by others and experience social stigmatization. This study investigated whether women with BPD exhibit negativity biases when evaluating their own and others' personality traits, and how these biases influence learning about others.

Methods Thirty women with BPD and thirty-one age- and intelligence-matched controls estimated and learned the personality of six individuals (learning profiles) by predicting their self-ratings on 40 personality traits, balanced across the Big Five personality dimensions. After each prediction, participants received feedback with the target's actual rating, allowing participants to gradually learn each profile's personality. Three profiles reflected BPD group personality patterns and three profiles reflected control group patterns. Crucially, participants were told that these profiles were from real individuals but were unaware of their clinical status.

Results As hypothesized, the BPD group rated themselves more negatively than controls, both at the trait level and on standardized personality measures (NEO-FFI, PID-5-BF). When evaluating others, both groups rated BPD profiles similarly, but the control group rated control profiles more favorably. For both groups, accuracy improved slightly over time for control profiles but not BPD profiles, suggesting that BPD trait patterns are inherently harder to learn. Computational modeling indicated that both groups used fine-grained learning strategies—regardless of profile type—with no credible group differences in learning mechanisms.

Conclusion These findings demonstrate that individuals with BPD exhibit pervasive negativity biases in self- and other-evaluations, yet retain intact social learning capacities. This suggests that interpersonal difficulties in BPD may stem more from negatively biased expectations than from deficits in learning. Moreover, the inherent difficulty in learning about BPD-like personality profiles—observed in both groups—may hinder mutual understanding and

*Correspondence:
Lisa M. Doppelhofer
lisamarie.doppelhofer@med.uni-heidelberg.de

Full list of author information is available at the end of the article

© The Author(s) 2025. **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/.

contribute to persistent social challenges. Importantly, the intact learning capacity points to a valuable therapeutic resource: targeting negative social expectations may help to reduce bias and improve social functioning in BPD.

Keywords Borderline personality disorder, Personality traits, Reinforcement learning, Computational modelling

Introduction

People differ widely in their personalities, which makes the ability to estimate and learn others' personality traits crucial for navigating social interactions and relationships. Personality traits are commonly defined as consistent patterns of thoughts, feelings, and behaviors that remain relatively stable over time and across different situations [1]. Traits of others are typically inferred through repeated interactions and observation. But impressions can also be formed rapidly and spontaneously, often based on minimal information and even without direct contact [2–8].

The process of evaluating and learning others' personality traits is closely intertwined with individuals' self-views [9–12]. Importantly, negative self-views combined with maladaptive social estimation and learning processes are thought to contribute to interpersonal dysfunction across various psychiatric disorders, including Borderline Personality Disorder (BPD) [13–18]. BPD is characterized by impulsivity, intense emotional instability, and difficulties in interpersonal relationships [19, 20]. These social difficulties often manifest as fewer, more unstable, and ambivalent interpersonal connections [21–24]. In addition, people with BPD frequently experience stigmatization—not only from the general public but also within the healthcare system [25–28]. One potential contributor to these interpersonal challenges may be atypical processes in how individuals with BPD evaluate and learn about others' personalities, as well as how others evaluate and learn about them. However, empirical and analytical approaches to investigate such maladaptive social learning mechanisms are surprisingly scarce and heterogeneous. In the present study, we provide a model-based framework to test how women with BPD perceive themselves and how they learn about the personalities of individuals with and without BPD. This framework also allowed us to examine how nonclinical control participants learn about the personalities of individuals with and without BPD.

While nonclinical populations typically maintain positive self-perceptions and self-related positivity biases [29–32], individuals with BPD tend to hold less positive, more negative, and more inconsistent self-views [14, 33–35]. This pattern is also reflected in self-report measures of the Big Five personality dimensions [36, 37]: patients with BPD score higher on Neuroticism and lower on socially desirable dimensions such as Agreeableness and Conscientiousness compared to controls [38, 39]. Similarly, on questionnaires assessing maladaptive personality trait domains, individuals with BPD have been shown to score higher across all domains than healthy controls [40] and clinical samples with other personality disorders [41]. Compared to controls, individuals with BPD focus more on negative self-relevant information and are particularly prone to internalize negative feedback from others [14, 33]. Importantly, individuals with BPD are also perceived more negatively by others. After a brief interaction, healthy controls tend to rate the personality traits of people with BPD more negatively than those of other healthy controls [14]. Individuals with BPD are rated higher in depression, emotional instability, and anxiety based on photographs compared to healthy controls [42], and are judged as less trustworthy, and less likeable based on short video clips [27, 43, 44].

Further evidence suggests that individuals with BPD struggle to recognize positive social signals and show hypersensitivity to negative social signals [20, 45]. In facial emotion recognition tasks, they exhibit a heightened sensitivity to threat-related cues [45–47] and deficits in recognizing neutral and positive facial expressions [48–50]. When evaluating others, they tend to view them as more negative, inconsistent, and aggressive compared to controls [14, 43, 51–53]. They also anticipate greater selfishness [54] and lower trustworthiness in others [55]. While previous studies have highlighted negativity biases in individuals with BPD, it remains unclear whether social estimation and learning processes are biased—and how such biases can be modelled.

Adaptive social estimation and learning require flexible representations and continuous adjustments of complex social knowledge structures [16]. Computational modeling offers a powerful approach for advancing research on these processes because it requires researchers to formalize theories, specify mechanisms, and derive precise predictions that can be explicitly tested [56, 57]. In this way, computational models move beyond descriptive accounts by providing explanatory, mechanistic frameworks with psychologically interpretable parameters [56, 58, 59].

Within this framework, reinforcement learning (RL; [60]) models have emerged as a particularly robust and parsimonious class of models for describing how humans learn from trial and error. In RL algorithms, learning is defined as a stepwise reduction of prediction errors (PEs)—the difference between an initial estimate (or prediction) and the subsequent feedback (or outcome). These PEs are used to step-by-step update future predictions. Crucially, these error signals correspond closely to neural responses in reward-related brain circuits [61–65],

providing strong evidence for the validity and importance of RL theory as a mechanistic account of human learning. RL theory has proven to be a powerful framework in explaining a wide range of cognitive and social tasks [59, 60, 66–71], and has increasingly been applied to elucidate key aspects of psychiatric disorders [72–79]. Previous studies have demonstrated that RL captures how humans learn about others' intentions [80], emotions [81], trustworthiness [82], and preferences [83]. Moreover, combining RL algorithms with conceptual knowledge highlights how humans manage the complexity of social environments: rather than relying solely on costly trial-and-error computations, people draw on abstract representations and prior expertise to generalize flexibly while minimizing cognitive effort [84].

Our own work has indicated that, in addition to RL, at least two additional concepts play an important role when learning about others' traits or preferences: reference points and granularity [85]. Reference points capture prior knowledge or assumptions about average members of specific groups and serve as a baseline when encountering new individuals. For example, when meeting a medical student for the first time, people may draw on their knowledge of the "average medical student" to generate initial predictions about that person's personality. Granularity, in turn, refers to the level of detail that people apply when updating these predictions after receiving new information. In our framework, updates are guided by the Big Five personality factors [36, 37]. A coarse-grained representation updates only at the level of broad factors. For instance, if the medical student is observed to be very organized, this information would raise the estimate of their overall Conscientiousness, which in turn increases expectations for all traits within that factor (e.g. reliability, diligence) equally. A fine-grained representation updates predictions at the level of individual traits, weighted by their similarity to the newly observed trait. In the same example, learning that the student is organized would strongly increase expectations about closely related traits such as diligence or reliability. It would also influence more distant traits, such as empathy from the Agreeableness factor, though to a much smaller extent. While the fine-grained model allows for more nuanced and flexible generalizations, it is also more cognitively complex. The coarse-grained model, in contrast, is more parsimonious and resource-efficient, and may be particularly adaptive when information about a new person is sparse or when only broad impressions are needed [85]. Recent studies confirm that people indeed exhibit conceptual trait knowledge, actively use it when making social judgments, and that these knowledge structures have neural correlates [86, 87]. Building on this foundation, our own work using computational models incorporating these two concepts, has shown that the learning behavior of non-clinical adults was best described by reference points and levels of generalization that were appropriately adjusted when learning about others [85]. We proposed that such models may also help identify impairments in social learning in clinical populations, which may stem from insufficiently adjusted reference points or generalizations when forming personality inferences [16]. Supporting this view, a previous study has shown that adolescents with autism spectrum disorder relied on their own preferences as reference points and showed reduced updating during the learning process [83].

Here, we use this modeling framework to test whether individuals with BPD show characteristic restrictions and negative biases in social learning. We adapted a social learning task from our previous studies [14, 31, 83, 85, 88] to investigate potential differences between individuals with BPD and controls in estimating and learning others' personality traits. Specifically, we examined how women with and without BPD (i.e., controls; CON) learned about the personality traits of both BPD and control groups. We proposed that individuals with BPD, would rely on their (negative) self-views as reference points and on overgeneralized coarse-grained representations [16]. In addition to reference points and granularity, we test extended model variants incorporate hierarchical estimation to recover the initial expectation for each trait prior to any learning (referred to as V0 estimates; see Methods section and Results section). These parameters capture participants' prior beliefs about average trait expressions and are estimated from the observed learning dynamics rather than specified a priori.

This study was preregistered, and we briefly summarize the preregistered hypotheses (https://osf.io/jwky9). The hypotheses are numbered [H1] to [H6] for easier reference throughout the article. In the methods section, we provide more technical details on these hypotheses. We predicted that individuals with BPD would rate themselves [H1] more negatively on trait adjectives and personality questionnaires (NEO-FFI, [36, 89]; PID-5-BF, [90–92]) than controls. In addition, we expected that the BPD group compared to the control group would also rate others more negatively on trait adjectives [H2]. Furthermore, we anticipated that controls would more accurately estimate the personality profiles of other controls compared to individuals with BPD [H3.1], and that individuals with BPD would demonstrate greater accuracy when estimating the profiles of other BPD participants than those of control participants [H3.2]. Based on our social learning framework [16], we hypothesized that the two groups would differ in the cognitive strategies during trait learning: we expected individuals with BPD to rely more on their self-ratings and coarse trait similarities, and controls to use average population ratings and

fine-grained item-level similarities—this pattern holding true regardless of the profile being evaluated [H4]. Additionally, we planned to test if individuals with BPD would exhibit a lower overall learning rate compared to controls [H5] and if they would learn more from negative PEs than from positive ones [H6].

Methods

Study procedures

This study is part of a broader project investigating social decision making in women with BPD, including additional behavioral and fMRI experiments, which are not part of the scope of the current study (see [93] for a published article on social decision-making in women with BPD in an overlapping sample of participants). The project received approval from the local ethics committee of the Faculty of Medicine at Heidelberg University (S-137/2020). The specific rationale, hypotheses and approach of this study were preregistered on the Open Science Framework on May 30, 2022 (https://osf.io/jwky9) prior to data analysis. Any protocol modifications are detailed in the methods section.

Data collection took place between October 2021 and May 2023. Participants in the control group were age-matched during the recruitment. A brief intelligence screening confirmed comparable cognitive skills between the groups (see Participants section). Participants were recruited in Heidelberg and nearby towns. We additionally recruited inpatients at the Department of General Psychiatry, University of Heidelberg. Participants in the control group received a base rate of €50 and those with BPD received a base rate of €60 (to account for the extended diagnostic interview time) for completing all experiments of the larger project. Following a telephone prescreening, eligible participants were invited to the study. During the in-person assessment, participants took part in a diagnostic interview, a brief cognitive task (see Clinical assessment section), and a series of self-report questionnaires using the German online survey platform SoSci Survey [94] (see Supplementary Material). They completed the behavioral tasks of the current study and unrelated behavioral experiments. Prior to participation, they were informed about the study procedures and provided written consent. All participants were fully debriefed at the end of the study.

Participants

Clinical assessments

We used two standardized diagnostic instruments: the Mini-DIPS Open Access [95], a shortened version of the Diagnostic Interview for Mental Disorders, to assess Axis I disorders; and the German version of the International Personality Disorder Examination (IPDE; [96, 97]) to evaluate BPD criteria. To be included in the BPD group, participants had to meet at least four IPDE criteria, allowing for the inclusion of sub-threshold cases. Control participants could meet no more than one IPDE criterion (see Supplementary Material for details). Additionally, all participants completed the mini-q [98], a three-minute speeded reasoning test, to ensure that cognitive skills are matched between the groups. All clinical interviews were conducted by trained raters, with typical sessions lasting approximately three hours.

Sample description

We analyzed all valid datasets for the current study and included 30 individuals with BPD and 31 controls (see Supplementary Material for inclusion and exclusion procedure). The control group was matched for age and cognitive abilities as measured with the mini-q (see Table 1). Participants in the BPD group differed significantly from the control group in their level of education, professional qualification, and employment status (see Table 2). The BPD group consisted of 30 participants, of whom 29 had been formally diagnosed with BPD prior to study enrollment. One participant had only a provisional BPD diagnosis from their psychotherapist, which was confirmed through the IPDE interview. Consistent with the heterogeneity of BPD, participants in the BPD group met diagnostic criteria through different combinations of symptoms in the IPDE. Importantly 73.3% endorsed the criterion of unstable and intense interpersonal relationship, confirming that interpersonal difficulties were present in the majority of cases. In addition, several other interpersonal or identity-related criteria (e.g., affective instability, identity disturbance, efforts to avoid abandonment) were endorsed by more than half of the sample, further underscoring the relevance of the current paradigm for studying social learning processes (the full distribution is provided in the Supplementary Material). In the BPD group, 86.67% of participants reported symptoms of comorbid diagnoses according to the Mini-DIPS interview (see Supplementary Material Table S1), most commonly recurrent major depressive disorder (50%), social anxiety disorder (43.33%), and post-traumatic stress disorder (33.33%). Approximately half of the BPD participants (53.33%) were taking psychotropic medication, primarily antidepressants (see Supplementary Material Table S2). Medication intake information was self-reported by participants.

Experimental procedure

We assessed participants' dimensional personality structure to test if participants in the BPD group evaluated themselves more negatively than participants in the control group. To this end, we assessed the Big Five personality dimensions (Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness) with the NEO

Table 1 Age and mini-q summary table

	BPD		CON		Welch's t-test					Wilcoxon rank sum test		
	M	*SD*	*M*	*SD*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*p*	*d*	*z*	*p*	*r*
Age	26.07	5.72	26.00	5.14	−0.048	57.873	−2.856	2.722	0.962	−0.012	0.196	0.845	0.025
Mini-q scores	34.30	12.18	34.03	8.20	0.100	50.613	−5.087	5.622	0.920	0.026	-0.094	0.925	0.012

Table 2 Demographic characteristics of both groups

	BPD (N = 30)		Control (N = 31)		*p*
	N	%	*N*	%	
Education					
Secondary general school (9 years)	1	3.33	0	0	< **0.005**
Intermediate school (10 years)	6	20	0	0	
High school diploma (13 years)	23	76.67	31	100	
Professional qualification					
Non, in professional training	7	23.33	15	48.39	< **0.001**
Non, not in professional training	4	13.33	0	0	
Apprenticeship	5	16.67	0	0	
Vocational school	9	30	2	6.45	
Technical school	0	0	1	3.23	
University	5	16.67	13	41.94	
Employment status					
Student	5	16.67	21	67.74	< **0.001**
In training	5	16.67	0	0	
Employed	10	33.33	8	25.81	
Unemployed	4	13.33	1	3.23	
Retired or homemaker	2	6.67	0	0	
Other	4	13.33	1	3.23	

p = *p*-values from Fisher's exact test for count data. Significant results in bold font

Five-Factor Inventory (NEO-FFI; [36, 89]). Additionally, we characterized the presence and frequency of five maladaptive personality domains (Negative Affect, Detachment, Antagonism, Disinhibition, Psychoticism) with the Personality Inventory for DSM-5–Brief Form (PID-5-BF; [90–92]).

The behavioral task consisted of two parts: 1) estimating and learning about the personality traits of six individuals, which we call "learning profiles," (Fig. 1) and 2) providing self-ratings on the same traits. During the first part, the learning phase, participants were shown six personality profiles. Each profile consisted of self-ratings on 40 personality traits. These 40 traits were selected to reflect items found in the Big Five personality model, with 20 positive and 20 negative traits (e.g., honest and anxious). Between 7 and 9 traits were shown per Big Five factor (see Supplementary Material Table S3 for the full list of original German and translated English trait words). Each profile was paired with a randomly assigned common German female name and an age within the study sample's range (see Supplementary Material for the specific names and ages).

The three BPD profiles were drawn from self-ratings of participants with BPD from a previous study [14]. These profiles were selected from participants whose trait ratings closely matched their group averages—specifically, within ± 1 SD on most traits (a maximum of six

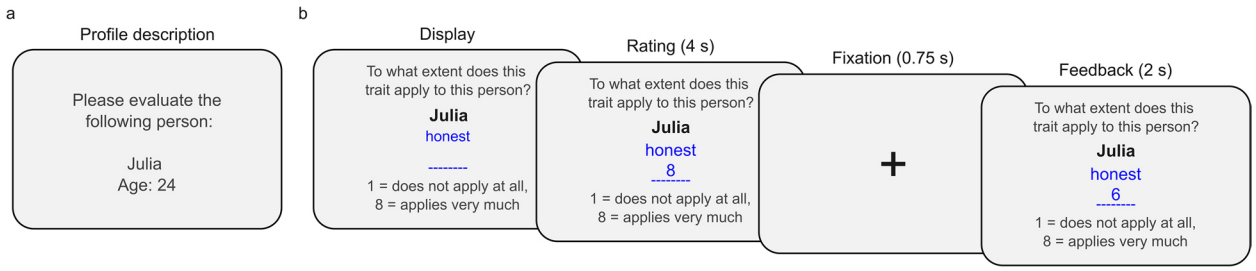


Fig. 1 Personality learning task procedure. **a** Example of a profile description presented at the beginning of a learning block, showing the person's (profile's) name and age. **b** Structure of a single trial within the learning block. On each trial, participants estimated the extent to which a trait word applied to a given person. After providing their rating (4 s), a fixation cross (0.75 s) appeared, followed by feedback (2 s) showing the profile's self-rating on the same trait word. This feedback allowed participants to gradually learn about the person and adjust their estimates for the following traits accordingly. A second fixation (0.75 s) followed the feedback, after which the next trait word was presented (not depicted). Each profile was evaluated on 40 trait words, with a total of six profiles presented (three BPD profiles, three control profiles). Profiles only differed in their names, ages and self-ratings (three reflecting BPD personality patterns and three reflecting control group personality patterns); they did not differ visually. Importantly, participants were not informed that the profiles represented either BPD or control individuals

traits exceeded this range). Due to a deviation from the intended implementation, the control profiles were not selected based on the ± 1 SD criterion. These profiles remained consistent with plausible control group patterns, although they did not strictly adhere to the preregistered selection parameters.

Participants were informed that the profiles represented real individuals from a previous study, but they were not told whether profiles came from a control group or a BPD group. Each profile learning block started with a brief description including the name and age of the person. Participants rated each profile's traits on an 8-point scale (1 = does not apply at all, 8 = applies very much), learning about one profile at a time. For each trial, one trait word (e.g., "honest") was displayed on the screen, and participants had 4 s to estimate the profile's self-rating. Following each estimate, 2 s of feedback were shown with the veridical rating from the person in question, allowing participants to learn about the profile's personality traits over time. Each profile learning block consisted of 40 trials (40 different traits), and all participants evaluated all six profiles (240 trials in total). In the second part of the task, participants rated themselves on the same 40 traits. The on-screen instructions are provided in the Supplementary Material.

The order of the trait words was randomized for each profile. The order of the profiles was CON, BPD, CON, BPD, BPD, CON (except for three participants). The order of the self-ratings was not randomized (similar to many questionnaires) to avoid inter-individual differences that are solely due to the order in which items are presented.

Analyses

Task specific exclusion procedures

As preregistered, we excluded trials with trait words (items) that participants did not understand. Participants could mark these words afterwards on a printout with all words. Eleven participants with BPD and 12 controls marked one or two items. That is, for each of these participants a maximum of two items were excluded. Additionally, for two participants in the BPD group, we chose to exclude one profile each, as they reported that the profile's name influenced their responses (this exclusion criterion was not preregistered).

Data preprocessing

We reverse-coded responses on negative items/traits for all PE, self-rating, and other-rating analyses that did not include item valence (e.g., a rating of '1' for a negative trait was transformed into a positive trait with a recalculated rating of '9 – 1 = 8').

Regression analyses

We used linear mixed effects (LME) models to test the preregistered hypotheses [H1] to [H3.2]. We hypothesized that individuals with BPD would rate themselves more negatively on personality traits and personality questionnaires than controls (NEO-FFI, PID-5-BF) [H1]. Additionally, we expected that individuals with BPD would rate others more negatively on personality traits than controls. That is, the BPD group's estimates of others' personality traits would be numerically smaller for positive trait words and larger for negative trait words compared with the control group's estimates [H2]. The study employed a mixed 2 × 2 design, with two groups (BPD/CON) and two types of learning profiles (BPD/CON). We hypothesized an interaction effect for absolute PEs, with specific expectations that control participants' estimates for control profiles would be closer to the true values (smaller PEs) than BPD participants' estimates [H3.1], and that BPD participants' estimates would be closer to the true values (smaller PEs) when learning

about other BPD participants compared to control profiles [H3.2].

We used the standard alpha $p < 0.05$ criterion for determining significant results, and hence rejecting the null hypotheses. All analyses were conducted in R. Linear mixed effects (LME) models were fitted with the package *lmerTest* [99] by restricted maximum likelihood (REML). We used the *tab_model* function from the *sjPlot* package [100] to report summary tables from the LME models with the Satterthwaite's approximation for computing degrees of freedom for *p*-values, standard errors and confidence intervals (CI). For all LME models we used the BPD group and the BPD profile type as reference categories. We calculated Type III analysis-of-variance (ANOVA) tables for the LME models with Satterwaite's method using the *anova* function in R. Note, that under treatment coding, main effects in the presence of interactions are interpreted relative to the reference levels of the other factors. Post-hoc comparisons were conducted by calculating estimated marginal means (EMMs) from the fitted LME model with the *emmeans* package [101]. All *p*-values for the post-hoc comparisons were adjusted with the *Tukey* method.

We report post-hoc tests for LME and ANOVA results, which reached significance at $p < 0.05$, in the results section, as these address our preregistered hypotheses. Summary tables for the LME models and the Type III ANOVAs are provided in the Supplementary Material.

Self-ratings on trait adjectives To test if individuals in the BPD group rate themselves more negatively than individuals in the control group [H1], we fitted an LME model and examined the effect of *Group* (BPD/CON) on the self-ratings. The model included *Group* as a fixed effect and random intercepts for both the 61 subjects and the 40 items. Ratings on negative items were reverse coded.

LME model formula:

$$\text{Self-ratings} \sim \text{Group} + (1|\ \text{Subject}) + (1|\ \text{Item})$$

In addition, we explored potential group differences between positive and negative items. We tested *Group* (BPD/CON) and *Valence* (positive/negative) and their interaction as fixed effects and random intercepts for subjects and items. Here, we explicitly contrasted the positive versus negative valence of the items (hence did not reverse code ratings on negative items). To investigate whether self-ratings differ between groups for positive or negative items, post-hoc comparisons of EMMs were calculated. This approach estimated the mean response for each *Group* within each level of *Valence*.

LME model formula:

$$\text{Self-ratings} \sim \text{Group} * \text{Valence} + (1|\ \text{Subject}) + (1|\ \text{Item})$$

Self-ratings on personality inventories To test whether the BPD group exhibited more unfavorable self-ratings on the Big Five [36, 37] and maladaptive personality traits [H1], as measured by the NEO-FFI [36, 89] and the PID-5-BF [90–92], we conducted MANOVAs and subsequent post-hoc t-tests. For the MANOVAs, Pillai's trace was computed. Welch's t-tests were performed, and *p*-values were adjusted for multiple comparisons using the false discovery rate (FDR) method. Effect sizes (Cohen's d) were calculated with the *effsize* [102] package.

Other-ratings We hypothesized that individuals with BPD would rate others more negatively—lower on positive traits and higher on negative traits—than controls [H2]. We fitted an LME model and tested the association between ratings of others' personality traits and the effect of *Group* (BPD/CON) and *Valence* (positive/negative), and their interaction. We tested *Group* and *Valence* and their interaction as fixed effects and random intercepts for subjects and items. Ratings on negative items were not reverse coded. To investigate whether other-ratings differ between groups for positive or negative items, we conducted pairwise comparisons.

LME model formula:

$$\text{Other-ratings} \sim \text{Group} * \text{Valence} + (1|\ \text{Subject}) + (1|\ \text{Item})$$

To explore whether other-ratings differed between groups (BPD/CON) depending on *Profile type* (BPD/CON), we fitted an LME model and tested the effects of *Group* and *Profile type*, and their interaction on other-ratings as fixed effects and random intercepts for subjects and items. Ratings on negative items were reverse coded. We calculated pairwise comparisons for *Group* within each level of *Profile type* (BPD/CON) to examine differences between the BPD and control group.

LME model formula:

$$\text{Other-ratings} \sim \text{Group} * \text{Profile type} + (1|\ \text{Subject}) + (1|\ \text{Item})$$

Prediction errors To calculate the PEs, we subtracted the participant's rating/prediction (P) from the profile's self-rating/feedback (F) for each trial (t). We used the absolute PE for all subsequent analyses.

$$PE_t = | F_t - P_t |$$

We expected that the control group would be more accurate than the BPD group in predicting control profiles [H3.1] and that the BPD group would be more accurate in predicting BPD profiles than CON profiles [H3.2]. We set up an LME model and conducted corresponding post-hoc tests. The LME model included the effects of *Group* (BPD/CON), *Profile type* (BPD/CON), and their interaction on PEs and random intercepts for subjects and items. We conducted pairwise comparisons to investigate whether PEs differed between groups across different levels of profile types and between profiles across different levels of group.

LME model formula:

$$\begin{aligned} PE \sim Group * Profile\ type \\ + (1|\ Subject) + (1|\ Item) \end{aligned}$$

In addition, we set up an LME model to test if PEs decrease over time (as a linear effect of trial number), as an indication of learning. We tested the effect of *trial* and random intercepts for subjects and items. The LME model is a more accurate procedure than the initially preregistered method of using Pearson correlation, which disregards the hierarchical nature of our data, potentially resulting in inflated estimates and biased standard errors [103, 104].

LME model formula:

$$PE \sim Trial + (1|\ Subject) + (1|\ Item)$$

To explore whether participants became more accurate in estimating the profiles across trials depending on the group and profile type, we set up an LME model and tested the effects of *Trial*, *Group* (BPD/CON), *Profile type* (BPD/CON), and their interactions on PEs. The model included random intercepts for subjects and items. We conducted pairwise comparisons of EMMs of linear trends to test whether the slopes differed between groups across the different levels of profiles.

LME model formula:

$$\begin{aligned} PE \sim Trial * Group * Profile\ type \\ + (1|\ Subject) + (1|\ Item) \end{aligned}$$

Computational models

We sought to understand how participants with and without BPD learn and adjust their expectations about profiles of people with and without BPD during the task by testing eight different computational models (see model description below). Conventional tests of mean and slope differences detect group effects but leave theoretical mechanisms unexamined (see [105] for a recent treatment). In contrast, computational models embed explicit mechanistic assumptions to bridge theory and statistical inference [56, 57, 106], and yield psychologically interpretable parameters that index latent cognitive processes instead of merely summarizing group-level outcomes [107, 108]. By modeling the full data-generation process, these approaches can distinguish between competing mechanisms that produce similar mean patterns [109, 110], and permit formal comparison of models embodying different theoretical assumptions, thereby driving systematic theory refinement [111, 112]. Motivated by these strengths, we tested eight models adapted from Frolichs et al. [85] organized into three families of generative assumptions.

Based on prior literature, we expected that individuals with BPD would rely more on self-ratings and coarse trait similarity (Model 4, see below), while controls would use population mean ratings and fine-grained similarity (Model 8, see below) [H4]. We expected these models to perform best regardless of the profiles participants learned about. Furthermore, if BPD participants' behavior could be explained by models including learning rates from trial-by-trial prediction errors, we expected that their learning rates for others' traits would be smaller than those of controls [H5]. Finally, if participants' behavior was best explained by models with separate learning rates for positive and negative prediction errors, we had planned to test models with separate learning rates [H6]. However, we refrained from testing [H6] due to limitations in the task design, which we had not adequately considered beforehand.

Below, we provide a brief conceptual overview of the three classes of computational models we employ. A more detailed mathematical explanation is included in the Supplementary Material.

No learning models In this category, participants are assumed not to update their beliefs over time. Instead, their responses rely on fixed reference points: either the average value observed in the overall population or the individual's own trait expressions. These reference points act as "default predictions" and remain unchanged despite new information.

Coarse granularity models These models propose that participants learn in a generalized, overarching way. Rather than tracking every individual trait, they update their overall expectations across broad personality dimensions, specifically the Big Five. When feedback differs from their expectations, participants adjust their beliefs

in the direction of the feedback. The speed of this adjustment is allowed to vary across individuals. The models also account for variability in responses by incorporating a degree of randomness or noise. Within this category, one model operates without any reference points, while others use either the population average or the individual's own trait expression as a reference point.

Fine granularity models The Fine Granularity Models provide the most detailed representation of the learning process, assuming that participants maintain distinct expectations for each individual trait. These models take into account the interconnections between traits, where learning about one trait can influence expectations about related traits. To achieve this, the models incorporate a set of known relationships between traits that were established in previous research [14, 31]. Similar to the coarse granularity models, the fine granularity models include versions that balance learning with reliance on a reference point (population or self) by weighting each source of information accordingly.

Model fitting We used the Stan programming language [113] and its R interface [114] to fit the models to the participants' data. Stan was chosen primarily for its strength in hierarchical estimation, which typically provides more accurate parameter estimates than non-hierarchical methods [71, 115]. In doing so, we considerably improved on our previous model fitting approach, in which models were fitted to each participant's data sequentially. Moreover, the Bayesian framework in Stan provides full uncertainty quantification for every parameter, facilitating direct comparisons of group-level effects (e.g., in initial expectations or learning rates) by sampling from the posterior difference-distribution of their estimated population means and examining coverage metrics such as the 95% credible interval. The technical details of our model fitting procedure, including specific algorithms and diagnostic criteria, are provided in the Supplementary Material.

Model comparison

Following Vehtari et al. [116], we compared the fit of our models by assessing how well they predicted data from new, unseen participants. Specifically, we employed a cross-validation approach (called leave-one-group-out-cross-validation, LOGO-CV) in which we systematically withheld one participant's data at a time, trained the model on the remaining participants, and then assessed its ability to predict the held-out data. In other words, each participant's data were excluded once, and the model's predictions for that participant were compared to the observed values. To quantify predictive accuracy, we averaged the squared errors across all participants, yielding an overall mean squared error (MSE) score that reflects each model's ability to generalize to new data.

To estimate the uncertainty in model comparisons, we calculated the standard errors of the differences between model scores. A model was considered a better fit than another if its performance difference exceeded the associated standard error by at least a factor of two. This criterion is approximately equivalent to the decision boundary for statistical significance commonly used in frequentist statistics [117]. A more detailed and mathematically thorough description of our model comparison approach is provided in the Supplementary Material.

Results

Regression analyses

Self-ratings on trait adjectives

In line with our hypothesis [H1] that individuals with BPD have more negative self-views than controls, we found that they rated themselves less favorably on the 40 personality traits. We found that the control group gave more favorable trait ratings, indicated by a significant effect of *Group* on self-ratings in the LME model ($b = 0.77$, 95% CI [0.56, 0.99], $SE = 0.11$, $t_{(58.76)} = 7.16$, $p < 0.001$; see Supplementary Table S4 for the model summary and Supplementary Table S5 for Type III ANOVA summary).

We additionally explored effects of trait valence (see Supplementary Tables S6-8 for full model summaries and post-hoc contrasts). Individuals with BPD rated themselves less positively on positive traits ($M_{BPD} = 5.79$, $SD_{BPD} = 0.582$, $M_{CON} = 6.28$, $SD_{CON} = 0.351$; EMMs contrast: $b = -0.49$, 95% CI [−0.71, −0.27], $SE = 0.11$, $t_{(121.67)} = -4.44$, $p < 0.001$) and more negatively on negative traits ($M_{BPD} = 3.89$, $SD_{BPD} = 0.653$, $M_{CON} = 2.84$, $SD_{CON} = 0.576$; EMMs contrast: $b = 1.05$, 95% CI [0.84, 1.27], $SE = 0.11$, $t_{(120.64)} = 9.57$, $p < 0.001$) compared to controls (Fig. 2).

Self-ratings on personality inventories

In line with our hypothesis [H1] that individuals with BPD rate themselves more negatively than controls on personality trait questionnaires, we found that participants with BPD rated themselves less favorably on the personality inventories NEO-FFI and PID-5-BF. We found a significant main effect of *Group* on the personality dimensions of the NEO-FFI ($V = 0.77$, $F(5,55) = 36.1$, $p < 0.001$). Post-hoc tests, corrected for multiple comparisons, revealed that individuals with BPD scored significantly higher than controls on Neuroticism, and significantly lower on Extraversion, Agreeableness, and Conscientiousness. We found no significant group difference in the Openness factor. The full MANOVA model summary and the post-hoc comparisons are reported in Supplementary Tables S9-10.

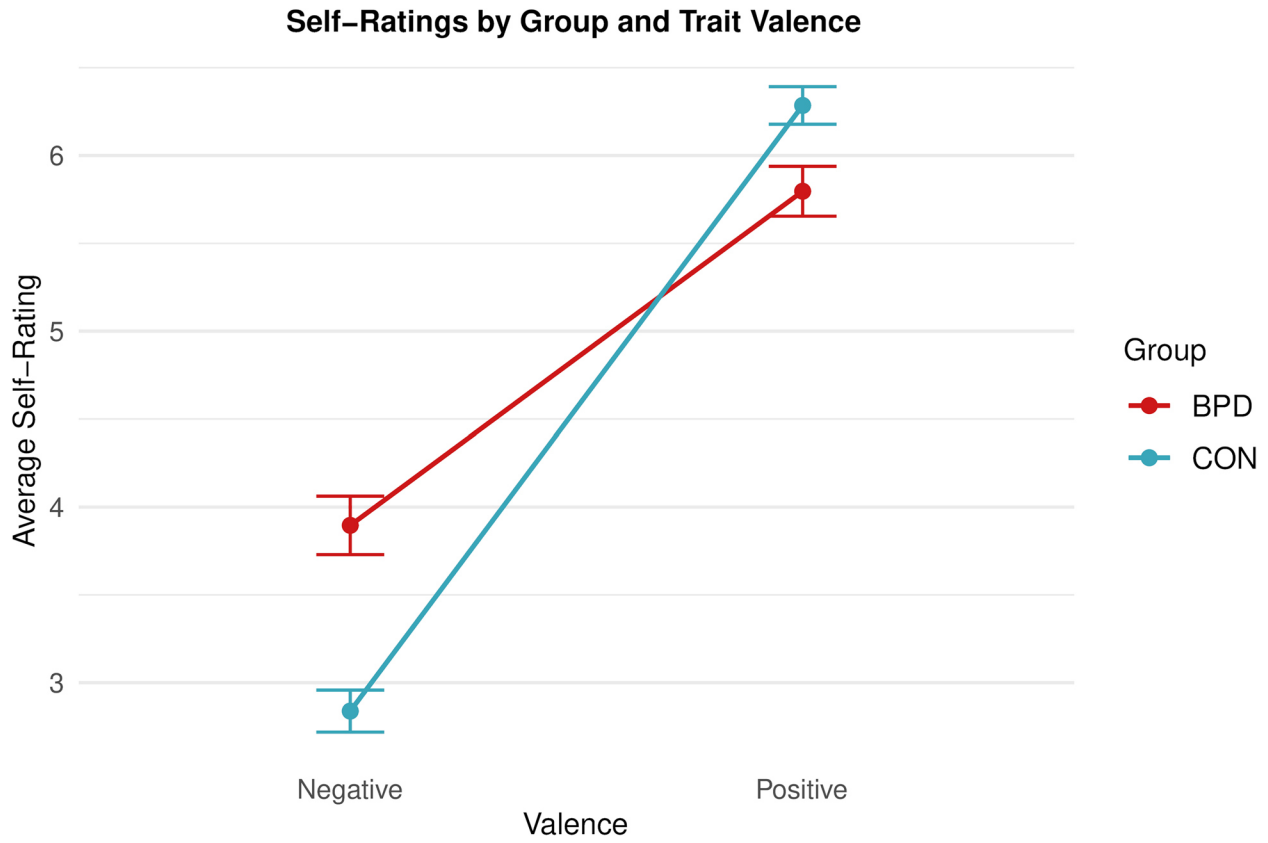


Fig. 2 Self-ratings on positive and negative trait words in the BPD group (red) and control group (blue). The graph indicates that the BPD group gave more unfavorable self-ratings, that is higher ratings on negative traits and lower ratings on positive traits (significant Group x Valence interaction in an LME model). Lines show observed means, error bars indicate 95% confidence intervals based on a normal distribution. Ratings on negative items are not reverse coded

Similarly, we found a significant main effect of *Group* on the PID-5-BF domain scores (V = 0.68, F(5,55) = 23.86, p < 0.001). After correcting for multiple comparisons, individuals with BPD scored significantly higher than controls across all five maladaptive trait domains of the PID-5-BF. A detailed model summary and post-hoc results are provided in Supplementary Tables S11-12.

Other-ratings

In contrast to our hypothesis [H2], the BPD and control groups did not differ in how favorably they rated others across profile types in the learning task (see Supplementary Tables S13-15 for full model summaries and post-hoc contrasts). Pairwise comparisons of group differences within each level of valence revealed no significant effects. Specifically, there was no significant group difference for negative items (M_{BPD} = 3.78, SD_{BPD} = 1.48; M_{CON} = 3.64, SD_{CON} = 1.41; EMMs contrast: b = 0.14, SE = 0.07, z = 1.83, p = 0.067) or for positive items (M_{BPD} = 5.67, SD_{BPD} = 1.33; M_{CON} = 5.76, SD_{CON} = 1.27; EMMs contrast: b = −0.09, SE = 0.07, z = −1.22, p = 0.221).

To explore whether potential group differences were driven by a specific profile type, i.e., who was rated in the learning task, we tested *Group* (BPD/CON) and *Profile type* (BPD/CON) on other-ratings (see Supplementary Tables S16-18 for full model summaries and post-hoc contrasts). We found that the control group shows more favorable other-ratings for control profiles only (M_{BPD} = 5.48, SD_{BPD} = 1.44; M_{CON} = 5.73, SD_{CON} = 1.32; EMMs contrast: b = −0.24, SE = 0.08, z = −3.18, p < 0.01) but not for the BPD profiles (M_{BPD} = 5.41, SD_{BPD} = 1.40; M_{CON} = 5.40, SD_{CON} = 1.37; EMMs contrast: b = 0.01, SE = 0.08, z = 0.17, p = 0.863; see Fig. 3).

Prediction errors

In line with our hypothesis [H3.1], we found that participants in the control group were more accurate in estimating the personality traits for control profiles than participants in the BPD group (Supplementary Tables S19-21 for full model summaries and post-hoc contrasts). Pairwise comparisons revealed that participants in the control group (M_{CON} = 1.26, SD_{CON} = 1.11) exhibited lower PEs for control profiles than participants in the BPD group (M_{BPD} = 1.42, SD_{BPD} = 1.21; EMMs contrast: b = 0.16, SE = 0.05, z = 3.00, p < 0.01). We additionally tested whether groups differed in estimating

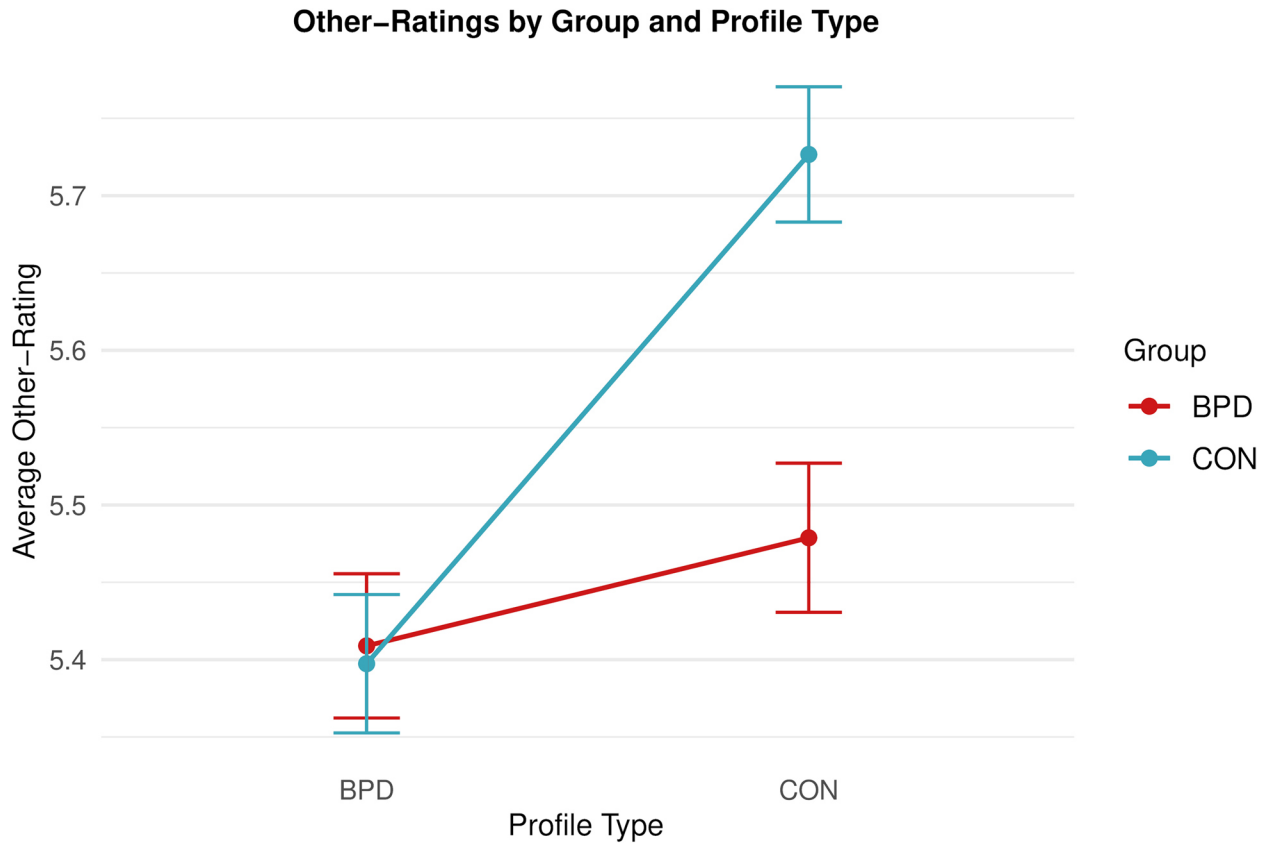


Fig. 3 Other-ratings of BPD profiles and control profiles in the BPD group (red) and the control group (blue). The graph indicates that participants in the control group rate the control profiles most favorably (significant *Group* and *Profile Type* interaction in an LME model, and significant post-hoc contrasts for the effect of *Group* on the control profiles, but not for the BPD profiles). Lines show observed means, error bars indicate 95% confidence intervals based on a normal distribution. Ratings on negative items are reverse coded

personality traits of BPD profiles and found no differences in PEs (M_{BPD} = 1.51, SD_{BPD} = 1.24; M_{CON} = 1.52, SD_{CON} = 1.25; EMMs contrast: b = −0.003, SE = 0.05, z = −0.06, p = 0.956). These exploratory findings suggest that group differences in PEs are specific to the control profiles, where the BPD group exhibits greater PEs than the control group (see Fig. 4).

Our hypothesis [H3.2] that individuals with BPD would be more accurate in estimating personality traits for BPD profiles than for control profiles was not supported by the data (see Supplementary Tables S19-20 and S22 for full model summaries and post-hoc contrasts). Pairwise comparisons revealed that individuals with BPD were more accurate in estimating personality traits for the control profiles (M = 1.42, SD = 1.21) than for BPD profiles (M = 1.51, SD = 1.24; b = 0.09, SE = 0.03, z = 3.32, p = 0.001). We found the same effect in the control group (CON profiles: M = 1.26, SD = 1.11; BPD profiles: M = 1.52, SD = 1.25; b = 0.26; SE = 0.03, z = 9.47, p < 0.001). Our findings suggest that both groups show lower prediction errors when learning about control profiles than when learning about BPD profiles, with the effect being more pronounced in the control group (see Fig. 4).

We tested whether PEs decreased over time, as an indication for learning, and found a small but significant negative effect of *Trial* (b = – 0.003, 95% *CI* [–0.005, –0.001], SE = 0.001, $t_{(13880)}$ = – 3.10, p < 0.01) across participants and profiles (see Supplementary Tables S23-24 for full model summaries).

In addition, we explored whether groups (BPD/CON) differed in PE reductions over time depending on the learning profile type (BPD/CON). We found that the PEs in both groups decreased significantly for the control profiles but not for the BPD profiles, as indicated by the 95% asymptotic confidence intervals that did not include zero (see Fig. 5). There were no group differences in PE reductions across or within learning profiles. The LME model summary and pairwise post-hoc tests for the slopes can be found in Supplementary Tables S25-S29.

Computational models

We tested learning in a more nuanced way by adapting and improving our previously established computational modeling framework [85]. We compared our previously introduced eight computational models in terms of their out-of-sample predictive accuracy as measured by their

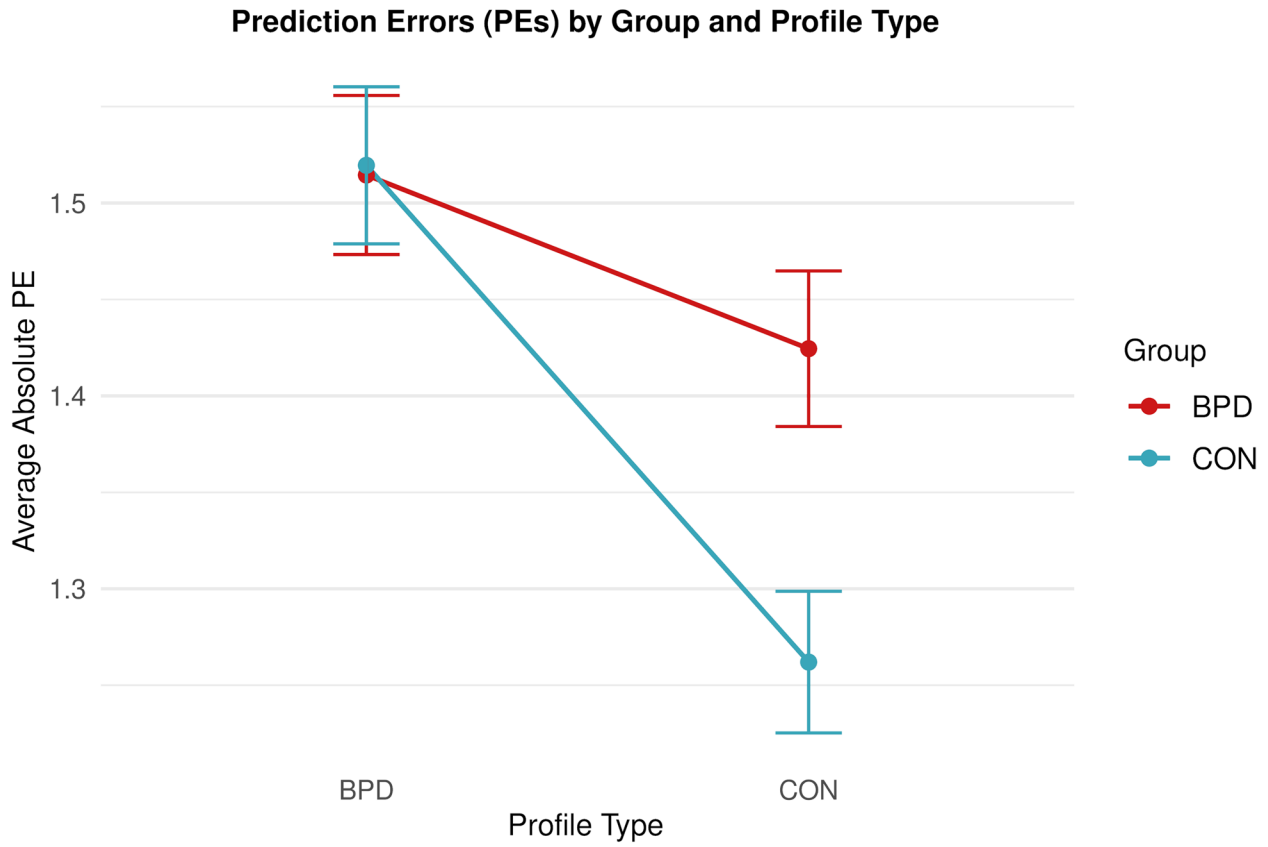


Fig. 4 Average absolute PEs (across all trials) on BPD profiles and control profiles in the BPD group (red) and the control group (blue). The graph indicates that both groups show lower prediction errors in the control profiles, with the effect being more pronounced in the control group (significant Group and Profile Type interaction in an LME model and significant post-hoc contrasts for the effect of Group on the control profiles, but not for the BPD profiles). Lines show observed means, error bars indicate 95% confidence intervals based on a normal distribution

Mean Squared Error. The results of the model comparison are shown in Fig. 6. In short, our preregistered hypothesis [H4] of different winning models, depending on the two groups and the two types of learning profiles, could not be supported. The fine granularity model with self reference points and the one without reference points outperformed all other models (all ΔMSE >2SE) but did not credibly differentiate themselves from each other (all ΔMSE < 2SE). We therefore specifically used the "Fine (No RP)" model to address [H5], i.e., to test for differences in learning rates between the two groups. No credible differences in the learning rates emerged. The average V0 exclusively differed between the control-on-control vs. the BPD-on-control condition (Δ = 0.42, 95% SPI = [0.19, 0.65]. Supplementary Table S30 presents all parameter comparisons.

Given limitations in the design, we refrain from reporting the preregistered hypothesis [H6] about differences between learning rates for positive and negative PEs. We did not adequately consider these design limitations when writing the preregistration.

Discussion

In this study, we examined whether women with BPD show negativity biases in trait assessments for self- and others and how these biases would affect their learning about other people's personalities. As hypothesized, we found that the BPD group rated themselves more negatively than the control group. When rating BPD profiles, the groups did not differ in other-ratings, but when rating control profiles, control participants rated others more positively on personality traits compared to the BPD group. Control participants were also more accurate in estimating the personality traits for control profiles than participants in the BPD group. We did not find this "in-group effect" for BPD profiles. Both groups were similarly inaccurate when estimating personality traits for the BPD profiles. Relatedly, the PEs in both groups showed a small but significant decrease across time for the control profiles but not for the BPD profiles (although the difference in PE decrease did not reach significance). These findings suggest that BPD profiles are harder to learn about than control profiles. Comparing different computational models revealed that the fine granularity models with either self-reference points or no reference points

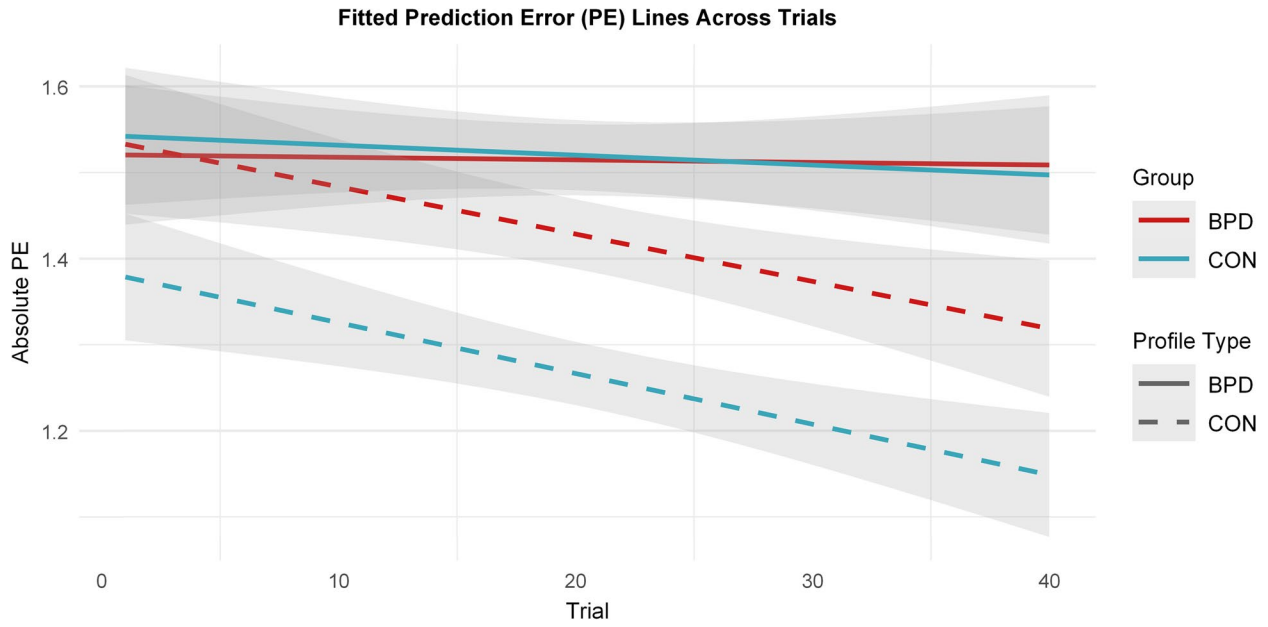


Fig. 5 Fitted linear trends of PE reductions across trials for both groups (BPD in red vs. CON in blue) and profile types (BPD in solid lines vs. CON in dashed lines). The graph indicates that both groups exhibit reductions in PEs across trials when learning about control profiles, but not when learning about BPD profiles. In addition, the BPD group starts with higher PEs than the control group when learning about control profiles. Lines represent linear model fits of PEs across trials for each condition, with 95% confidence intervals (CIs) shaded around the lines. Note that individual data points are not shown; the lines represent model-predicted trends. Post-hoc tests of the LME model indicated that the slopes for control profiles were significantly different from zero for both groups, whereas the slopes for the BPD profiles were not. Post-hoc contrasts further revealed group differences for the control profiles, but not for the BPD profiles

exhibited the best fit in both groups regardless of profile. While these two models outperformed all other models, they could not be distinguished from each other.

BPD group shows negatively biased self-ratings

Consistent with our hypothesis, individuals in the BPD group rated themselves more negatively across 40 personality traits compared to controls—reporting lower endorsement of positive traits and higher endorsement of negative traits.

These findings are in line with previous studies showing that individuals with BPD describe themselves with more negative and less positive trait words [14, 34, 118, 119], and are consistent with a broader literature indicating a negative self-concept and heightened self-related negative emotions in BPD [35]. For instance, Winter et al. [120] found that individuals with BPD judged positive and neutral words less positively than healthy controls if the words referred to themselves, highlighting the role of self-referential processing. Moreover, numerous studies have demonstrated that BPD is associated with lower self-esteem [121–125], heightened levels of self-criticism [13], trait-shame proneness [126], and self-disgust [127–130].

These findings may reflect differences in self-motives: while healthy individuals are typically motivated to maintain a stable and positive self-concept, showing a greater propensity to internalize favorable feedback [29–32], individuals with BPD show a reduced or absent positivity bias and heightened sensitivity to unfavorable feedback [14, 33–35, 131]. Such heightened sensitivity to negative self-relevant information may reinforce negative self-perceptions, potentially contributing to emotional dysregulation [132] and interpersonal dysfunction [133], while hindering the formation of positive relationships [17].

To validate the observed differences in self-ratings on personality traits, we further assessed whether BPD participants rated themselves more negatively on two established personality measures: the NEO-FFI [36, 89] and the PID-5-BF [90–92]. On the NEO-FFI, individuals with BPD scored significantly higher than controls on Neuroticism—a dimension associated with emotional instability, anxiety, and self-consciousness. Consistent with prior findings, the BPD group also scored significantly lower on Extraversion, Agreeableness, and Conscientiousness—personality factors that are typically associated with socially desirable traits such as sociability, cooperativeness, and reliability. These results align with previous research showing that individuals with BPD tend to exhibit higher Neuroticism and lower Agreeableness and Conscientiousness compared to healthy controls [38, 39, 134, 135]. Lower Extraversion has also been reported, albeit with somewhat less consistency across studies

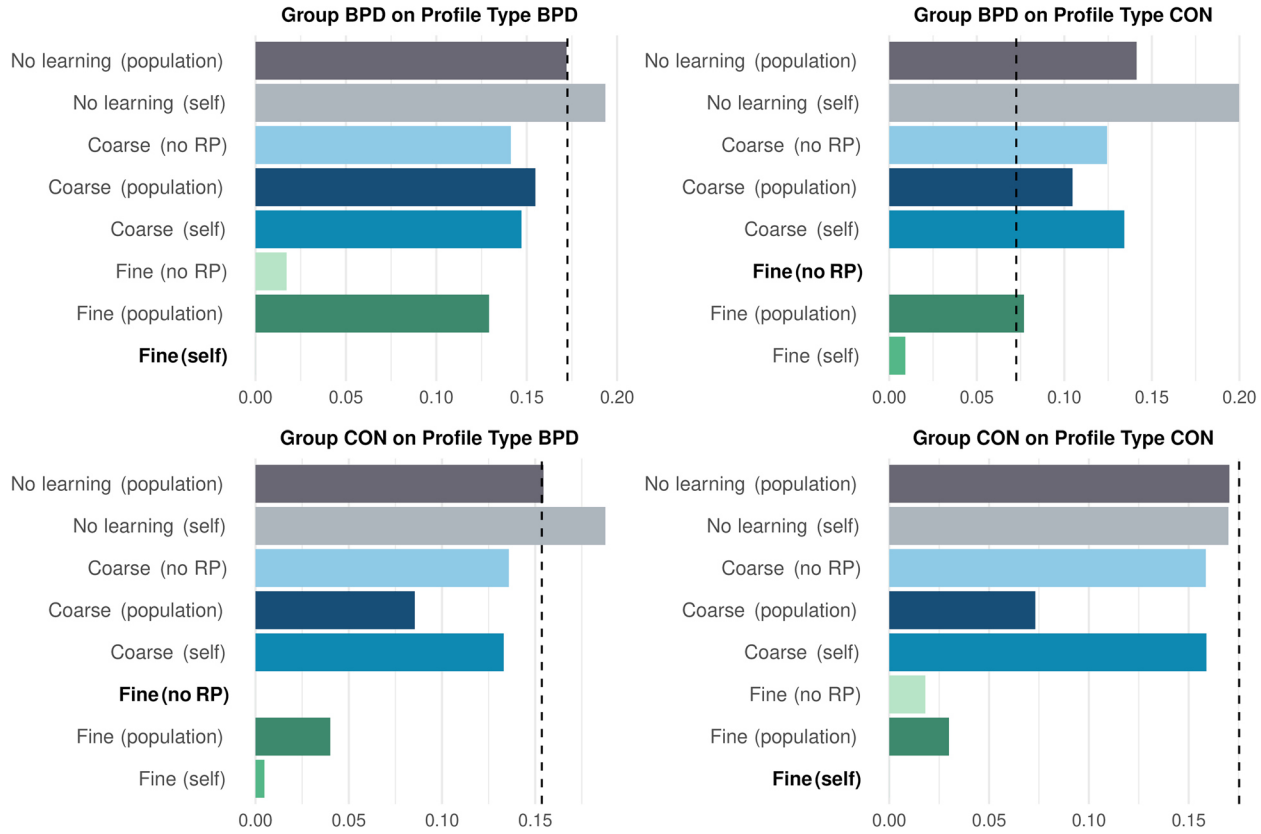


Fig. 6 Barplots showing the Mean Squared Error (MSE) out-of-sample predictive performance differences of models compared to the best performing model across conditions (covering every pairing of profile and participant type). Performance differences were estimated by fitting eight models with different assumptions about the learning process to experimental data. The performance of a naive mean-prediction is given as a baseline (dashed line). No learning models are depicted in shades of gray, coarse-grained learning models in shades of blue, and fine-grained learning models in shades of green. Lower MSE values indicate better fit. The plots show that the fine granularity models outperform all other models in all conditions (see main text for complementary statistical analyses)

[136, 137]. Openness, meanwhile, appears largely unrelated to BPD [38, 39].

When examining maladaptive personality traits using the PID-5-BF, individuals with BPD scored significantly higher than controls across all five pathological domains: Negative Affect, Detachment, Antagonism, Disinhibition, and Psychoticism. These results align with prior findings demonstrating that BPD is associated with elevated levels of maladaptive traits relative to both healthy and clinical comparison groups [41, 138]. Notably, PID-5 profiles in BPD are characterized by prominent features of negative affectivity and impulsivity, which are core aspects of the disorder's clinical presentation and differ in other clinical groups [41, 138].

Taken together, our findings provide robust support for the hypothesis that individuals with BPD exhibit more negative self-views across a range of evaluative contexts. These pervasive negative self-evaluations may play a key role not only in the emotional but also in the interpersonal difficulties characteristic of BPD.

Negative biases of individuals with BPD generalize to others

In contrast to our hypothesis, the groups did not differ significantly in their overall evaluations of others in the learning task. However, when dissociating profile types, we found that control participants evaluated control profiles more favorably than participants with BPD. In contrast, the BPD group showed comparable ratings for both BPD and control profiles. Importantly, while participants were unaware of whether the profiles were from patients with BPD or controls during the learning task, they did receive trial-by-trial feedback on the profiles in question. Our results suggest that the more negative evaluations of BPD profiles by BPD and control participants were influenced by the trait characteristics inherent to the profiles rather than by explicit knowledge or stereotypes about the BPD group. The group differences in how participants rated control profiles specifically showed that BPD participants maintained negative views of control profiles, despite receiving feedback that the other people in question rated themselves more favorably.

Our results are consistent with previous research indicating that individuals with BPD tend to view others as more negative and less trustworthy [14, 43, 51–55]. Notably, one study reported no differences in how individuals with BPD evaluated the valence of other-referential stimuli, suggesting that a negative bias for others may depend on whether the information is self-relevant or socially meaningful [120]. We found that both control and BPD participants rated BPD profiles similarly, indicating that the negative views align with the trait information they received on BPD profiles (which are inherently more negative). Previous research also showed that even in the absence of explicit trait feedback, participants evaluate individuals with BPD more negatively based on photographs and videos [27, 42–44]. We show that these negative evaluations are reinforced similarly through trait feedback, even in the absence of visual cues or diagnostic labels. This suggests that the trait pattern of people with BPD may lead to more negative impressions by others—even when others are unaware of the individual's mental health status.

BPD profiles are harder to learn than control profiles across groups

Both groups were more accurate in assessing control profiles than BPD profiles, and this effect was more pronounced in the control group. In line with our hypothesis, control participants were more accurate (i.e., showed slightly lower PEs) when evaluating control profiles than participants in the BPD group. However, contrary to our expectations, participants in the BPD group were not more accurate when evaluating BPD profiles compared to control profiles.

Participants in both groups became slightly more accurate in their personality trait assessment across trials, as indicated by a small, general decrease in PEs. Further analysis revealed that this effect was dependent on the profile type rather than the participant group. Specifically, both groups improved in accuracy when evaluating control profiles, but not when evaluating BPD profiles. Interestingly, the improvement (i.e., PE slopes) was similar across groups, which may suggest that learning about others based on feedback is similar in the BPD and control group. The BPD group started with lower initial accuracy (i.e., higher PEs), which may account for their overall reduced task performance. Surprisingly, BPD participants did not improve in accuracy when evaluating BPD profiles over time. This suggests that learning to assess personality traits in individuals with BPD is challenging not only for controls but also for those with BPD themselves. The absence of a group difference in improvement is noteworthy. Despite beginning with lower initial accuracy, individuals with BPD adapted their judgments over time to a similar extent as controls. Highlighting such intact capacities is important, as it underscores that social-cognitive processes in BPD are not uniformly impaired. Preserved feedback-based updating may provide a foundation for adaptive learning under conditions that are less complex and more realistic than the present task.

Although the overall effects were small, the observed decrease in PEs is consistent with a feedback-based learning process. Interpretation should nevertheless be cautious, given the modest size of the effects. As we discuss further in the limitations section, the small magnitude of these effects may partly reflect the complexity and limited realism of the current task design. Previous studies have shown that similar paradigms can capture social learning effects more robustly [80–83, 85], but given the demands of our paradigm, the current findings may be more cautiously interpreted as an improvement in accuracy that hints at an underlying learning process, rather than a strong confirmation of learning per se. Future refinements of the task design should help capture this process more clearly and with greater robustness.

Both groups rely on a fine granularity learning model to improve predictions of others over time

Learning in both groups was best captured by models that integrate reinforcement learning with fine-grained social knowledge structures. Notably, models incorporating fine granularity consistently outperformed no learning models (which assume no belief updating over time) or coarse granularity models (which assume a generalized learning across broad personality dimensions), regardless of group or profile type.

Contrary to our initial hypothesis—that participants in the control group would rely on fine-grained similarity and participants in the BPD group on coarse-grained similarity—the results suggest that both groups used detailed knowledge of personality trait similarities when learning about others.

Using extended models, we tried to estimate participants' initial expectations for each trait (V0 estimates), capturing prior beliefs about average trait expressions. Notably, average V0 values differed only credibly between groups when evaluating control profiles in the learning task.

While our modelling framework offers a nuanced way to assess social learning, the model comparisons did not suggest group differences in the learning process itself. Rather, individuals with BPD may be influenced more by their negative baseline expectations about themselves and others than by a dysfunctional learning process. Notably, this framework can be extended to other types of learning processes [16].

Limitations and outlook

As with any experimental design, this study has several limitations that should be acknowledged.

First, the task employed in this study may have been overly difficult for participants. In future research, we will reduce complexity by presenting fewer profiles, allowing more time for responses, and selecting profiles that are more easily distinguishable. With these adjustments, we hope to ensure that the task better captures participants' evaluation and learning processes.

Second, the current design lacked realism and direct social interactions, which limits its ecological validity. To address this, future studies should enhance the realism of the stimuli—for example, by including photographs and richer background information for each profile. In addition, incorporating more naturalistic approaches, such as live social interactions, could better approximate the dynamics of real-world personality learning. While our study emphasizes the importance of using a controlled baseline paradigm without confounding social cues, we also highlight the relevance of extending this work by incorporating more realistic stimuli to better capture the process of learning about other people's personalities.

Third, as noted in the discussion, these design limitations may help explain why the learning effects observed in this study were small and why the exploratory power was limited. The overall weak learning effect might have limited our ability to detect significant group differences in the learning mechanisms. At the same time, even subtle effects can be informative in computational modeling, as they enable the estimation of underlying parameters and provide a basis for testing mechanistic hypotheses. An additional strength of this work is the introduction of a novel paradigm that integrates reinforcement learning with pre-existing knowledge in a clinical population. Our framework can be readily adapted to other populations and learning contexts. With further refinements to the task design, this paradigm may serve as a valuable starting point for systematically testing social learning processes in clinical populations.

Fourth, the sample consisted exclusively of female participants learning about female profiles, which limits the generalizability of the findings. Additionally, our sample size may limit statistical power and the detection of smaller effects. Nonetheless, a key strength of this study is the use of a well-characterized sample of women with BPD, carefully matched with controls on age and intelligence. This provides a foundation for future work using dimensional diagnostic approaches.

Conclusion

Our study provides compelling evidence that individuals with BPD exhibit pervasive negative biases not only in their self-evaluations, but also in their perceptions of others' personalities. Despite these biases, both groups relied on fine-grained social learning models, suggesting that the biased evaluations in individuals with BPD are likely driven by more negative baseline expectations rather than impaired learning mechanisms, although this conclusion warrants further investigation. The learning effects observed were modest and should be interpreted with caution, but the absence of a deficit in feedback-based updating is noteworthy. Individuals with BPD were able to adapt their judgments over time to a similar extent as controls, underscoring that their capacity for feedback-based social learning remains intact. Highlighting this preserved ability provides an important counterbalance to deficit-focused narratives. At the same time, profiles reflecting BPD traits remained consistently harder to assess accurately across groups, which may contribute to a self-reinforcing cycle: individuals with BPD hold more negative expectations about themselves and others, and are also more likely to be viewed negatively by others. This reciprocal negativity may exacerbate interpersonal difficulties. Importantly, the intact learning capacity observed suggests a potential therapeutic resource: interventions that target and recalibrate initial social expectations could help mitigate these biases and enhance social outcomes for individuals with BPD.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40479-025-00320-4.

Supplementary Material 1.

Acknowledgements

We thank Clara Beitz, Clara Heilmann, Inga von Kortzfleisch, Sita Sielaff, and Johanna Löloff for help with data acquisition and Marie-Luise Otte for advising on the diagnostic interviews. For the publication fee we acknowledge financial support by Heidelberg University.

Authors' contributions

LMD collected the data, performed data analyses, and wrote the original draft. RP performed modeling analyses. KMMF designed the study and performed data analyses. GR wrote the original draft and supervised the project. SCH provided essential support with the clinical sample and revised the manuscript. CWK designed the study, wrote the original draft, supervised the project, and realized the funding. All authors reviewed and edited the manuscript and gave final approval for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL. This project was supported by the Emmy Noether Research Group grant (392443797) from the German Research Foundation (DFG). This manuscript also draws on an ongoing US-German collaboration between Drs. Rosenblau and Korn, supported by the NIMH (R01MH132236) and Bundesministerium für Bildung und Forschung (BMBF), respectively. Additionally, this project was endorsed by the German Center for Mental Health (DZPG).

Data availability

The data and code for this study are publicly available on Github (https://github.com/dnhi-lab/PersonalityLearningBPD_2025).

Declarations

Ethics approval and consent to participate

The project was approved by the local ethics committee (Faculty of Medicine, Heidelberg University, S-137/2020). All participants were briefed on the study procedure and gave written informed consent before participating.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of General Psychiatry, Center for Psychosocial Medicine, University of Heidelberg, Heidelberg, Germany
[2]German Center for Mental Health (DZPG), partner site Mannheim/Heidelberg/Ulm, Heidelberg, Germany
[3]Section Social Neuroscience, Department of General Adult Psychiatry, Heidelberg University Hospital, Voßstraße 4, Heidelberg 69115, Germany
[4]Department of Psychological and Brain Sciences, The George Washington University, Washington, D.C, USA
[5]Autism and Neurodevelopmental Disorders Institute, The George Washington University, Washington, D.C, USA

Received: 7 July 2025 / Accepted: 5 October 2025
Published online: 09 February 2026

References

1. Kandler C, Bleidorn W. Personality Differences and Development: Genetic and Environmental Contributions. Int Encycl Soc Behav Sci. Elsevier; 2015. p. 884–90. https://doi.org/10.1016/B978-0-08-097086-8.25011-3.
2. Borkenau P, Liebler A. Trait inferences: Sources of validity at zero acquaintance. J Pers Soc Psychol. 1992;62:645–57. https://doi.org/10.1037/0022-3514.62.4.645.
3. Bott A, Brockmann L, Denneberg I, Henken E, Kuper N, Kruse F, et al. Spontaneous Trait Inferences From Behavior: A Systematic Meta-Analysis. Pers Soc Psychol Bull. 2024;50:78–102. https://doi.org/10.1177/01461672221100336.
4. Carlston DE, Skowronski JJ. Savings in the relearning of trait information as evidence for spontaneous inference generation. J Pers Soc Psychol 1994;66:840–56. https://doi.org/10.1037/0022-3514.66.5.840.
5. Funder DC. Accurate Personality Judgment. Curr Dir Psychol Sci. 2012;21:177–82. https://doi.org/10.1177/0963721412445309.
6. Jaksic C, Schlegel K. Accuracy in Judging Others' Personalities: The Role of Emotion Recognition, Emotion Understanding, and Trait Emotional Intelligence. J Intell. 2020;8:34. https://doi.org/10.3390/jintelligence8030034.
7. Little AC, Perrett DI. Using composite images to assess accuracy in personality attribution to faces. Br J Psychol. 2007;98:111–26. https://doi.org/10.1348/000712606X109648.
8. Naumann LP, Vazire S, Rentfrow PJ, Gosling SD. Personality Judgments Based on Physical Appearance. Pers Soc Psychol Bull. 2009;35:1661–71. https://doi.org/10.1177/0146167209346309.
9. Albright L, Kenny DA, Malloy TE. Consensus in personality judgments at zero acquaintance. J Pers Soc Psychol. 1988;55:387–95. https://doi.org/10.1037/0022-3514.55.3.387.
10. Cho JC, Knowles ED. I'm like you and you're like me: Social projection and self-stereotyping both help explain self–other correspondence. J Pers Soc Psychol. 2013;104:444–56. https://doi.org/10.1037/a0031017.
11. Thielmann I, Spadaro G, Balliet D. Personality and prosocial behavior: A theoretical framework and meta-analysis. Psychol Bull. 2020;146:30–90. https://doi.org/10.1037/bul0000217.
12. Wood D, Harms P, Vazire S. Perceiver effects as projective tests: What your perceptions of others say about you. J Pers Soc Psychol. 2010;99:174–90. https://doi.org/10.1037/a0019390.
13. Kopala-Sibley DC, Zuroff DC, Russell JJ, Moskowitz DS, Paris J. Understanding heterogeneity in borderline personality disorder: Differences in affective reactivity explained by the traits of dependency and self-criticism. J Abnorm Psychol. 2012;121:680–91. https://doi.org/10.1037/a0028513.
14. Korn CW, La Rosée L, Heekeren HR, Roepke S. Social feedback processing in borderline personality disorder. Psychol Med. 2016;46:575–87. https://doi.org/10.1017/S003329171500207X.
15. Niedtfeld I, Renkewitz F, Mädebach A, Hillmann K, Kleindienst N, Schmahl C, et al. Enhanced memory for negative social information in borderline personality disorder. J Abnorm Psychol. 2020;129:480–91. https://doi.org/10.1037/abn0000540.
16. Rosenblau G, Frolichs K, Korn CW. A neuro-computational social learning framework to facilitate transdiagnostic classification and treatment across psychiatric disorders. Neurosci Biobehav Rev. 2023;149:105181. https://doi.org/10.1016/j.neubiorev.2023.105181.
17. Van Schie CC, Whiting L, Grenyer BFS. How negative self-views may interfere with building positive relationships: An experimental analogue of identity dysfunction in borderline personality disorder. PLOS ONE. 2024;19:e0301196. https://doi.org/10.1371/journal.pone.0301196.
18. Werner AM, Tibubos AN, Rohrmann S, Reiss N. The clinical trait self-criticism and its relation to psychopathology: A systematic review – Update. J Affect Disord. 2019;246:530–47. https://doi.org/10.1016/j.jad.2018.12.069.
19. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Fifth Edition. American Psychiatric Association; 2013. https://doi.org/10.1176/appi.books.9780890425596.
20. Gunderson JG, Herpertz SC, Skodol AE, Torgersen S, Zanarini MC. Borderline personality disorder. Nat Rev Dis Primer. 2018;4:18029. https://doi.org/10.1038/nrdp.2018.29.
21. Bouchard S, Sabourin S. Borderline personality disorder and couple dysfunctions. Curr Psychiatry Rep. 2009;11:55–62. https://doi.org/10.1007/s11920-009-0009-x.
22. Lazarus SA, Cheavens JS, Festa F, Zachary Rosenthal M. Interpersonal functioning in borderline personality disorder: A systematic review of behavioral and laboratory-based assessments. Clin Psychol Rev. 2014;34:193–205.
23. Liebke L, Bungert M, Thome J, Hauschild S, Gescher DM, Schmahl C, et al. Loneliness, social networks, and social functioning in borderline personality disorder. Personal Disord Theory Res Treat. 2017;8:349–56. https://doi.org/10.1037/per0000208.
24. Stepp SD, Pilkonis PA, Yaggi KE, Morse JQ, Feske U. Interpersonal and Emotional Experiences of Social Interactions in Borderline Personality Disorder. J Nerv Ment Dis. 2009;197:484–91. https://doi.org/10.1097/NMD.0b013e3181aad2e7.
25. Amestoy ME, Best MW, Ruocco AC, Uliaszek AA. Borderline personality disorder stigma: Examining the effects of diagnostic disclosure, behavior, and gender as sources of stigma in the general population. Personal Disord Theory Res Treat. 2024;15:254–63. https://doi.org/10.1037/per0000672.
26. Aviram RB, Brodsky BS, Stanley B. Borderline Personality Disorder, Stigma, and Treatment Implications. Harv Rev Psychiatry. 2006;14:249–56. https://doi.org/10.1080/10673220600975121.
27. Hepp J, Störkel LM, Kieslich PJ, Schmahl C, Niedtfeld I. Negative evaluation of individuals with borderline personality disorder at zero acquaintance. Behav Res Ther. 2018;111:84–91. https://doi.org/10.1016/j.brat.2018.09.009.
28. Herpertz SC, Schneider I, Renneberg B, Schneider A. Patients with personality disorders in everyday clinical practice—implications of the ICD-11. Dtsch Ärztebl Int. 2022. https://doi.org/10.3238/arztebl.m2022.0001.
29. Elder JJ, Davis TH, Hughes BL. A Fluid Self-Concept: How the Brain Maintains Coherence and Positivity across an Interconnected Self-Concept While Incorporating Social Feedback. J Neurosci. 2023;43:4110–28. https://doi.org/10.1523/JNEUROSCI.1951-22.2023.
30. Hughes BL, Beer JS. Protecting the Self: The Effect of Social-evaluative Threat on Neural Representations of Self. J Cogn Neurosci. 2013;25:613–22. https://doi.org/10.1162/jocn_a_00343.
31. Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR. Positively Biased Processing of Self-Relevant Social Feedback. J Neurosci. 2012;32:16832–44. https://doi.org/10.1523/JNEUROSCI.3016-12.2012.
32. Leary MR. Motivational and Emotional Aspects of the Self. Annu Rev Psychol. 2007;58:317–44. https://doi.org/10.1146/annurev.psych.58.110405.085658.
33. Van Schie CC, Chiu C-D, Rombouts SARB, Heiser WJ, Elzinga BM. Stuck in a negative me: fMRI study on the role of disturbed self-views in social feedback processing in borderline personality disorder. Psychol Med. 2020;50:625–35. https://doi.org/10.1017/S0033291719000448.
34. Vater A, Schröder-Abé M, Weißgerber S, Roepke S, Schütz A. Self-concept structure and borderline personality disorder: Evidence for negative compartmentalization. J Behav Ther Exp Psychiatry. 2015;46:50–8. https://doi.org/10.1016/j.jbtep.2014.08.003.

35. Winter D, Bohus M, Lis S. Understanding Negative Self-Evaluations in Borderline Personality Disorder—a Review of Self-Related Cognitions, Emotions, and Motives. Curr Psychiatry Rep. 2017;19:17. https://doi.org/10.1007/s11920-017-0771-0.
36. Costa PTJr, McCrae RR. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory Professional Manual. Psychological Assessment Resources; 1992.
37. Goldberg LR. An alternative "description of personality": The Big-Five factor structure. J Pers Soc Psychol. 1990;59:1216–29. https://doi.org/10.1037/0022-3514.59.6.1216.
38. Samuel D, Widiger T. A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis☆. Clin Psychol Rev. 2008;28:1326–42. https://doi.org/10.1016/j.cpr.2008.07.002.
39. Saulsman LM, Page AC. The five-factor model and personality disorder empirical literature: A meta-analytic review. Clin Psychol Rev. 2004;23:1055–85. https://doi.org/10.1016/j.cpr.2002.09.001.
40. D'Adda F, Sighinolfi G, Mitolo M, Scala M, Guidi L, Motta L, et al. Neurobiological correlates of personality dimensions in borderline personality disorder using graph analysis of functional connectivity. Sci Rep. 2025;15:12623. https://doi.org/10.1038/s41598-025-85989-x.
41. Calvo N, Valero S, Sáez-Francàs N, Gutiérrez F, Casas M, Ferrer M. Borderline Personality Disorder and Personality Inventory for DSM-5 (PID-5): Dimensional personality assessment with DSM-5. Compr Psychiatry. 2016;70:105–11. https://doi.org/10.1016/j.comppsych.2016.07.002.
42. Daros AR, Ruocco AC, Rule NO. Identifying Mental Disorder from the Faces of Women with Borderline Personality Disorder. J Nonverbal Behav. 2016;40:255–81. https://doi.org/10.1007/s10919-016-0237-9.
43. Hepp J, Kieslich PJ, Schmitz M, Schmahl C, Niedtfeld I. Negativity on two sides: Individuals with borderline personality disorder form negative first impressions of others and are perceived negatively by them. Personal Disord Theory Res Treat. 2021;12:514–25. https://doi.org/10.1037/per0000412.
44. Oltmanns TF, Friedman JNW, Fiedler ER, Turkheimer E. Perceptions of people with personality disorders based on thin slices of behavior. J Res Personal. 2004;38:216–29. https://doi.org/10.1016/S0092-6566(03)00066-7.
45. Bertsch K, Krauch M, Stopfer K, Haeussler K, Herpertz SC, Gamer M. Interpersonal Threat Sensitivity in Borderline Personality Disorder: An Eye-Tracking Study. J Personal Disord. 2017;31:647–70. https://doi.org/10.1521/pedi_2017_31_273.
46. Bertsch K, Gamer M, Schmidt B, Schmidinger I, Walther S, Kästel T, et al. Oxytocin and Reduction of Social Threat Hypersensitivity in Women With Borderline Personality Disorder. Am J Psychiatry. 2013;170:1169–77. https://doi.org/10.1176/appi.ajp.2013.13020263.
47. Domes G, Czieschnek D, Weidler F, Berger C, Fast K, Herpertz SC. Recognition of Facial Affect in Borderline Personality Disorder. J Personal Disord. 2008;22:135–47. https://doi.org/10.1521/pedi.2008.22.2.135.
48. Fenske S, Lis S, Liebke L, Niedtfeld I, Kirsch P, Mier D. Emotion recognition in borderline personality disorder: effects of emotional information on negative bias. Borderline Personal Disord Emot Dysregulation. 2015;2:10. https://doi.org/10.1186/s40479-015-0031-z.
49. Izurieta Hidalgo NA, Oelkers-Ax R, Nagy K, Mancke F, Bohus M, Herpertz SC, et al. Time course of facial emotion processing in women with borderline personality disorder: an ERP study. J Psychiatry Neurosci. 2016;41:16–26. https://doi.org/10.1503/jpn.140215.
50. Thome J, Liebke L, Bungert M, Schmahl C, Domes G, Bohus M, et al. Confidence in facial emotion recognition in borderline personality disorder. Personal Disord Theory Res Treat. 2016;7:159–68. https://doi.org/10.1037/per0000142.
51. 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. https://doi.org/10.1016/j.brat.2009.02.003.
52. Hepp J, Kieslich PJ, Wycoff AM, Bertsch K, Schmahl C, Niedtfeld I. Mouse-tracking reveals cognitive conflict during negative impression formation in women with BorderlinePersonality Disorder or Social Anxiety Disorder. PLOS ONE. 2021;16:e0247955. https://doi.org/10.1371/journal.pone.0247955.
53. Scheunemann J, Jelinek L, Biedermann SV, Lipp M, Yassari AH, Kühn S, et al. Can you trust this source? Advice taking in borderline personality disorder. Eur Arch Psychiatry Clin Neurosci. 2023;273:875–85. https://doi.org/10.1007/s00406-022-01539-w.
54. Lévay EE, Bajzát B, Unoka ZS. Expectation of Selfishness From Others in Borderline Personality Disorder. Front Psychol. 2021;12:702227. https://doi.org/10.3389/fpsyg.2021.702227.
55. Masland SR, Schnell SE, Shah TV. Trust Beliefs, Biases, and Behaviors in Borderline Personality Disorder: Empirical Findings and Relevance to Epistemic Trust. Curr Behav Neurosci Rep. 2020;7:239–49. https://doi.org/10.1007/s40473-020-00220-7.
56. Guest O, Martin AE. How Computational Modeling Can Force Theory Building in Psychological Science. Perspect Psychol Sci. 2021;16:789–802. https://doi.org/10.1177/1745691620970585.
57. Van Rooij I, Baggio G. Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. Perspect Psychol Sci. 2021;16:682–97. https://doi.org/10.1177/1745691620970604.
58. Lewandowsky S, Farrell S. Computational Modeling in Cognition: Principles and Practice. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.; 2011. https://doi.org/10.4135/9781483349428.
59. Niv Y, Langdon A. Reinforcement learning with Marr. Curr Opin Behav Sci. 2016;11:67–73. https://doi.org/10.1016/j.cobeha.2016.04.005.
60. Sutton RS, Barto A. Reinforcement learning: an introduction. Nachdruck. Cambridge, Massachusetts: The MIT Press; 2014.
61. Montague P, Dayan P, Sejnowski T. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 1996;16:1936–47. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996.
62. Schultz W, Apicella P, Ljungberg T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci. 1993;13:900–13. https://doi.org/10.1523/JNEUROSCI.13-03-00900.1993.
63. Pagnoni G, Zink CF, Montague PR, Berns GS. Activity in human ventral striatum locked to errors of reward prediction. Nat Neurosci, 2002;5:97–8. https://doi.org/10.1038/nn802.
64. Schönberg T, Daw ND, Joel D, O'Doherty JP. Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making. J Neurosci. 2007;27:12860–7. https://doi.org/10.1523/JNEUROSCI.2496-07.2007.
65. Niv Y. Reinforcement learning in the brain. J Math Psychol. 2009;53:139–54. https://doi.org/10.1016/j.jmp.2008.12.005.
66. Charpentier CJ, O'Doherty JP. The application of computational models to social neuroscience: promises and pitfalls. Soc Neurosci. 2018;13:637–47. https://doi.org/10.1080/17470919.2018.1518834.
67. Chierchia G, Soukupová M, Kilford EJ, Griffin C, Leung J, Palminteri S, et al. Confirmatory reinforcement learning changes with age during adolescence. Dev Sci. 2023;26:e13330. https://doi.org/10.1111/desc.13330.
68. Dayan P, Niv Y. Reinforcement learning: The Good, The Bad and The Ugly. Curr Opin Neurobiol. 2008;18:185–96. https://doi.org/10.1016/j.conb.2008.08.003.
69. Lockwood PL, Klein-Flügge MC. Computational modelling of social cognition and behaviour—a reinforcement learning primer. Soc Cogn Affect Neurosci. 2020:nsaa040. https://doi.org/10.1093/scan/nsaa040.
70. Rodriguez Buritica JM, Eppinger B, Heekeren HR, Crone EA, Van Duijvenvoorde ACK. Observational reinforcement learning in children and young adults. Npj Sci Learn. 2024;9:18. https://doi.org/10.1038/s41539-024-00227-9.
71. Zhang L, Lengersdorff L, Mikus N, Gläscher J, Lamm C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. Soc Cogn Affect Neurosci. 2020;15:695–707. https://doi.org/10.1093/scan/nsaa089.
72. Czekalla N, Schröder A, Mayer AV, Stierand J, Stolz DS, Kube T, et al. Neurocomputational Mechanisms Underlying Maladaptive Self-Belief Formation in Depression. 2024. https://doi.org/10.1101/2024.05.09.593087.
73. Liebenow B, Jones R, DiMarco E, Trattner JD, Humphries J, Sands LP, et al. Computational reinforcement learning, reward (and punishment), and dopamine in psychiatric disorders. Front Psychiatry. 2022;13:886297. https://doi.org/10.3389/fpsyt.2022.886297.
74. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. 2011;14:154–62. https://doi.org/10.1038/nn.2723.
75. Ossola P, Garrett N, Biso L, Bishara A, Marchesi C. Anhedonia and sensitivity to punishment in schizophrenia, depression and opiate use disorder. J Affect Disord. 2023;330:319–28. https://doi.org/10.1016/j.jad.2023.02.120.
76. Pike AC, Robinson OJ. Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. JAMA Psychiatry. 2022;79:313. https://doi.org/10.1001/jamapsychiatry.2022.0051.
77. Reiter AMF, Heinze H-J, Schlagenhauf F, Deserno L. Impaired Flexible Reward-Based Decision-Making in Binge Eating Disorder: Evidence from Computational Modeling and Functional Neuroimaging. Neuropsychopharmacology. 2017;42:628–37. https://doi.org/10.1038/npp.2016.95.

78. Schuermann B, Kathmann N, Stiglmayr C, Renneberg B, Endrass T. Impaired decision making and feedback evaluation in borderline personality disorder. Psychol Med. 2011;41:1917–27. https://doi.org/10.1017/S003329171000262X.
79. Waite EE, Savalia T, Cohen AL, Haliczer LA, Huffman S, Dixon-Gordon KL. Borderline personality disorder and learning: The influences of emotional state and social versus nonsocial feedback. J Affect Disord. 2024;363:474–82. https://doi.org/10.1016/j.jad.2024.07.072.
80. Diaconescu AO, Mathys C, Weber LAE, Kasper L, Mauer J, Stephan KE. Hierarchical prediction errors in midbrain and septum during social learning. Soc Cogn Affect Neurosci. 2017;12:618–34. https://doi.org/10.1093/scan/nsw171.
81. Zaki J, Kallman S, Wimmer GE, Ochsner K, Shohamy D. Social Cognition as Reinforcement Learning: Feedback Modulates Emotion Inference. J Cogn Neurosci. 2016;28:1270–82. https://doi.org/10.1162/jocn_a_00978.
82. Fareri DS, Chang LJ, Delgado MR. Computational Substrates of Social Value in Interpersonal Collaboration. J Neurosci. 2015;35:8170–80. https://doi.org/10.1523/JNEUROSCI.4775-14.2015.
83. Rosenblau G, Korn CW, Dutton A, Lee D, Pelphrey KA. Neurocognitive Mechanisms of Social Inferences in Typical and Autistic Adolescents. Biol Psychiatry Cogn Neurosci Neuroimaging. 2021;6:782–91. https://doi.org/10.1016/j.bpsc.2020.07.002.
84. Hackel LM, Kalkstein DA, Mende-Siedlecki P. Simplifying social learning. Trends Cogn Sci. 2024;28:428–40. https://doi.org/10.1016/j.tics.2024.01.004.
85. Frolichs KMM, Rosenblau G, Korn CW. Incorporating social knowledge structures into computational models. Nat Commun. 2022;13:6205. https://doi.org/10.1038/s41467-022-33418-2.
86. Stolier RM, Hehman E, Freeman JB. Trait knowledge forms a common structure across social cognition. Nat Hum Behav. 2020;4:361–71. https://doi.org/10.1038/s41562-019-0800-6.
87. Thornton MA, Tamir DI. The Organization of Social Knowledge Is Tuned for Prediction. In: Gilead M, Ochsner KN, editors. Neural Basis Ment., Cham: Springer International Publishing; 2021. p. 283–97. https://doi.org/10.1007/978-3-030-51890-5_14.
88. Rosenblau G, Korn CW, Pelphrey KA. A Computational Account of Optimizing Social Predictions Reveals That Adolescents Are Conservative Learners in Social Contexts. J Neurosci. 2018;38:974–88. https://doi.org/10.1523/JNEUROSCI.1044-17.2017.
89. Borkenau P, Ostendorf F. NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (NEO-FFI; 2. neu normierte und vollständig überarb. Aufl.). Göttingen: Hogrefe; 2007.
90. American Psychiatric Association. The Personality Inventory for DSM–5—Brief Form (PID-5-BF)—Adult. Available from: https://www.psychiatry.org/File%20Library/Psychiatrists/Practice/DSM/APA_DSM5_The-Personality-Inventory-For-DSM-5-Brief-Form-Adult.pdf.
91. Krueger RF, Derringer J, Markon KE, Watson D, Skodol AE. Initial construction of a maladaptive personality trait model and inventory for DSM-5. Psychol Med. 2012;42:1879–90. https://doi.org/10.1017/S0033291711002674.
92. Zimmermann J, Altenstein D, Krieger T, Holtforth MG, Pretsch J, Alexopoulos J, et al. The Structure and Correlates of Self-Reported DSM-5 Maladaptive Personality Traits: Findings From Two German-Speaking Samples. J Personal Disord. 2014;28:518–40. https://doi.org/10.1521/pedi_2014_28_130.
93. Doppelhofer LM, Löloff J, Neukel C, Herpertz SC, Korn CW. Cooperative decision-making in borderline personality disorder: insights from a preregistered study using a comprehensive economic task battery. Borderline Personal Disord Emot Dysregul. 2025;12(1):24. https://doi.org/10.1186/s40479-025-00295-2.
94. Leiner DJ. SoSci Survey (Version 3.1.06) 2019.
95. Margraf J, Cwik JC. Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen. 2017. https://doi.org/10.13154/RUB.102.91.
96. Loranger AW, Sartorius N, Andreoli A. The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration International Pilot Study of Personality Disorders. Arch Gen Psychiatry. 1994;51:215. https://doi.org/10.1001/archpsyc.1994.03950030051005.
97. Loranger AW, Sartorius N, Andreoli A, Berger P, Buchheim P, Channabasavanna SM, et al. Deutschsprachige Fassung der International personality disorder examination: IPDE. Genf WHO; 1998.
98. Baudson TG, Preckel F. mini-q: Intelligenzscreening in drei Minuten. Diagnostica. 2016;62:182–97. https://doi.org/10.1026/0012-1924/a000150.
99. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. J Stat Softw. 2017;82. https://doi.org/10.18637/jss.v082.i13.
100. Lüdecke D. sjPlot: Data Visualization for Statistics in Social Science. 2024.
101. Lenth RV. emmeans: Estimated Marginal Means, aka Least-Squares Means. 2024. https://CRAN.R-project.org/package=emmeans
102. Torchiano M. Effsize - a package for efficient effect size computation. 2016. https://doi.org/10.5281/ZENODO.1480624.
103. Degenholtz HB, Bhatnagar M. Introduction to Hierarchical Modeling. J Palliat Med. 2009;12:631–8. https://doi.org/10.1089/jpm.2009.9595.
104. Gelman A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. Technometrics. 2006;48:432–5. https://doi.org/10.1198/004017005000000661.
105. Haines N, Kvam PD, Irving L, Smith CT, Beauchaine TP, Pitt MA et al. A tutorial on using generative models to advance psychological science: lessons from the reliability paradox. Psychol Methods. 2025. https://doi.org/10.1037/met0000674.
106. Suppes P. Models of Data. Stud. Log. Found. Math., vol. 44, Elsevier; 1966, p. 252–61. https://doi.org/10.1016/S0049-237X(09)70592-0.
107. White CN, Ratcliff R, Starns JJ. Diffusion models of the flanker task: Discrete versus gradual attentional selection. Cognit Psychol. 2011;63:210–38. https://doi.org/10.1016/j.cogpsych.2011.08.001.
108. Wiecki TV, Sofer I, Frank MJ. HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Front Neuroinformatics. 2013;7. https://doi.org/10.3389/fninf.2013.00014.
109. Estes WK. The problem of inference from curves based on group data. Psychol Bull. 1956;53:134–40. https://doi.org/10.1037/h0045156.
110. Farrell S, Lewandowsky S. Computational Modeling of Cognition and Behavior: 1st ed. Cambridge University Press; 2018. https://doi.org/10.1017/CBO9781316272503.
111. Turner BM, Forstmann BU, Love BC, Palmeri TJ, Van Maanen L. Approaches to analysis in model-based cognitive neuroscience. J Math Psychol. 2017;76:65–79. https://doi.org/10.1016/j.jmp.2016.01.001.
112. Whelan R. Effective Analysis of Reaction Time Data. Psychol Rec. 2008;58:475–82. https://doi.org/10.1007/bf03395630.
113. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. J Stat Softw. 2017;76. https://doi.org/10.18637/jss.v076.i01.
114. Gabry J, Češnovar R, Johnson A, Bronder S. cmdstanr: R Interface to "CmdStan." 2025.
115. Lee MD, Wagenmakers E-J. Bayesian cognitive modeling: a practical course. Cambridge, United Kingdom New York: Cambridge University Press; 2014. https://doi.org/10.1017/CBO9781139087759.
116. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. 2015. https://doi.org/10.48550/ARXIV.1507.04544.
117. Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017;355:584–5. https://doi.org/10.1126/science.aal3618.
118. Auerbach RP, Tarlow N, Bondy E, Stewart JG, Aguirre B, Kaplan C, et al. Electrocortical Reactivity During Self-referential Processing in Female Youth With Borderline Personality Disorder. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016;1:335–44. https://doi.org/10.1016/j.bpsc.2016.04.004.
119. Beeney JE, Hallquist MN, Ellison WD, Levy KN. Self–other disturbance in borderline personality disorder: Neural, self-report, and performance-based evidence. Personal Disord Theory Res Treat. 2016;7:28–39. https://doi.org/10.1037/per0000127.
120. Winter D, Herbert C, Koplin K, Schmahl C, Bohus M, Lis S. Negative Evaluation Bias for Positive Self-Referential Information in Borderline Personality Disorder. PLOS ONE. 2015;10:e0117083. https://doi.org/10.1371/journal.pone.0117083.
121. Bungert M, Liebke L, Thome J, Haeussler K, Bohus M, Lis S. Rejection sensitivity and symptom severity in patients with borderline personality disorder: effects of childhood maltreatment and self-esteem. Borderline Personal Disord Emot Dysregulation. 2015;2:4. https://doi.org/10.1186/s40479-015-0025-x.
122. Kanter JW, Parker CR, Kohlenberg RJ. Finding the self: A behavioral measure and its clinical implications. Psychother Theory Res Pract Train. 2001;38:198–211. https://doi.org/10.1037/0033-3204.38.2.198.
123. Roepke S, Schröder-Abé M, Schütz A, Jacob G, Dams A, Vater A, et al. Dialectic behavioural therapy has an impact on self-concept clarity and facets of self-esteem in women with borderline personality disorder. Clin Psychol Psychother. 2011;18:148–58. https://doi.org/10.1002/cpp.684.
124. Rüsch N, Lieb K, Göttler I, Hermann C, Schramm E, Richter H, et al. Shame and Implicit Self-Concept in Women With Borderline Personality Disorder. Am J Psychiatry. 2007;164:500–8. https://doi.org/10.1176/ajp.2007.164.3.500.

125. Abela JRZ, Payne AVL, Moussaly N. Cognitive Vulnerability to Depression in Individuals With Borderline Personality Disorder. J Personal Disord. 2003;17:319–29. https://doi.org/10.1016/j.comppsych.2017.09.003
126. Ritter K, Vater A, Rüsch N, Schröder-Abé M, Schütz A, Fydrich T, et al. Shame in patients with narcissistic personality disorder. Psychiatry Res. 2014;215:429–37. https://doi.org/10.1016/j.psychres.2013.11.019.
127. Abdul-Hamid S, Denman C, Dudas RB. Self-Relevant Disgust and Self-Harm Urges in Patients with Borderline Personality Disorder and Depression: A Pilot Study with a Newly Designed Psychological Challenge. PLoS ONE. 2014;9:e99696. https://doi.org/10.1371/journal.pone.0099696.
128. Clarke A, Simpson J, Varese F. A systematic review of the clinical utility of the concept of self-disgust. Clin Psychol Psychother. 2019;26:110–34. https://doi.org/10.1002/cpp.2335.
129. Dudas RB, Mole TB, Morris LS, Denman C, Hill E, Szalma B, et al. Amygdala and dlPFC abnormalities, with aberrant connectivity and habituation in response to emotional stimuli in females with BPD. J Affect Disord. 2017;208:460–6. https://doi.org/10.1016/j.jad.2016.10.043.
130. Schienle A, Leutgeb V, Wabnegger A. Symptom severity and disgust-related traits in borderline personality disorder: The role of amygdala subdivisions. Psychiatry Res Neuroimaging. 2015;232:203–7. https://doi.org/10.1016/j.pscychresns.2015.04.002.
131. Klein MH, Wonderlich SA, Crosby R. Self-Concept Correlates of the Personality Disorders. J Personal Disord. 2001;15:150–6. https://doi.org/10.1521/pedi.15.2.150.19214.
132. Neacsiu AD, Herr NR, Fang CM, Rodriguez MA, Rosenthal MZ. Identity Disturbance and Problems With Emotion Regulation Are Related Constructs Across Diagnoses. J Clin Psychol. 2015;71:346–61. https://doi.org/10.1002/jclp.22141.
133. Gunderson JG, Lyons-Ruth K. BPD's Interpersonal Hypersensitivity Phenotype: A Gene-Environment-Developmental Model. J Personal Disord. 2008;22:22–41. https://doi.org/10.1521/pedi.2008.22.1.22.
134. Dukalski B, Suslow T, Egloff B, Kersting A, Donges U-S. Implicit and explicit self-concept of neuroticism in borderline personality disorder. Nord J Psychiatry. 2019;73:159–68. https://doi.org/10.1080/08039488.2019.1582694.
135. Wischniewski J, Brüne M. How Do People With Borderline Personality Disorder Respond to Norm Violations? Impact of Personality Factors on Economic Decision-Making. J Personal Disord. 2013;27:531–46. https://doi.org/10.1521/pedi_2012_26_036.
136. Fowler CJ, Sharp C, Kalpakci A, Madan A, Clapp J, Allen JG, et al. A dimensional approach to assessing personality functioning: examining personality trait domains utilizing DSM-IV personality disorder criteria. Compr Psychiatry. 2015;56:75–84. https://doi.org/10.1016/j.comppsych.2014.09.001.
137. Zweig-Frank H, Paris J. The five-factor model of personality in borderline and nonborderline personality disorders. Can J Psychiatry. 1995;40(9):523–6.
138. Fowler JC, Madan A, Allen JG, Patriquin M, Sharp C, Oldham JM, et al. Clinical utility of the DSM-5 alternative model for borderline personality disorder: differential diagnostic accuracy of the BFI, SCID-II-PQ, and PID-5. Compr Psychiatry. 2018;80:97–103.

Publisher's Note

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

4 Study II: Social decision-making and expectations

Social decision-making and expectations in women with borderline personality disorder: evidence from a novel cooperation task and fMRI (manuscript)

Lisa M. Doppelhofer

Abstract

Background: Borderline Personality Disorder (BPD) is characterized by interpersonal instability and persistent difficulties in interpersonal relationships. Empirical findings on cooperation in BPD are mixed. It is unclear under which conditions cooperative behavior differs from controls and whether such differences co-occur with more pessimistic expectations about an interaction partner's behavior. Neuroimaging evidence from interactive social decision-making paradigms in BPD is sparse, but related work suggests fronto-limbic dysregulation and altered recruitment of prefrontal mentalizing networks. The present study examined how a partner's need for help and the personal risk of helping shape cooperative decisions, partner-related expectations, and their neural correlates in women with BPD compared with controls.

Methods: Women with BPD and age- and intelligence-matched controls completed two versions of a newly developed, repeated social decision-making task framed as a virtual foraging scenario with a risk of "starvation". The decision task was performed during functional magnetic resonance imaging (fMRI), whereas the expectation task was a behavioral variant outside the scanner. Participants were told they were playing with another person (a confederate), although the partner's choices were pre-generated and did not vary between participants. On each trial, participants chose between a risky Play option that could yield larger gains but also larger losses and that was more likely to succeed when both players selected it, and a Wait option that incurred a small but certain cost. Importantly, the social meaning of the Play option varied by context. In dependent-partner trials, choosing Play increased the partner's chance of survival and thus constituted a helping choice (more other-regarding). In independent-partner trials, the partner's survival was not contingent on the participant's decision, turning Play into a more mixed-motive option (self- and other-regarding). In addition, personal risk (high vs. low) was manipulated. In the expectation task, explicit expectations about the partner's likely Play choices were assessed in a role-reversed variant in which the participant depended on the partner's help. Behavioral analyses tested how Play choices and partner expectations varied by partner context and risk across groups using generalized linear mixed-effects models (GLMMs). The fMRI analyses focused on choice-phase related neural responses as a function of partner context and risk, with an emphasis on prefrontal networks implicated in social valuation and cognitive control.

Results: Across groups, participants chose the cooperative option (Play) substantially more often when the partner depended on their help than when the partner was independent, indicating preserved sensitivity to another person's need. Behavioral group differences were risk-dependent: compared with controls, individuals with BPD chose Play less often primarily under low-risk conditions across partner contexts, suggesting reduced engagement in potentially beneficial actions when personal costs were low rather than a global cooperation deficit. In the expectations task, BPD participants anticipated lower partner cooperation specifically when they depended on their partner's help. Neurally, partner context modulated choice-phase activity, with stronger medial/orbitofrontal engagement in the independent-

relative to dependent-partner conditions. Group differences were most evident in lateral prefrontal regions, including the inferior frontal gyrus (IFG), with reduced engagement in BPD under low-risk and independent-partner decisions.

Conclusion: Cooperative responding in BPD appears largely intact when partner need is explicit but is accompanied by pessimistic social expectations and reduced action engagement (Play) under low risk. IFG differences suggest altered recruitment of lateral prefrontal control mechanisms as a potential contributor to context-dependent social decision-making. Together, these findings suggest that interpersonal difficulties in BPD may stem less from a global cooperation deficit than from biased expectations about the partner's behavior and reduced context-sensitive modulation, underscoring the importance of distinguishing these components when assessing cooperation.

Keywords: borderline personality disorder, social decision-making, cooperation, expectations, risk sensitivity, medial prefrontal cortex, lateral prefrontal cortex

Introduction

Cooperative decisions require balancing personal costs against the potential benefits of helping, forming expectations about others' behavior, and managing situational uncertainty. In Borderline Personality Disorder (BPD), negatively biased social expectations and heightened sensitivity to ambiguous cues, together with dysregulated fronto-limbic-prefrontal circuits, may alter cooperative behavior. Yet, to our knowledge, no studies have examined cooperative choices and partner-related expectations jointly within interactive paradigms that allow both processes to be dissociated. Here, we introduce a novel repeated, cooperation-based task that manipulates partner dependence (i.e., whether an interaction partner depends on help) and personal risk and combines it with explicit expectation measures and functional magnetic resonance imaging (fMRI), enabling us to test how social context shapes cooperative choices, partner expectations, and prefrontal recruitment in BPD.

Cooperation is a fundamental component of human social life. It enables individuals to achieve mutual goals, maintain stable interpersonal relationships, and form social networks grounded in trust and reciprocity. Successful cooperation requires not only a willingness to incur costs for others but also the ability to form accurate expectations about others' intentions and to regulate the social emotions that arise during interaction. Disruptions in any of these processes can impair social functioning and contribute to persistent interpersonal difficulties.

BPD is characterized by impulsivity, intense emotional instability, and pervasive difficulties in social relationships (American Psychiatric Association, 2013; Gunderson et al., 2018). Individuals with BPD report fewer and more unstable and ambivalent interpersonal connections compared with the general population (Bouchard & Sabourin, 2009; Lazarus et al., 2020; Liebke et al., 2017; Stepp et al., 2009). These interpersonal difficulties co-occur with alterations in socio-cognitive and affective processing, which may both contribute to and arise from repeated negative social experiences. Empirical studies indicate that individuals with BPD struggle to recognize positive social cues and display heightened sensitivity to negative or ambiguous signals (Bertsch et al., 2017; Domes et al., 2008; Fenske et al., 2015; Izurieta Hidalgo et al., 2016; Thome et al., 2016). They also tend to evaluate others more negatively and perceive them as more aggressive or malevolent (Barnow et al., 2009; Hepp, Kieslich, Schmitz, et al., 2021; Hepp, Kieslich, Wycoff, et al., 2021; C. W. Korn et al., 2016; Scheunemann et al., 2023). Consistent with these biases, individuals with BPD expect greater selfishness from others (Lévay et al., 2021) and judge them to be less trustworthy (for a review see Masland et al., 2020). Related tendencies include heightened rejection sensitivity (Cavicchioli & Maffei, 2020; Poggi et al., 2019; Renneberg et al., 2012; Staebler, Renneberg, et al., 2011) and a reduced sense of social connectedness even when being included (De Panfilis et al., 2015; Domsalla et al., 2014; Liebke et al., 2018; Staebler, Helbing, et al., 2011).

Evidence from economic game paradigms illustrates that these interpersonal tendencies can translate into altered social decision-making, but findings are mixed. Several studies report difficulties with

sustaining or repairing cooperation in BPD (Hula et al., 2018; King-Casas et al., 2008; Saunders et al., 2015; Unoka et al., 2009) and patterns point to more self-interested or economically "rational" choices (Jeung et al., 2016), whereas other work finds no overall group differences in prosocial preferences or cooperative tendencies (Graumann et al., 2025; Lévay et al., 2021; Preuss et al., 2016). One reason for these discrepancies may be that different paradigms tap into different forms of cooperation. Individuals with BPD may show relatively preserved *active* cooperation (e.g., willingness to give or share) but greater difficulties in *reactive* cooperation (e.g., forgiving norm violations or re-engaging after unfair treatment) (Hepp et al., 2014; Niedtfeld & Kroneisen, 2020; Thielmann et al., 2014). Moreover, cooperative behavior in BPD appears highly sensitive to situational cues. For instance, factors such as emotional states or social feedback can shape cooperative responses: negative emotional primes have been shown to reduce trustworthiness appraisals in individuals with elevated BPD symptoms (Masland & Hooley, 2020), and impairments in social behavior may be particularly pronounced following unexpected positive or fair interactions or feedback (Abramov et al., 2020; De Panfilis et al., 2019; Liebke et al., 2018; Niedtfeld & Kroneisen, 2020). Such effects may be most detectable in interactions where trust and reciprocity unfold dynamically across time. Accordingly, multi-round designs have repeatedly reported altered trajectories of cooperation in BPD (Abramov et al., 2020; King-Casas et al., 2008; Unoka et al., 2009), with behavioral patterns shifting across phases of interaction (Abramov et al., 2020; Hula et al., 2018; King-Casas et al., 2008). Our own work aligns with this perspective, showing that in anonymous one-shot economic games, individuals with BPD often behave similarly to controls, with comparable active and reactive cooperation as well as similar social preferences and fairness perceptions (Doppelhofer et al., 2025). Taken together, these findings suggest that baseline cooperative tendencies may be relatively preserved in BPD, whereas differences emerge primarily in socially meaningful or repeated interaction contexts.

At a neural level, social decision-making in nonclinical individuals consistently implicates the medial prefrontal cortex (mPFC) as a central region (Frith & Singer, 2008; Hiser & Koenigs, 2018; Lee & Harris, 2013; Lockwood & Wittmann, 2018; Ruff & Fehr, 2014), including in strategic game tasks (Chang et al., 2011; Cutler & Campbell-Meiklejohn, 2019; Decety et al., 2004; McCabe et al., 2001; Sanfey, 2007). The mPFC comprises multiple anatomically and functionally distinguishable subregions that contribute to decision-making via partially dissociable processes (Klein-Flügge et al., 2022; Ongur, 2000; Wittmann et al., 2018). It is centrally involved in theory of mind and mentalizing, enabling individuals to represent and predict others' beliefs, intentions, and actions (Amodio & Frith, 2006; Frith & Frith, 2012; Saxe & Powell, 2006; Schurz et al., 2014). More ventral mPFC regions additionally play a key role in subjective valuation of diverse and often abstract rewards, including socially relevant rewards (Bartra et al., 2013; Haber & Knutson, 2010; Hare et al., 2009; Levy & Glimcher, 2012; Strombach et al., 2015). Cooperative decisions additionally recruit reward-related regions such as the ventral striatum (Bhanji & Delgado, 2014; Cutler & Campbell-Meiklejohn, 2019; Sanfey, 2007) and

regions implicated in the processing of social information and affective signals, including the anterior insula and midcingulate cortex (Apps et al., 2013; Bellucci et al., 2020; Lamm et al., 2011).

Although relatively few fMRI studies have directly examined social decision-making or economic games in BPD, converging evidence from social-cognitive and emotion-processing paradigms indicates systematic alterations in neural networks that support social functioning. These studies consistently point to dysregulation within fronto-limbic circuits (Mier et al., 2013; Schurz et al., 2024; Tebartz Van Elst et al., 2003) involved in social cognition and emotional processing. A robust pattern is heightened reactivity of limbic regions, including the amygdala, accompanied by impaired habituation to affective stimuli and increased sensitivity to both negative and neutral social cues (Donegan et al., 2003; Herpertz et al., 2001; Koenigsberg, Siever, et al., 2009; Lischke et al., 2017; Minzenberg et al., 2007; Schulze et al., 2019). This limbic hyperreactivity is often accompanied by reduced engagement of medial and lateral prefrontal regions, suggesting altered top-down regulation of emotional responses (Koenigsberg, Siever, et al., 2009; Mauchnik & Schmahl, 2010; Minzenberg et al., 2007). Beyond affective dysregulation, task-based social-cognition studies further suggest atypical engagement of mentalizing networks: individuals with BPD show increased activation in regions such as the mPFC, temporoparietal junction, and precuneus during self-other processing (Beeney et al., 2016), and heightened mPFC/dorsal anterior cingulate cortex responses during social interaction contexts including exclusion and inclusion (Domsalla et al., 2014; Ruocco et al., 2010). Together, these findings suggest that neural alterations in BPD are not limited to fronto-limbic mechanisms in affective dysregulation, but also include atypical mPFC engagement during social inference and social interaction – both of which are likely to shape cooperative behavior.

However, few fMRI studies have examined how these systems are recruited during repeated, interactive social decisions in BPD. To address this gap, we developed a social decision-making task framed as a virtual foraging scenario involving the risk of "starvation," in which participants believed they were interacting with another person. The task systematically varied social context – whether the partner's survival depended on the participant's decision or whether the partner was independent (creating a mixed-motive situation) – as well as personal risk. In addition, in a behavioral variant of the task we assessed explicit expectations about the partner's behavior in a role-reversed version of the interaction.

Using this novel framework, we examined whether individuals with BPD differ from age- and intelligence-matched controls (CON) in their cooperative decisions, partner-related expectations, response times, and the recruitment of (medial) prefrontal regions during social choice. This integrative paradigm allows us to clarify how social context and personal risk shape cooperative decision-making and partner expectations in BPD.

Hypotheses

This study was preregistered (https://osf.io/sd2tm), and we briefly summarize the preregistered hypotheses here. The hypotheses are numbered H1 to H6 for reference throughout the article, with full technical details provided in the ***Methods*** section. Note that some preregistered hypotheses are not listed here (i.e., H4-H5 and H7 are missing from the numbering) because they concern additional analyses and outcomes that will be addressed in a separate project.

At the behavioral level, we hypothesized that all participants would cooperate more when their partner depended on them **(H1.1)** and when cooperating was safe rather than risky **(H1.2)**. We further expected these effects to interact, such that participants would cooperate most in safe situations when their partner was dependent **(H1.3)**. We expected individuals with BPD to show reduced cooperation, particularly under dependent-partner and high-risk conditions **(H1.4)**, and to expect less cooperation from their partner compared to controls **(H2)**. We also predicted that participants with BPD would respond more slowly when cooperation involved risk **(H3).**

At a neural level, we expected greater mPFC activity when cooperation was needed by the partner **(H6.1)**, and a group difference in mPFC activity specifically under risky cooperation in the dependent-partner condition **(H6.2)**.

Methods

Ethics and preregistration

We preregistered this study on November 18, 2022, on the Open Science Framework (https://osf.io/sd2tm) during the data collection process but before any analyses of the data. This study is part of a larger project on social decision-making in women with BPD. All participants who participated in this study also took part in other behavioral experiments that are preregistered and reported elsewhere (https://osf.io/w3kz5 and https://osf.io/jkvh8; Doppelhofer et al., 2025). A few minor deviations from the preregistration are explained and reported in the respective methods sections.

Data collection procedures

We performed a power analysis using G*Power, based on the effect sizes reported in a previous study (Christopher W. Korn et al., 2016) on group differences in updating for undesirable feedback in BPD patients and healthy controls. To detect a group difference with 80% power for a one-tailed test at an alpha level of 0.05, a sample size of 30 participants per group (N = 60) was deemed sufficient. Accounting for anticipated dropout rates of approximately 20% in the BPD group and 10% in the control

group, we aimed to collect 30 valid datasets per group—defined as datasets that fulfilled our preregistered inclusion criteria.

Data collection began in October 2021 and was completed in May 2023. Participants were recruited through online advertisements, flyers, and existing databases. For the BPD group, additional recruitment was carried out via local psychotherapists in Heidelberg and neighboring towns, as well as through inpatient units at the Department of General Psychiatry, University of Heidelberg. Control participants were recruited exclusively from Heidelberg.

For the full task battery, participants in the control group received a base compensation of €50, while participants in the BPD group received €60, reflecting the additional time required for diagnostic interviews. All participants could earn additional bonuses based on their performance during experimental tasks. Participants who completed only a subset of experiments were compensated at a standard hourly rate. Travel allowances were provided for BPD participants residing outside Heidelberg.

Participants

Clinical assessments

Axis I disorders were assessed using the short version of the open-access Diagnostic Interview for Mental Disorders (Mini-DIPS Open Access ; Margraf & Cwik, 2017), while BPD diagnostic criteria were evaluated with the German version of the International Personality Disorder Examination (IPDE; Loranger et al., 1994, 1998). All clinical interviews were conducted by trained raters and typically lasted approximately three hours for participants in the BPD group and around one and a half hours for control participants. To estimate general cognitive ability, all participants completed a three-minute speeded reasoning test (mini-q; Baudson & Preckel, 2016). Additional measures included the Beck Depression Inventory-II (BDI-II ; Beck et al., 2006) to assess depressive symptoms, the Brief Symptom Checklist (BSCL; Franke, 2017) to evaluate general psychopathological symptom severity over the past seven days, and the Borderline Symptom List-23 (BSL-23; Bohus et al., 2009) to assess BPD symptom severity. Impairments in personality functioning were measured using the Level of Personality Functioning Scale–Brief Form (LPFS-BF; Spitzer et al., 2021; Weekers et al., 2019), and personality dysfunction was assessed with the Personality Inventory for DSM-5–Brief Form (PID-5-BF; Bach et al., 2020; Zimmermann et al., 2014).

Inclusion and exclusion criteria

We recruited female participants aged 18 to 40 years who were fluent in German. BPD participants were required to meet at least four BPD criteria according to the IPDE, allowing for the inclusion of subthreshold cases with a prior clinical diagnosis. Control participants were matched to BPD participants

by age during the recruitment process. All participants underwent a telephone screening to determine eligibility before being invited to participate in the study.

Participants were excluded from the study based on the following general criteria: presence of severe neurological disorders or history of severe head or brain injuries, severe physical illness, a body mass index (BMI) below 15, severe visual impairments, cognitive disorders, a lifetime diagnosis of schizophrenia, schizoaffective disorder, or bipolar disorder, a current diagnosis of narcissistic or avoidant personality disorder, acute suicidality, enrollment in psychology studies beyond the first semester, and standard contraindications for fMRI participation (e.g., certain metal implants, pregnancy). Additional exclusion criteria were applied for the two groups. For the BPD patient group, individuals with current moderate to severe alcohol or drug dependence or the use of psychotropic medication (except antidepressants, anxiolytics, on-demand medication, and Quetiapine up to 100 mg extended-release and 50 mg immediate-release prescribed for sleep) were excluded. For the control group, further exclusion criteria included current alcohol or drug dependence, a lifetime diagnosis of any personality disorder, schizophrenia, schizoaffective disorder, or bipolar disorder, any other psychiatric diagnosis within the past five years, current or recent (within five years) psychological or psychiatric treatment, the presence of more than one current BPD symptom as assessed by the IPDE, current clinically relevant symptoms determined via diagnostic interview and questionnaires, and the use of psychotropic medication.

We applied additional preregistered exclusion criteria for the experimental tasks. Participants were excluded if they failed to understand the task—indicated either by explicitly stating confusion or by responding in a random manner. Only participants with at least two complete behavioral and (where applicable) fMRI runs after applying these criteria were included in the analysis. Individual trials were excluded if no response was recorded within the allotted time. Experimental runs were excluded if more than 25% of trials were missing due to non-responses; this led to the exclusion of three runs from one participant in the control group. In the fMRI data, runs were also excluded if participants exhibited consistent excessive head movement (i.e., > 3 mm), resulting in the exclusion of one run from one BPD participant, one run from one control participant, two runs from another BPD participant, and three runs from another control participant. One control participant was excluded from all analyses because, after these exclusions, fewer than two complete runs remained.

Sample description

We analyzed all valid datasets for the current study, including 31 participants with BPD and 34 controls (see section **Inclusion and exclusion criteria**). Participants in both groups were aged between 18 and 39 years, with no significant age differences between groups (see **Table 1** for details). General cognitive abilities assessed with the mini-q did not differ significantly between groups (**Table 1**). However,

participants in the BPD group differed significantly from controls in terms of educational attainment, professional qualification, and employment status (**Supplementary Table S1**).

In the BPD group, 28 participants had previously received a formal BPD diagnosis. The remaining three had provisional BPD diagnoses from their treating psychiatrist or psychotherapist; all three were confirmed to meet diagnostic criteria based on the IPDE interview (i.e., ≥ 5 BPD criteria). Additionally, 87.1% of participants in the BPD group reported symptoms of comorbid psychiatric conditions according to the Mini-DIPS interview (**Table 2**), most commonly recurrent major depressive disorder (45.16%), social anxiety disorder (45.16%), and post-traumatic stress disorder (35.48%). Nearly half of the BPD participants reported taking psychotropic medication, primarily antidepressants (**Supplementary Table S2**). Medication intake information was self-reported. The BPD group and the control group differed significantly in BDI-II, BSCL, BSL-23, IPDE, LPFS, and PID-5-BF (see **Table 3** for details).

Table 1. Age and mini-q summary table

	BPD		Control		Welch's t-test						Wilcoxon rank sum test		
	M	*SD*	*M*	*SD*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*p*	*d*	*z*	*p*	*r*
Age	26.23	5.75	24.62	5.29	-1.170	61.104	-4.357	1.140	0.247	-0.291	-1.165	0.244	0.145
Mini-q scores	33.97	12.68	37.00	10.59	-1.041	58.697	-8.861	2.796	0.302	-0.260	-0.920	0.357	0.114

Table 2. Comorbid diagnoses of the BPD group according to the mini-DIPS diagnostic interview

Comorbid Diagnosis	*N*	**%**
Panic disorder	3	9.68
Agoraphobia with panic	1	3.23
Specific phobia	5	16.13
Social anxiety disorder	14	45.16
Major depressive episode	1	3.23
Major depressive disorder, recurrent	14	45.16
Persistent depressive disorder	3	9.68
Body dysmorphic disorder	1	3.23
Posttraumatic stress disorder	11	35.48
Bulimia nervosa	1	3.23
Binge eating disorder	2	6.45
Illness anxiety disorder	2	6.45
Substance abuse, alcohol	1	3.23
Substance abuse, cannabis	2	6.45
No axis I diagnosis	4	12.90

Table 3. Clinical characteristics of both groups

Questionnaire	BPD (*N* = 35)		CON (*N* = 50)		Welch's t-test						Wilcoxon rank sum test		
	M	*SD*	*M*	*SD*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*p*	*d*	*z*	*p*	*r*
BDI-II	28.93	12.46	3.48	3.54	10.801	33.244	20.656	30.241	**< 0.001**	2.779	6.565	**< 0.001**	0.827
BSCL	1.61	0.74	0.17	0.21	10.321	33.387	1.162	1.732	**< 0.001**	2.655	6.571	**< 0.001**	0.828
BSL-23	1.86	0.84	0.19	0.27	10.366	34.302	1.343	1.997	**< 0.001**	2.665	6.631	**< 0.001**	0.835
IPDE	13.63	2.06	0.24	0.66	34.051	34.444	12.592	14.190	**< 0.001**	8.754	7.141	**< 0.001**	0.900
LPFS	35.97	5.23	18.21	5.05	13.675	59.974	15.158	20.352	**< 0.001**	3.452	6.581	**< 0.001**	0.829
PID-5-BF	38.13	8.22	11.42	6.72	14.039	56.182	22.898	30.520	**< 0.001**	3.558	6.644	**< 0.001**	0.837

Note. *CI*: 95% confidence intervals with lower limits (*LL*) and upper limits (*UL*); *d*: Cohen's d (effect size for Welch's t-test); *r*: effect size *r* for Wilcoxon rank sum test; BDI-II: Beck Depression Inventory-II (total score); BSCL: Brief Symptom Checklist (General Severity Index); BSL-23: Borderline Symptom List-23 (mean score); IPDE-Dimensional: International Personality Disorder Examination (dimensional score); LPFS: Level of Personality Functioning Scale-Brief Form (total score); PID-5-BF: Personality Inventory for DSM-5-Brief Form (total score); significant results in bold font

Experimental procedures

Participants were invited to two experimental sessions. The first session was conducted in a behavioral laboratory, where they completed a diagnostic assessment, a set of online questionnaires (**Supplementary Table S3**), training for the decision task, and additional unrelated behavioral experiments.

In the second session, participants were led to believe that they would be performing the experimental task together with another person, with the participant completing the task inside a magnetic resonance imaging (MRI) scanner and the other person in a nearby behavioral lab. They were introduced to a female confederate (hereafter referred to as the "partner") and jointly received instructions for the decision task. After the scanning session, participants took a short break before receiving instructions for the expectation task, again alongside the confederate. They were then taken to separate behavioral labs to complete the expectation task and a series of follow-up questions (e.g., about their task strategy). They were also asked open-endedly whether they had any doubts about the study setup. A few participants reported minor doubts, but none indicated being certain that the setup differed from what they had been told. At the end of the session, participants were fully debriefed.

Both tasks were presented using the MATLAB toolbox Psychtoolbox (www.psychtoolbox.org). Trial and block orders were pseudorandomized. Specifically, four randomized item lists were generated from the original item pool using the MATLAB function randperm, with one randomized list presented per block. To pseudorandomize block order, we manually created four different block order sequences. The specific block order shown to each participant was determined by participant number.

Decision task

The decision task (**Figure 1**) was framed as a virtual foraging game in which participants were instructed that their goal—as well as their partner's—was to "survive" by maintaining at least one energy point across short sequences of trials called mini-blocks.

Each trial began with a fixation period (1-3 seconds; $M = 1.54$, $SD = 0.72$), after which participants had 3.5 seconds to choose between two options: Wait or Play. The Wait option, represented by a black square, resulted in a guaranteed loss of one energy point. The Play option could lead to a gain of one energy point, or a loss of two, depending on the current environment and the joint decisions of the participant and their partner.

Two distinct environments determined the outcome probabilities of the Play option, each occurring with an equal chance (50%). In Environment 1, the probability of gaining energy was 20% if only one player chose to play and 40% if both did. In Environment 2, these probabilities increased to 40% and 80% respectively. Thus, successful outcomes for Play were more likely when both players chose it –

highlighting the cooperative structure of the task. Outcome probabilities for the Play option were visualized using pie charts, where colored segments indicated the winning and losing probabilities (**Figure 1**).

The mini-blocks consisted of either one trial or two consecutive trials, referred to as days. In two-day mini-blocks, trials were separated by a second fixation period (1-3 seconds; $M = 1.39$, $SD = 0.70$) and participants began the second trial with the energy state resulting from the outcome of the first trial. The outcome phase (1.75 seconds) revealed both players' choices and outcomes (shown as numerical energy gains or losses and respective changes in the energy bars). If a player reached zero energy, their bar turned red. On day two, if either the participant or their partner had no remaining energy, they could not make a choice, but the trial still continued.

At the start of each mini-block, both players received initial energy levels ranging from 1 to 3. Participants were told these levels were randomly assigned, but in fact, they were pre-selected to define specific task conditions (see section **Main task conditions** below). Unknown to the participants, their partner's decisions were pre-programmed based on an optimal policy that maximized joint rewards. The full task consisted of four blocks, containing 36 one-day mini-blocks and 18 two-day mini-blocks.

Importantly, while participants were aware that their decisions could affect their partner in certain trials (and vice versa), they were not explicitly instructed to cooperate or compete. Monetary rewards were based solely on the participant's own survival and were unaffected by the partner's survival.

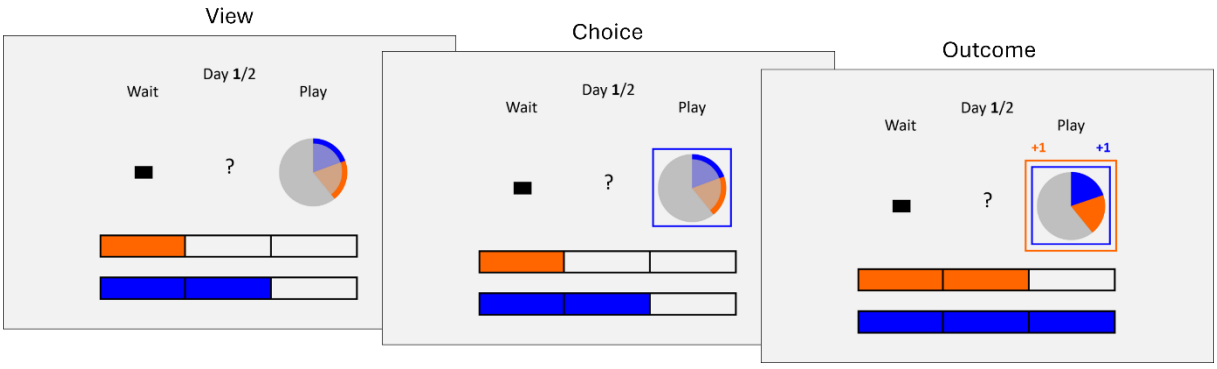


Figure 1. Decision task overview. On each trial, participants choose between a Wait option and a Play option, aiming to maintain a non-zero energy level by the end of each mini-block. Mini-blocks consist of one or two trials ("days"). The display shows the current day, number of days in the mini-block, choice options, success probabilities for the Play option (shown as a pie chart, with the colored segments indicating the chance of gaining energy), and the energy levels of the participant (blue bar) and partner (orange bar). Choices are highlighted (blue = participant; orange = partner), and outcomes are shown after each decision.

Main task conditions

The task comprised two orthogonal experimental factors: the partner factor (social context), which varied whether the partner required help (dependent- vs. independent-partner) and the risk factor, which varied the participant's potential cost when cooperating (high- vs. low-risk) (see **Figure 2**). In the dependent-partner conditions, the partner had only one energy point, placing them at immediate risk of "death" (reaching zero energy) if they did not choose the Play option (as waiting would result in a deterministic loss of one point). The participant, however, had more than one energy point and therefore a choice between waiting and playing. Importantly, by selecting the Play option, they could increase the joint probability of winning, thereby cooperating with the partner. In that sense, the partner's survival largely depended on the participant's choice and choosing Play therefore constituted a helping choice (i.e., a mostly other-regarding choice). In the independent-partner conditions, both the participant and the partner had more than one and an equal number of energy points. In this case, neither is at immediate risk, eliminating the partner's dependency on the participant's decision. The Play choice in independent-conditions could still yield larger gains if both chose it but none of the players' immediate survival depended on it and it was therefore a more mixed-motive option (i.e., a self- and other-regarding choice). Additionally, and orthogonal to the partner factor, the risk factor was manipulated by varying the participant's energy level: in the high-risk conditions, the participant had two energy points and thus risked falling to zero (i.e., "death") if they chose Play and lost, whereas they had three energy points in the low-risk conditions, ensuring that they would retain at least one energy point even after a loss. Note that the naming of factors and conditions in this manuscript differs from the preregistration for improved clarity and readability.

Using these orthogonal experimental factors, we aimed to assess whether participants were sensitive to their partner's outcomes. The partner factor manipulated only the partner's energy level – comparing a dependent partner (one energy point) to an independent partner (sufficient energy) – while keeping the participant's energy level constant. Therefore, any behavioral difference between these two conditions would suggest that participants were influenced by the partner's need for help, indicating concern for their partner's outcome. The risk factor varied the participant's own level of risk, distinguishing between high risk (possibility of reaching zero energy) and low risk (no such risk) when choosing the cooperative option (i.e., Play). Differences between these conditions would reflect a variation in participants' risk-taking behavior. Finally, an interaction between the two factors would suggest that participants' willingness to cooperate was influenced by their risk-taking behavior.

Expectation task

The second task was a modified version of the decision task, differing in two key aspects. First, the initial energy states of the participant and the partner were reversed. In this version, the partner always

began with an energy level equal to or higher than the participant's energy level. In the dependent-partner conditions, the participant started with only one energy point, while the partner had either 2 or 3 – placing the participant in a position of needing help. Second, each trial included an additional estimation phase, in which participants were asked to indicate their expectation about their partner's choice – specifically, whether they believed the partner had chosen Play or Wait. Participants were given 3.5 seconds to make this estimation.

Using the same orthogonal factors—with reversed energy levels and the expectation judgment in place of the participant's cooperation choice—we aimed to test whether participants expected their partner to cooperate when help was needed (dependent- partner vs. independent-partner conditions), and whether this expectation was influenced by the level of risk the partner would incur (high-risk vs. low-risk conditions) by cooperating.

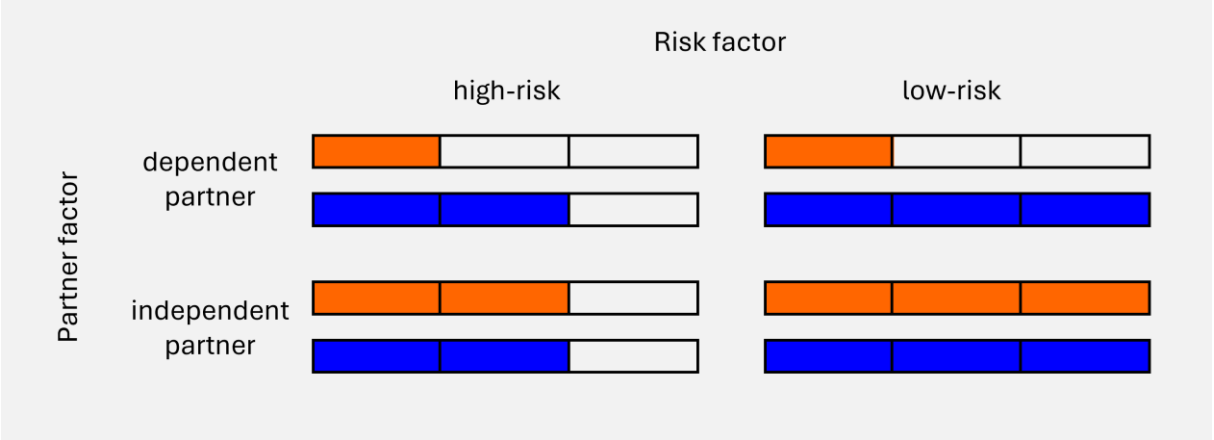


Figure 2. Main task conditions of the decision task. The figure depicts the initial energy distributions between the participant and the partner, which define the main experimental conditions. Energy levels range from 0 to 3 and are represented as horizontal bars, with filled segments indicating the current energy state. In the dependent-partner conditions, the partner has only one energy point and thus depends on the participant's help to increase the winning probability; in the independent-partner conditions, both players have the same number of energy points, and neither is at immediate risk. The risk conditions reflect the participant's potential cost of cooperating (i.e., choosing the Play option): in the high-risk conditions, playing may reduce their energy to zero, while in the low-risk conditions, energy can never drop below one.

MRI acquisition

MRI data were collected using a 3T Siemens Magnetom Prisma scanner equipped with a 64-channel head coil. T2*-weighted functional images were acquired using a gradient-echo echo-planar imaging (EPI) sequence with the following parameters: voxel size = 2.0 × 2.0 × 2.0 mm^3; repetition time

(TR) = 1350 ms; echo time (TE) = 30 ms; flip angle (FA) = 50°; field of view (FoV) = 224 x 224 mm^2; 42 axial slices; 25% interslice gap; GRAPPA acceleration factor = 4; 374 volumes; acquisition time = 8 min 34 s. A dual-echo gradient field map was also obtained to correct for geometric distortions in the functional images. Parameters were as follows: voxel size = 2.0 × 2.0 × 2.0 mm^3; TR = 518 ms; TE1 = 6.12 ms; TE2 = 8.58 ms; FA = 40°; FoV = 224 × 224 mm^2; 42 slices; acquisition time = 7 min 23 s.

High-resolution structural images were acquired with a T1-weighted magnetization-prepared rapid gradient echo (MPRAGE) sequence with the following parameters: voxel size = 1.0 × 1.0 × 1.0 mm^3; TR = 2300 ms; TE = 2.98 ms; inversion time (TI) = 1100 ms; FA = 9°; FoV = 256 x 192 mm^2; 240 slices; acquisition time = 1 min 44 s. All images were aligned to the anterior–posterior commissure axis using the AutoAlign feature, and prescan normalization was applied.

Analyses

Hypotheses list

Behavioral hypotheses

Decision task

H1. We hypothesized that individuals with BPD would cooperate less than controls, particularly when cooperation incurs a personal cost. Note that we define cooperation as choosing the Play option in a way that increases the partner's chance of survival.

We tested this using a mixed-effects model with choice (Play vs. Wait) as the dependent variable, and group (BPD vs. CON), partner factor (dependent vs. independent), and risk factor (high-risk vs. low-risk) as fixed effects, including their interactions. This analysis was conducted on the decision task performed inside the MRI scanner.

- **H1.1.** We hypothesized that participants would cooperate more when the partner depended on them (main effect of partner factor: more Play choices in the dependent-partner conditions).
- **H1.2.** We hypothesized that participants would cooperate more in low-risk than in high-risk conditions (main effect of risk factor).
- **H1.3.** We expected a risk x partner interaction, with reduced cooperation in high-risk versus low-risk trials, particularly in the dependent-partner conditions.
- **H1.4.** We expected a group x partner x risk interaction, such that the interaction described in H1.3 would be more pronounced in the BPD group than in the control group.

Expectation task

H2. We hypothesized that individuals with BPD would expect less cooperation from their partner than controls. Note that here, expecting help is defined as expecting the partner to choose Play (vs. Wait) when help is needed.

We tested this using the same mixed-effects model applied to participants' judgments about their partner's choices (Play vs. Wait) in the expectation task. We predicted a group x partner interaction, with BPD participants expecting less cooperation when they depend on their partner (i.e., in the dependent-partner condition).

Response time

H3. We hypothesized that individuals with BPD would respond more slowly overall and particularly in dependent-partner conditions when cooperation involved risk.

We tested this using the same mixed-effects model on participants' response time (RT) as the dependent variable in the decision task. We predicted a main effect of group (slower RTs in the BPD group) and a group x partner x risk interaction, with the strongest effect (slower RTs) in high-risk, dependent-partner trials among BPD participants.

fMRI hypotheses

H6. We hypothesized that decisions in dependent-partner conditions would be associated with increased mPFC activity.

We modeled BOLD responses using a general linear model (GLM) that included group, partner, risk, and their interaction.

- **H6.1.** We expected greater mPFC activity in dependent- versus independent-partner trials across groups (main effect of partner factor).
- **H6.2.** We expected differences in mPFC activation between the BPD and the control group to be more pronounced in high-risk versus low-risk trials, specifically in the dependent-partner condition (group x partner x risk interaction).

Behavioral analyses

All behavioral analyses were conducted in R (R Core Team, 2024). The decisions during the choice and expectation tasks were analyzed using the lme4 (Bates et al., 2015), car (Fox & Weisberg, 2019), and emmeans (Lenth, 2024) packages.

All models for participants' choices (decisions and expectations) were fitted as generalized linear mixed-effects models (GLMMs) with a binomial distribution and logit link (Laplace approximation), including a random intercept for participant to account for repeated measures.

Predictors were effect-coded (sum-to-zero contrasts), centering the intercept at the grand mean and allowing main and interaction effects to be interpreted as deviations from the overall mean.

Type III Wald χ^2 tests (car::Anova) were used to assess significance of main and interaction effects, and estimated marginal means (EMMs; emmeans) were used to derive interpretable effect sizes. Pairwise contrasts were computed on the log-odds scale and reported as odds ratios (ORs) with 95% confidence intervals.

Analysis dataset

As preregistered, analyses were restricted to the final day of each mini-block (i.e., day 1/1 and day 2/2), because the distinction between high- and low-risk trials is less meaningful on day 1/2 (e.g., the low-risk condition on day 1/2 is not truly low risk if energy must be sustained for the subsequent trial).

Deviations from the preregistration

We note two minor deviations/additions from our preregistration: First, although the preregistration stated that we would analyze proportions of Play choices, we instead analyzed trial-level data using a GLMM. This approach is more appropriate for the nested structure of the data and aligns with our original intent. Second, we tested H1-H4 using reduced models to avoid overfitting (see below for details).

Decision task

Outcome and predictors

The binary outcome was participants' choices in the decision task (coded as Play vs. Wait). Categorical predictors were group (BPD vs. CON), partner (dependent vs. independent), and risk (high vs. low).

Model specification

Three nested models were fit:

Model 1 (full; preregistered):

choice ~ group * partner * risk + (1 | participant)

Model 2 (no three-way interaction):

choice ~ group * partner + group * risk + partner * risk + (1 | participant)

Model 3 (no partner x risk interaction):

choice ~ group * partner + group * risk + (1 | participant)

Model comparison and inference

Likelihood ratio tests (LRT; anova(…, test = "Chisq")) were used to compare nested models. We compared the full Model 1 to a reduced Model 2 to evaluate the group x partner x risk interaction (**H1.4**). As the three-way interaction did not improve model fit, hypotheses **H1.1-H1.3** were tested using Model 2, which retains all two-way interactions and reduced overfitting. As an additional check of **H1.3** (partner x risk interaction), we compared Model 2 to Model 3 (see **Supplementary Tables S4-S8** for full model summaries and model comparisons).

Follow ups and effect sizes

(1) Main effects of partner and risk (**H1.1-H1.2**). Given the significant interactions in Model 2, we assessed the global interpretability of the partner and risk main effects by testing Bonferroni-adjusted pairwise contrasts of EMMs to verify whether the effect directions were consistent and the differences significant across all combinations of the other factors.
(2) Risk x partner interaction (**H1.3**). We conducted Bonferroni-adjusted pairwise comparisons of EMMs comparing high- versus low-risk conditions at each level of partner.
(3) Group effect (**H1.4**). Since the three-way interaction was not significant, we interpreted group differences using Model 2. We obtained Bonferroni adjusted pairwise contrasts of EMMs comparing the BPD versus control group at each level of risk.

Expectation task

Outcome and predictors

The binary outcome was participants' expectations about whether their partner would help them (coded as expected partner choice to Play vs. Wait).Predictors were identical to the decision task: group, partner, and risk.

Model specification

Three nested models were fit following the same logic as the decision task:

Model 1 (full; preregistered):

expectation ~ group * partner * risk + (1 | participant)

Model 2 (no three-way interaction):

expectation ~ group * partner + group * risk + partner * risk + (1 | participant)

Model 3 (group x partner only):

expectation ~ group * partner + (1 | participant)

Model 3 was tested because the preregistered hypothesis (**H2**) specifically focused on the group x partner interaction and did not involve the risk factor.

Model comparison and inference

Likelihood ratio tests were used to compare nested models. Removing the three-way interaction from the preregistered Model 1 did not significantly reduce the model fit. However, excluding the risk factor and its two-way interactions (Model 2 vs. Model 3) significantly reduced the model fit. Consequently, Model 2, which includes all main effects and two-way interactions, was used for all subsequent analyses (see **Supplementary Tables S9-S12** for full model summaries and model comparison).

Follow ups and effect sizes

(1) Group x partner interaction (**H2**). To test whether participants in the BPD group would expect less cooperation from their partner than controls, specifically in the dependent-partner

condition, we obtained Bonferroni-adjusted pairwise contrasts of EMMs comparing the BPD and control groups at each partner condition, averaged over risk levels.

Response time analyses

RTs were analyzed to examine differences between groups and across conditions in both the cooperation and expectation tasks. RTs were log-transformed to correct for positive skew and analyzed using linear mixed-effects models (LMMs) implemented in the lmerTest package (Kuznetsova et al., 2017). Each model included a random intercept for participant to account for repeated measures. Type III F-tests with Satterthwaite's approximation (implemented in the base stats package) were used to assess the significance of main and interaction effects.

We specified three nested models to maintain comparability with the behavioral analyses:

Model 1 (full; preregistered):

log(RT) ~ group * partner * risk + (1 | participant)

Model 2 (no three-way interaction):

log(RT) ~ group * partner + group * risk + partner * risk + (1 | participant)

Model 3 (no partner x risk interaction):

log(RT) ~ group * partner + group * risk + (1 | participant)

Model comparisons were conducted using likelihood-ratio tests (LRT; anova(…, test = "Chisq")).

RTs decision task

For the choice task, we tested the preregistered hypothesis (**H3**) that participants with BPD would respond more slowly overall and particularly under high-risk dependent-partner conditions. The LRTs indicated that the three-way interaction did not improve model fit (Model 1 vs. Model 2), but retaining the partner x risk interaction (Model 2 vs. Model 3) significantly improved fit. Therefore, Model 2 was used for inference (see **Supplementary Tables S13-S17** for full model summaries and model comparison).

Given the significant partner x risk interaction, we followed up with Bonferroni-adjusted pairwise contrasts of EMMs comparing RTs for dependent- and independent-partner conditions in each level of

risk, averaged over groups. EMM contrasts were computed on the response scale yielding ratios of geometric mean RTs that express proportional changes in response speed.

RTs expectation task

Analogous models were fit to the expectation task to explore whether the RTs differ between groups and across conditions. Model comparisons showed that neither the three-way (Model 1 vs. Model 2) nor the partner x risk interaction (Model 2 vs. Model 3) improved model fit. Thus, the more parsimonious Model 3 was used for inference. The same Type III ANOVA procedure was applied to assess main effects and interactions (see **Supplementary Tables S18-S22** for full model summaries and model comparison). Because only the partner main effect was significant, we computed the corresponding pairwise contrasts of the partner factor on the response scale to quantify the proportional differences in RTs.

fMRI analyses

fMRI preprocessing

Imaging data preprocessing was performed using SPM12 (www.fil.ion.ucl.ac.uk/spm) running on MATLAB 2024a (The MathWorks Inc, 2024). The first two scans of each run were discarded to allow for T1 equilibration. A voxel displacement map (VDM) was calculated from the phase and magnitude field map images to correct for geometric distortions. Functional images were realigned to the first image and unwarped using the VDM. The volumes were slice-timing corrected to the middle slice of each run. Each participant's T1-weighted structural scan (defaced prior to analysis to ensure anonymity) was co-registered to the mean functional image and segmented into tissue types. Normalization parameters were applied to warp structural and functional images to Montreal Neurological Institute (MNI) template space. Finally, normalized functional images were smoothed using an isotropic Gaussian kernel with a full width at half maximum (FWHM) of 6mm.

fMRI data analysis

Whole-brain univariate fMRI analysis

For the first-level analysis, we modeled each participant's blood oxygenation level-dependent (BOLD) response during the choice and outcome phases using a GLM implemented in SPM. We defined six regressors for the choice phase conditions. Four regressors of interest corresponded directly to the behavioral model factors: trials involving high-risk in the dependent-partner condition, low-risk in the dependent-partner condition, high-risk in the independent-partner condition, and low-risk in the

independent-partner condition. Additionally, we included two nuisance regressors: one regressor for trials in which the participant had exactly one energy point (task-understanding control-condition), and one regressor that grouped all remaining trials (including energy point combinations outside the predefined conditions, missed response trials, and trials with zero energy points). Choice events were modeled as boxcar regressors corresponding to the full choice period (duration = 3.5 s), convolved with the canonical hemodynamic response function (HRF). The outcome phase was modeled with two regressors (duration = 1.75s) that distinguished positive (>0) and negative (≤0) energy outcomes for the participant (independent of the partner's outcome), also convolved with the HRF. Six realignment parameters per run were included as nuisance regressors, a high-pass filter of 128s was applied to remove low-frequency signal drifts, and each run included its own constant term to model baseline differences.

Group-level effects were assessed using two flexible factorial models in SPM. The choice GLM included the factors group (BPD vs. CON), partner (dependent vs. independent), risk (high vs. low) and estimated the main effects, all two-way interactions (group x partner, group x risk, partner x risk), and the three-way interaction (group x partner x risk).

The outcome GLM included the factors group (BPD vs. CON) and outcome for the participant (positive vs. negative) and estimated the main effects and the group x outcome interaction. Both models included subject as a random effect to account for repeated measures.

For both the choice (2 x 2 x 2) and outcome (2 x 2) models, contrasts were defined by placing weights on the design-matrix columns to obtain valid marginal effects under unequal group sizes, following the procedure described by Gläscher & Gitelman (2008, *Contrast weights in flexible factorial design with multiple groups of subjects*, tutorial, available on ResearchGate). Group was coded using ±1/n within each group, partner and risk used within-factor difference weights averaged across groups in proportion to sample size, and interactions were specified as standard differences-of-differences. Whole-brain inference was performed using cluster-level correction based on Gaussian random field theory (cluster-defining threshold $p < .001$ and a cluster-probability of $p < .05$ family-wise error (FWE) correction). Anatomical labels for significant clusters were assigned based on the Automatic Anatomical Labeling atlas AAL3 (AAL3; Rolls et al., 2020) using the toolbox bspmview (Spunt, 2016).

We made minor adjustments to the preregistered analysis to ensure appropriate modeling of all task events. Specifically, all trials were included in the analysis rather than only the last day of each mini-block to allow for complete representation of task-related variance. In addition, a second nuisance regressor was modeled to account for variance associated with unmodeled task events.

Univariate ROI analysis

Preregistered univariate regions of interest (ROIs) analyses were conducted to test group x partner x risk effects within medial prefrontal regions. Two ROIs were defined a priori. The mPFC ROI was derived

from a meta-analysis that used a clustering approach to identify neural patterns involved in social-cognitive processes (Schurz et al., 2021). Specifically, we selected the mPFC cluster (including the anterior cingulate cortex, ACC) from the activation map representing common activation during strategic games involving cooperative and competitive interactions. The dorsomedial prefrontal cortex (dmPFC) ROI was based on a previous study on heuristic and optimal decision-making strategies (Korn & Bach, 2018). We used the peak voxel reported for the dmPFC cluster showing a positive relationship between BOLD activity and success probability (reflecting the heuristic policy computations) in a sequential decision-making task. To ensure midline location, we shifted the x-coordinate of the reported peak from x = 8 to x = 0 while keeping y = 29 and z = 50 constant, and created a spherical ROI with a 10 mm radius centered on the adjusted coordinates. Both ROIs were extracted using the MarsBaR toolbox (Brett et al., 2002) and resampled to match the resolution of the functional data.

For each ROI, we extracted the first-level contrast estimates corresponding to the four partner x risk conditions (high-risk and dependent-partner, low-risk and dependent-partner, high-risk and independent-partner, low-risk and independent-partner). These estimates were analyzed using LMMs implemented in R with the lmerTest package (Kuznetsova et al., 2017) and the emmeans package (Lenth, 2024) for follow-up comparisons. Group (BPD vs. CON), partner (dependent vs. independent), and risk (high vs. low) were treated as categorical predictors. All predictors were effect-coded using sum-to-zero contrasts, such that main effects and interactions reflect deviations from the grand mean rather than contrasts with a single reference category. The preregistered model included fixed effects of group, partner, and risk and their interactions, with a random intercept for participant.

ROI Model:

ROI estimate ~ group * partner * risk + (1 | participant)

Statistical inference was based on Type III F-tests with Satterthwaite's approximation (implemented in the base stats package). Significant interactions were followed up with pairwise contrasts of EMMs with false discovery rate (FDR) correction.

In addition to parameter-extraction analyses, we conducted voxel-wise mask-based cluster inference within each ROI, restricting the statistical search volume to the preregistered ROI mask in SPM12. Mask-based analyses followed the same procedure as the whole-brain analysis: clusters were considered significant if they survived family-wise error (FWE) correction at $p < 0.05$ at the cluster level, using a voxel-wise cluster-defining threshold of $p < 0.001$. Whole-brain and ROI analyses were thus performed in parallel using identical contrasts and correction thresholds.

Results

Behavioral results

Decision task

*M*ain Effect of Partner Dependence (H1.1)

There was a significant main effect of the partner factor (Model 2, Type III ANOVA; **Table 4**). The partner factor also interacted significantly with risk but not with group.

Participants were substantially more likely to cooperate in the dependent-partner condition compared to the independent-partner condition. Averaged over groups and risk (EMMs contrasts), participants were more than ten times as likely to cooperate when help was needed (dependent-partner condition) than when it was not needed (independent-partner condition; $OR = 11.50$, 95% CI [10.16, 13.01], $p < 0.001$).

EMMs within each group x risk combination pointed in the same direction (dependent > independent) and all EMMs contrasts remained significant after Bonferroni correction (p-adj < 0.001 for all contrasts; see **Supplementary Tables S23-S24**), supporting a robust global effect of the partner factor.

In line with our hypothesis, participants cooperated substantially more in the dependent-partner condition compared to the independent-partner condition, and this effect was robust across all group x risk contexts (see **Figure 3a** for visualization of participant cooperative choices across groups and conditions).

Table 4. Model summary of the effects of group, partner, and risk and their interactions on (Play) choices (Type III Wald Chi-Square Tests for decision task Model 2)

Term	*Chisq*	*df*	*p*
(Intercept)	10.141	1	**0.001**
Group	2.962	1	0.085
Partner	1,500.332	1	**0.000**
Risk	6.623	1	**0.010**
Group x Partner	3.516	1	0.061
Group x Risk	16.211	1	**0.000**
Partner x Risk	57.627	1	**0.000**

Main Effect of Risk (H1.2)

We found a significant main effect of risk (Model 2, Type III ANOVA; **Table 4**), which interacted with both group and partner.

The direction of the risk effect reversed across contexts, and therefore the significant main effect of risk, was not globally interpretable (see **Supplementary Tables S25-S26** for EMMs**)**. Instead, risk influenced choices in a context dependent manner. Play choices decreased under high risk when the partner was dependent, whereas the opposite tendency emerged when partner was independent (see partner x risk interaction below for details).

Thus, our second hypothesis, that participants would cooperate more in low-risk compared to high-risk conditions, was only partially supported. Although we found a significant effect of risk, there was no global effect across group x partner contexts.

Partner × Risk Interaction (H1.3)

The partner x risk interaction was significant in Model 2 (Type III ANOVA; **Table 4**). An LRT comparing Model 2 (with the partner x risk interaction) to Model 3 (without it) confirmed that including the interaction improved the model fit ($\chi^2(1) = 58.89$, p-adj < 0.001; **Supplementary Table S8**).

Averaged over groups, EMM contrasts showed that in the dependent-partner condition, cooperation was lower at high risk compared to low risk (*OR* = 0.547, 95% *CI* [0.476, 0.629], p-adj < 0.001). In contrast, in the independent-partner condition, cooperation was higher at high risk compared to low risk (*OR* = 1.347, 95% *CI* [1.074, 1.688], p-adj = 0.006; **Supplementary Table S27**).

These results indicate that risk influenced the partner effect. Consistent with our hypothesis, risk reduced cooperation when the partner depended on help, whereas the opposite pattern emerged when the partner did not need help.

Group × Partner × Risk Interaction (H1.4)

The three-way interaction was not significant in the full model (Model 1; *OR* = 1.04, p = 0 .230; **Supplementary Table S4**), and an LRT indicated that removing the three-way term did not reduce model fit ($\chi^2(1) = 1.44$, p = 0.230; **Supplementary Table S7**). Accordingly, we based inference on Model 2, which included the lower-order interactions. In Model 2, the group x risk interaction was significant (**Table 4**), indicating that risk-related changes in choices differed between groups when averaged across partner conditions.

When averaged over partner levels, we found that groups did not differ at high risk (*OR* = 1.344, 95% *CI* [0.687, 2.629], *p*-adj = 0.647). However, at low risk, controls cooperated more than participants in the BPD group (*OR* = 2.063, 95% *CI* [1.047, 4.063], *p*-adj = 0.033; **Supplementary Table S28**).

While these findings do not support our hypothesis that the partner x risk interaction would be more pronounced in the BPD group than in controls, we did find that controls cooperated (chose Play) more than BPD participants under low-risk conditions when averaged across partner levels (**Figure 3b**).

Summary

We found that cooperation was primarily shaped by the factor partner dependence. Differences between the BPD and control groups only appeared under low-risk conditions, where the control group cooperated more than the BPD group.

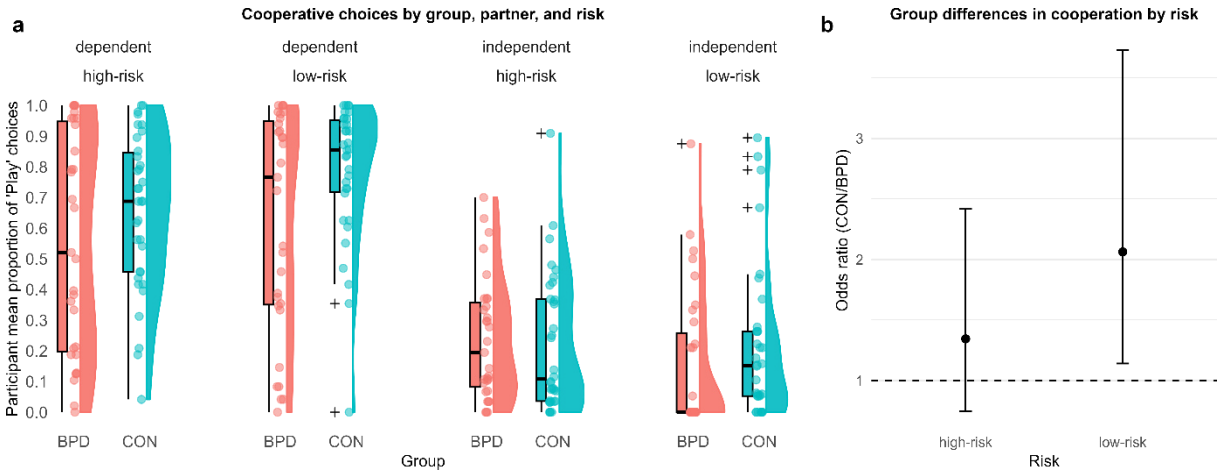


Figure 3. Average cooperative choices and model-based group effects. **(a)** Raincloud plots show participants' mean proportions of Play choices by group (BPD, CON), shown separately for each partner condition (dependent, independent) and risk condition (high, low). These participant-level averages are descriptive and are not intended for statistical inference about group differences. Points show participant means; boxplots indicate the median and interquartile range, and half-violins show smoothed distributions. **(b)** Model-based estimates of the group effect by risk (averaged over partner), expressed as odds ratios (ORs; CON/BPD) derived from mixed-effects logistic regression on trial-level data (Model 2) with post-hoc contrasts. Points represent estimated ORs and error bars show 95% confidence intervals (back-transformed from the log-odds scale); the dashed line at *OR* = 1 indicates no group difference. Consistent with the mixed-model results, the control group showed higher odds of choosing Play than the BPD group at low risk, whereas the group difference was not significant at high risk.

Expectation task

Group x Partner Interaction (H2)

An LRT comparing the preregistered full model (Model 1) to a reduced model without the three-way term (Model 2) showed no significant loss of fit ($\chi^2(1) = 1.578$, $p = 0.209$; **Supplementary Table S12**). In contrast, removing the Risk factor and its two-way interactions (Model 2 vs. Model 3) significantly worsened model fit ($\chi^2(3) = 345.724$, $p < 0.001$; **Supplementary Table S12**). Model 2, which included all main effects and two-way interactions, therefore provided best fit and was used for the follow-up analyses.

A type III ANOVA on Model 2 revealed significant main effects of partner and risk but not group (**Table 5**). We also found significant group x partner and partner x risk interactions, whereas the Group x Risk interaction was not significant.

When averaged over risk levels, groups differed significantly in the dependent-partner condition (*OR* = 2.108, 95% *CI* [1.221, 3.640], $p = 0.004$), but not in the independent-partner condition (*OR* = 0.751, 95% *CI* [0.393, 1.433], $p = 0.641$; **Supplementary Table S29**). Specifically, controls expected approximately twice as much cooperation from their partners as participants with BPD did when they dependent on their partner's help (see **Figure 4a** for visualization of participant expectations about partner cooperative choices across groups and conditions and **Figure 4b** for the specific group effect).

Table 5. Model summary of the effects of group, partner, and risk and their interactions on (Play) partner expectations (Type III Wald Chi-Square Tests for expectation task Model 2)

Term	*Chisq*	*df*	*p-value*
(Intercept)	42.095	1	**<0.001**
Group	0.859	1	0.354
Partner	756.400	1	**<0.001**
Risk	118.161	1	**<0.001**
Group x Partner	26.775	1	**<0.001**
Group x Risk	3.478	1	0.062
Partner x Risk	22.536	1	**<0.001**

Because we had no a priori hypothesis for a group x risk effect on expectations, we treated this analysis as exploratory. The omnibus group x risk interaction was not significant (**Table 5**), indicating no reliable evidence that the risk effect differed between groups. To assess whether participants in each group we conducted exploratory post-hoc contrasts to confirm whether each group nevertheless responded to the risk manipulation, we examined within-group post-hoc contrasts. These showed that both controls and participants with BPD were sensitive to risk in the same direction, expecting less cooperation under high risk than under low risk (controls: *OR* = 0.301, 95% *CI* [0.227, 0.401], *p*-adj < 0.001; BPD: *OR* = 0.403, 95% *CI* [0.307, 0.528], *p*-adj < 0.001; **Supplementary Table S30**).

In line with our hypothesis, patients with BPD expected their partners to cooperate less than controls in situations where their survival depended on their partner's help.

Exploratory contrasts further indicated that both groups were similarly sensitive to risk, expecting less partner cooperation under high-risk than low-risk conditions.

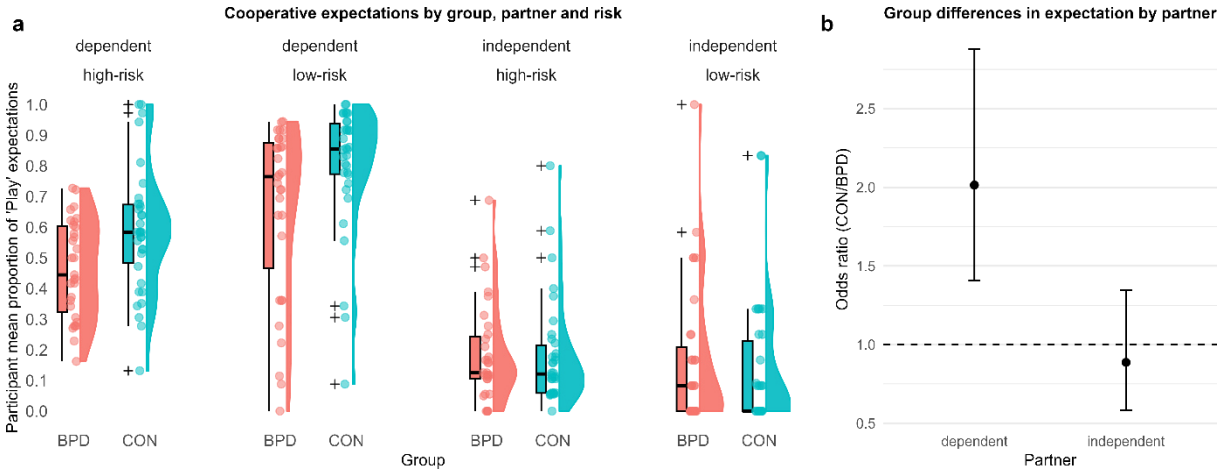


Figure 4. Average expectations about cooperative partner choices and model-based group effects. **(a)** Raincloud plots show participants' average partner expectations (i.e., whether the partner will choose Play) by group (BPD, CON), shown separately for each partner condition (dependent, independent) and risk condition (high, low). These participant-level averages are descriptive and are not intended for statistical inference about group differences. Points show participant means; boxplots indicate the median and interquartile range, and half-violins show smoothed distributions. **(b)** Model-based estimates of the group effect by partner (averaged over risk), expressed as odds ratios (ORs; CON/BPD) derived from mixed-effects logistic regression on trial-level data (Model 2) with post-hoc contrasts. Points represent estimated ORs and error bars show 95% confidence intervals (back-transformed from the log-odds scale); the dashed line at OR = 1 indicates no group difference. Consistent with the mixed-model results, the control group showed higher odds of expecting the partner to choose the cooperative option Play than the BPD group for dependent-partner conditions, whereas the group difference was not significant for independent-partner conditions.

Response time analyses

RTs decision task

We tested the preregistered hypothesis that participants with BPD would respond more slowly overall and, in particular, in high-risk and dependent-partner trials.

An LRT comparing the full model (Model 1) to a reduced model without the three-way interaction (Model 2) showed no significant loss of fit ($\chi^2(1) = 1.505$, $p = 0.220$; **Supplementary Table S16**). In contrast, removing the partner x risk interaction (Model 3) significantly worsened model fit ($\chi^2(1) = 65.113$, $p < 0.001$; **Supplementary Table S16**). Model 2, which included all main effects and two-way interactions, therefore provided the best fit and was used for subsequent inference.

A Type III ANOVA on Model 2 revealed significant main effects of partner ($F(1,9004.3) = 330.52$, $p < 0.001$) and risk ($F(1, 9004.3) = 9.95$, $p = 0.002$), as well as a significant partner x risk interaction ($F(1, 9004.2) = 65.31$, $p < 0.001$). The effects of group ($F(1, 63.4) = 0.70$, $p = 0.406$) and all interactions involving group were non-significant (all $p > 0.5$; **Supplementary Table S17**).

Follow-up EMMs confirmed that participants responded about 8% more slowly in high-risk than low-risk trials when the partner was dependent (*ratio* = 1.08, 95% *CI* [1.06, 1.10], $p < 0.001$), but about 3% faster under high risk when the partner was independent (*ratio* = 0.97, 95% *CI* [0.94, 0.99], $p = 0.006$; **Supplementary Table S31**).

In line with our hypothesis, participants were slower in risky cooperative contexts, particularly when the partner depended on their help, although this effect did not differ by group.

RTs expectation task

To explore whether the RTs differ between groups when participants made expectation judgments about their partners' choices, we applied the same models. LRTs indicated that removing the three-way interaction (Model 1 vs. Model 2) did not significantly reduce model fit ($\chi^2(1) = 1.874$, $p = 0.171$) and dropping the partner x risk term (Model 2 vs. Model 3) also did not worsen the fit ($\chi^2(1) = 2.744$, $p = 0.098$; **Supplementary Table S21**). The more parsimonious Model 3, including main effects and the group x partner, and group x risk interactions, was therefore retained for inference.

A Type III ANOVA on Model 3 revealed a significant main effect of partner ($F(1, 4431.4) = 16.89$, $p < 0.001$), indicating that responses were slower in the dependent-partner condition. No other main effects or interactions reached significance (all $p > 0.05$; **Supplementary Table S22**).

Follow-up pairwise comparisons of the partner factor confirmed that participants responded about 13% more slowly in the dependent-partner condition compared to the independent-partner condition (*ratio* = 1.13, 95% *CI* [1.06, 1.19], $p < 0.001$).

When making judgments about their partner's choices, participants took longer when they depended on their partner compared to when they did not depend on their partner, irrespective of group or risk.

fMRI results

Choice phase (decision phase)

We examined whether social context and risk were related to BOLD responses in the mPFC. We implemented a GLM with the factors group (BPD vs. CON), partner (dependent vs. independent), and risk (high vs. low) during the choice phase. We hypothesized that mPFC activation would be greater in the dependent- compared to the independent-partner condition across groups (**H6.1**), and that mPFC activity would differ between groups for high- versus low-risk trials, specifically in the dependent-partner condition (**H6.2)**. All reported results are whole-brain FWE corrected at the cluster-level ($p < 0.05$) using a voxel-wise cluster-defining threshold of $p < 0.001$.

Main effect of social context (partner)

During the choice phase, the dependent-partner condition, compared to the independent-partner condition was associated with significantly increased BOLD signal in several regions (**Figure 5**). Higher activation was found in the right inferior frontal gyrus (IFG, opercular and triangular parts; peak voxel: $t = 6.136$, MNI: 36, 6, 28), the left precentral gyrus extending into the left IFG (opercular and triangular parts; peak voxel: $t = 5.819$, MNI: -36, 20, 20), and bilateral superior frontal gyrus (left: $t = 5.448$, MNI: -24, 2, 48; right: t = 4.878, MNI: 24, 52, 28). Additionally, a large posterior cluster spanning the right to left middle occipital gyrus and including the bilateral precuneus showed increased activity (peak voxel: t = 9.422, MNI: 36, -88,14). Full statistics for all significant clusters are reported in **Supplementary Table S32**.

In contrast to our hypothesis, the independent-partner condition (independent > dependent) elicited significantly higher BOLD signal in mPFC regions (**Figure 6**), including the medial orbitofrontal /superior frontal gyrus (peak voxel: $t = 6.326$, MNI: 6, 52, -12). Increased activity was also observed in memory- and salience related regions, including the right parahippocampal gyrus extending into the hippocampus and insula (peak voxel: $t = 5.973$, MNI = 18, -8, -20) and the left hippocampus (peak voxel: $t = 5.129$, MNI: -24, -16, -20). Additional increases were detected in lateral prefrontal regions,

including the right IFG (triangular part; peak voxel: t = 5.798, MNI: 52, 38, 2), the left IFG (pars orbitalis; peak voxel: t = 4.836, MNI: -38, 36, -12), the left IFG (triangular part; peak voxel t = 4.698, MNI: -52, 34, 0), and the left middle cingulate/paracingulate gyrus (peak voxel: t = 4.691, MNI: -8, 46, 26). Full statistics are provided in **Supplementary Tables S33**.

Mask-based ROI inference

To increase sensitivity within the preregistered mPFC ROI, we additionally performed voxelwise mask-based cluster correction. No voxels survived mask-based FWE correction for the dependent > independent contrast. For the independent > dependent contrast, one significant cluster emerged within the left medial superior frontal gyrus, extending into the pregenual ACC (peak voxel: t = 4.576, MNI: -8, 46, 26; **Supplementary Table S34**). No voxels within the preregistered dmPFC ROI survived mask-based FWE correction for either direction of the partner contrast.

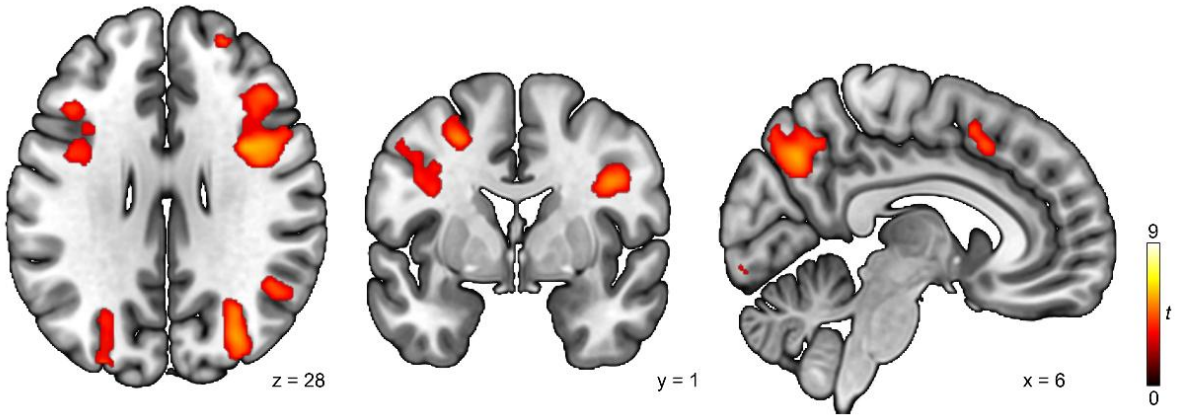


Figure 5. Main effect of social context (dependent > independent partner). Group-level statistical parametric map for the contrast dependent partner > independent partner during the choice phase, overlaid on the MNI152 template. The dependent (vs. independent) partner condition showed increased activity in lateral frontal regions (including bilateral inferior frontal gyrus), with additional engagement of middle and superior frontal areas and posterior regions including the precuneus and middle occipital cortex. Results are cluster-level FWE-corrected ($p < 0.05$) using a voxel-wise cluster-defining threshold of $p < 0.001$; colors indicate t values.

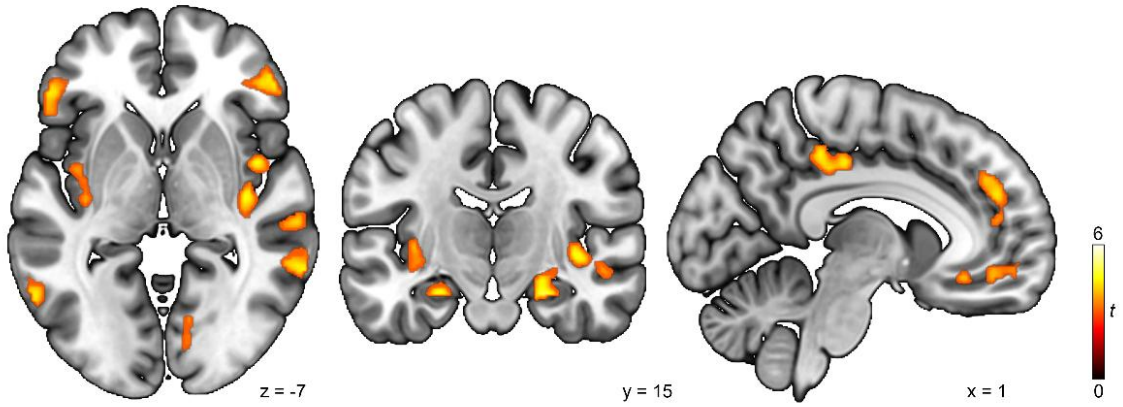


Figure 6. Main effect for social context (independent > dependent partner). Group level- statistical parametric map for the contrast independent partner > dependent partner during the choice phase, overlaid on the MNI152 template. The independent (vs. independent) partner condition showed increased activity in medial prefrontal regions (including medial orbitofrontal/medial superior frontal cortex, extending into midcingulate/pregenual ACC), alongside recruitment of lateral frontal regions and memory/salience-related regions, including the insula and bilateral hippocampus. Results are cluster-level FWE-corrected ($p < 0.05$) using a voxel-wise cluster-defining threshold of $p < 0.001$; colors indicate t values.

Group x partner x risk interaction

For the preregistered contrast (BPD > CON, dependent > independent, high > low), no clusters survived whole-brain cluster-level FWE correction. Thus, we did not observe the expected increase in mPFC activation for the BPD group during high-risk trials in the dependent partner condition. In contrast, the opposite contrast (CON > BPD, independent > dependent, low > high) revealed significant clusters in prefrontal and lateral regions. Higher BOLD signal was found in the left and extending into right superior frontal gyrus, medial orbital (peak voxel: $t = 4.795$, MNI: -4, 32 -12) and bilateral IFG (left: $t = 4.435$, MNI: -52, 28, 12; right: $t = 4.320$, MNI: 52, 24,12). Full statistics for all significant clusters are provided in **Supplementary Table S35**.

Our results suggest greater prefrontal engagement in controls under low-risk, independent-partner conditions (**Figure 7**).

Mask-based ROI inference

No voxels within the preregistered mPFC and dmPFC ROI survived mask-based FWE correction for either correction of the three-way interaction.

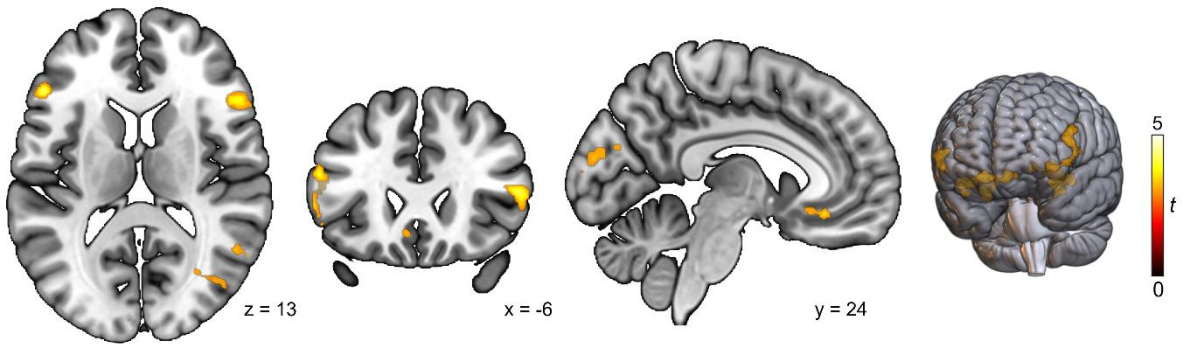


Figure 7. Group x partner x risk interaction during the choice phase. Group-level statistical parametric maps for the three-way interaction. The group x partner x risk interaction (CON > BPD x independent > dependent x low > high) revealed significant clusters in prefrontal cortex, including medial orbitofrontal/superior frontal regions and bilateral inferior frontal gyri. The preregistered contrast in the opposite direction (BPD > CON x dependent > independent, high < low) yielded no clusters surviving whole-brain cluster-level FWE correction. Maps are overlaid on the MNI152 template. Results are cluster-level FWE-corrected ($p < 0.05$) using a voxel-wise cluster-defining threshold of $p < 0.001$; colors indicate t values.

Exploratory analyses

In addition to the preregistered analyses, we conducted exploratory whole-brain analyses to examine other main and interaction effects.

Main effect of risk

Exploratory analyses revealed widespread effects of risk levels on BOLD activity. High- compared to low-risk trials were associated with increased activation in the bilateral insula (left: $t = 7.959$, MNI: -34, 26, 0; right: $t = 7.613$, MNI: 34, 22, 10); the left insular cluster extended into the left IFG. An additional cluster was observed in the middle cingulate cortex (peak voxel: $t = 6.781$, MNI: -6, 20, 36).

The reverse contrast (low > high) showed greater activation in prefrontal and temporoparietal regions, including the left middle frontal gyrus (peak voxel: $t = 7.320$, MNI: -44, 38 16), right IFG (triangular part; peak voxel: $t = 6.711$, MNI: 44, 40, 0), and bilateral supramarginal gyri. All significant clusters are reported in **Supplementary Tables S36** and **S37**.

Mask-based ROI inference

When restricting the statistical search volume to the preregistered mPFC ROI, the high > low contrast revealed one significant cluster located in the left middle cingulate/paracingulate gyri (peak voxel: $t = 6.781$, MNI: -6, 20, 36), extending into the bilateral medial superior frontal gyrus. For the low > high contrast, a significant cluster emerged in the right medial superior frontal gyrus (peak voxel: $t = 5.341$, MNI: 12, 64, 18) extending into the left medial superior frontal gyrus (**Supplementary Tables S38** and **S39**; see **Figure 8** for significant mPFC ROI effects). No clusters within the preregistered dmPFC ROI survived mask-based FWE correction for either direction of the risk contrast.

Two-way interactions

We also explored all contrasts for two-way interactions. No significant effects were found for the group x partner interactions (both contrasts), and no relevant effects emerged for the group x risk interaction apart from a single cluster in the calcarine fissure for the BPD > CON contrast (see **Supplementary Table S40**). For the Partner x Risk interaction, no effects were observed for the dependent > independent contrasts, whereas the independent > dependent contrast revealed activation in the right precuneus (peak voxel: $t = 5.901$, MNI: 6, -56, 20), the left caudate nucleus (peak voxel: $t = 5.390$, MNI: -10, 24, 16), and the subgenual ACC ($t = 5.015$, MNI: 0, 36, -4). Full results for all two-way interaction contrasts are reported in **Supplementary Table S41**.

Mask-based ROI inference

Within the preregistered mPFC ROI, only one contrast showed a significant effect: the independent > dependent, low > high partner x risk contrast revealed a cluster in the subgenual ACC (peak voxel: $t = 5.015$, MNI: 0, 36, -4), extending into bilateral supracallosal, subgenual, and pregenual ACC (**Supplementary Table S42**; see **Figure 8** for significant mPFC ROI effects). No other two-way interaction contrasts produced significant clusters within the mPFC mask. No voxels within the dmPFC ROI survived mask-based FWE correction for any of the two-way interaction contrasts.

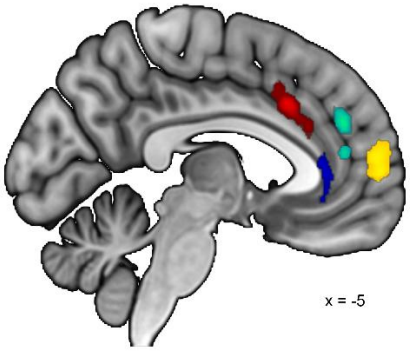


Figure 8. Exploratory overlay of significant mPFC ROI effects. For visualization purposes, significant clusters from all contrasts surviving mask-based cluster inference within the preregistered mPFC mask derived from a meta-analytic clustering approach for social-cognitive processes (Schurz et al., 2021) are overlaid on the MNI152 template for qualitative comparison. The overlays highlight bilateral midcingulate activation for the high > low risk contrast (red), bilateral medial superior frontal activation for low > high risk (yellow), left-lateralized medial superior frontal activation for independent > dependent partner (turquoise), and bilateral anterior cingulate activation for the partner x risk contrast (independent > dependent, low > high; blue). Each colored region reflects voxels significant for that contrast when restricting the search volume to the mPFC mask (cluster-level FWE $p < 0.05$, voxel-wise cluster-defining threshold $p < 0.001$).

Outcome phase (feedback phase)

To examine neural responses to the feedback, we implemented a separate GLM modeling the outcome phase, including the factors group (BPD vs. CON) and outcome valence (positive vs. negative). This model estimated the main effects of outcome and group, as well as the group x outcome interaction. All reported results are whole-brain FWE corrected at the cluster level ($p < 0.05$) using a voxel-wise threshold of $p < 0.001$.

Exploratory analyses of the outcome phase revealed significant effects of outcome valence on BOLD activity. Positive compared to negative outcomes were associated with increased activation in a large cluster encompassing several cortical and subcortical regions. The peak voxel of this cluster was centered in the left caudate nucleus (t =9.420, MNI: -18, 0, 26), extending bilaterally into the dorsal striatum and middle cingulate cortex, and medial and lateral prefrontal regions, consistent with striatal reward processing. Additional significant clusters were found in frontal regions including the left IFG (parts orbitalis; peak voxel: t = 5.058, MNI: -38, 42, -12) and left superior frontal gyrus (peak voxel: t = 4.498, MNI: -22, 32, 54; **Supplementary Table S43**).

The reverse contrast (negative > positive) showed greater activation for regions typically involved in salience and aversive outcome processing. Specifically, in the bilateral insula, with the right insula cluster extending into the triangular part of the right IFG (peak voxel: t = 6.155, MNI: 50, 20, 6) and the left insula cluster extending into the triangular part of the left IFG (peak voxel: t = 5.707, MNI: -40, 20, 2), and the right ACC (peak voxel: t = 5.325, MNI: 6, 32, 28). Additional activation for negative outcomes was observed in the left superior frontal gyrus (medial; peak voxel: t = 5.383, MNI: -8, 50, 16), right middle temporal gyrus (peak voxel: t = 5.640, MNI: 54, 2, -20), left middle temporal gyrus (two clusters: t = 5.344, MNI: -56, -18, -12; t = 5.340, MNI: -56, -36- 2), and the left fusiform gyrus (peak voxel: t = 5.603, MNI: -26, -68, -10; **Supplementary Table S44**).

Importantly, there was no significant main effect of group on outcome-related activity at the corrected whole-brain level. The group x outcome interaction (BPD > CON, positive > negative) yielded one significant cluster in the right superior parietal gyrus (peak voxel: t = 4.148, MNI: 28, -66, 60; **Supplementary Table S45**). No other interaction effects survived correction.

Regions of interest (ROI)

Parameter estimates from the preregistered mPFC and dmPFC ROIs were entered into linear mixed-effects models including group, partner, and risk and their interactions.

mPFC ROI

The analysis revealed no significant main effects of group or risk (both $p > 0.19$), and no significant two-way interactions (all $p > 0.69$). There was a significant main effect of partner ($F(1,189) = 5.02$, $p = 0.026$). However, pairwise comparisons of partner effects within group x risk did not survive FDR correction (all adjusted $p > 0.67$), and no effects involving group or risk were significant. Thus, none of the preregistered hypotheses were supported within the mPFC ROI. Full statistical results are reported in **Supplementary Table S46**.

dmPFC ROI

The analysis revealed no significant main effects of group, partner, or risk (all $p > 0.06$), no significant two-way interactions (all $p > 0.13$), and no significant three-way interaction ($F(1,189) = 3.06$, $p = 0.082$). Accordingly, the dmPFC did not show reliable modulation by partner condition, risk, or group. Full statistical results are provided in **Supplementary Table S47**.

Discussion

The present study investigated how a partner's need for help and the personal risk of helping shape cooperative behavior and its neural correlates in women with BPD and matched controls. In a separate behavioral task, we additionally assessed explicit expectations about a partner's willingness to help, allowing us to compare how social context and risk influence what participants do and what they expect others to do.

Overall, cooperative behavior was primarily driven by whether the partner needed help. Across groups, participants chose the cooperative option (Play) more often when their partner was in need than when the partner was independent. Group differences were more subtle than hypothesized and emerged most clearly under low-risk conditions: across partner contexts, participants with BPD chose the cooperative option (Play) significantly less often than controls when the personal cost was low. Importantly, these group differences were expressed differently for choices and expectations. Whereas choices differed mainly as a function of risk between groups, participants with BPD expected less help than controls specifically when they themselves depended on their partner (across risk conditions). At the neural level, we did not observe the expected mPFC increase when the partner needed help. Instead, medial prefrontal activity, together with lateral frontal and orbitofrontal activity, was more pronounced in the opposite, more mixed-motive context when the partner was independent. Neural group differences emerged selectively under low-risk, independent-partner decisions, where individuals with BPD showed reduced IFG and medial orbitofrontal engagement relative to controls, mirroring the reduced action engagement when personal costs were minimal.

Partner dependence robustly promotes cooperation across groups

Consistent with our first hypothesis (H1.1), participants chose the cooperative option (Play) more frequently when their partner depended on their help than when the partner was independent of their help. This effect was robust across risk levels and groups, underscoring sensitivity to another person's needs. Importantly, choosing Play in the present task is neither a purely cooperative nor a purely self-interested action. In all conditions, Play increased the probability of a positive outcome for both players and thus involved a mixture of self- and other-regarding motives. Partner dependence amplified the helping component of this choice by making the partner's survival contingent on the participant's action,

thereby providing a clear socially prescribed response (to help). By contrast, in the independent-partner conditions, the decision is less normatively constrained and more mixed-motive, requiring participants to balance self- and other-regarding motives. Because the partner manipulation varied only the partner's energy level while holding the participant's own energy level constant, the increased selection of Play in the dependent-partner condition is most parsimoniously interpreted as enhanced other-regarding motivation.

This effect emerged across repeated interactions with the same partner and across different risk levels and groups, demonstrating two key points. First, it underscores that humans are ready to incur personal costs when their actions have direct and salient consequences for another person's payoff even in a stylized game. When responsibility for another's outcome was explicit, participants reliably increased their willingness to help, consistent with extensive evidence that humans are strongly motivated to cooperate and help others even at personal costs (Axelrod & Hamilton, 1981; Fehr & Rockenbach, 2004; Rand et al., 2012; Tomasello, 2009; Warneken & Tomasello, 2006).

Second, these findings indicate that cooperative behavior can be preserved in individuals with BPD even across repeated decisions. Participants with BPD showed intact sensitivity to another person's need when social contingencies were explicit and cooperation took the form of unidirectional helping rather than reciprocal exchange. This pattern aligns with evidence that active cooperation seems to be relatively preserved in individuals with BPD (Doppelhofer et al., 2025; Graumann et al., 2025; Hepp et al., 2014; Lévay et al., 2021; Preuss et al., 2016; Thielmann et al., 2014), whereas impairments are more pronounced in reactive or reciprocity-based contexts (De Panfilis et al., 2019; King-Casas et al., 2008; Thielmann et al., 2014). In addition, these results extend prior findings from anonymous or single-trial economic games (Doppelhofer et al., 2025) and support the notion that baseline cooperative capacity in BPD is largely intact even in repeated interactions under conditions that minimize ambiguity, retaliation, or trust repair demands.

Risk modulates cooperation in a context-dependent manner

Our hypothesis that cooperation would generally be higher under low-risk than high-risk conditions (H1.2) was only partially supported and did not generalize across contexts. Although risk significantly influenced behavior, its effect was contingent on partner dependence. Consistent with our hypothesis (H1.3), when the partner depended on help, higher personal risk (vs. low risk) reduced the likelihood of choosing Play, suggesting a trade-off between concern for the partner and self-protection. In this context, increased personal risk likely reduced the willingness to engage in helping behavior, in line with a small body of prior work in nonclinical samples suggesting that personal risk can attenuate prosocial actions when helping entails potential costs (Bode et al., 2015; Liu et al., 2023; Vieira et al., 2020; Wilbanks et al., 2005). Notably, response times increased in these dependent-partner decisions, particularly under

high-risk conditions, suggesting heightened engagement with choices that carried both greater personal cost and direct consequences for another person's outcome.

In contrast, when the partner was independent, higher risk was associated with more frequent Play choices. In this context, the helping component of Play was weaker and less urgent, and increased selection of Play likely reflected greater willingness to engage in a risky option for potential personal benefit. This pattern aligns with well-established findings that humans are often willing to trade off risk and select uncertain options in pursuit of higher potential gains (Christopoulos et al., 2009; Cohen et al., 2007; Daw et al., 2006; Schultz et al., 2008).

Taken together, these findings underscore that the same action can reflect different motivational processes depending on social context. Rather than exerting a uniform effect, risk modulated how self- and other-regarding considerations were weighted during decision-making. Consistent with prior theoretical and empirical accounts, the relationship between risk and helping appears to be context-sensitive and shaped by factors such as individual risk preferences, social closeness, and situational constraints (Blankenstein & Van Duijvenvoorde, 2019; Christopoulos et al., 2009; Kahneman & Tversky, 1979; Schultz et al., 2011; Telzer et al., 2011; Trepel et al., 2005).

Selective behavioral group differences under low risk

Contrary to our hypothesis that individuals with BPD would show reduced cooperation particularly under high-risk conditions (H1.4), we did not observe a group x partner x risk interaction in choice behavior. Significant group effects emerged selectively along the dimension of personal risk. When averaged across partner dependence, individuals with BPD and controls did not differ in their likelihood of choosing the Play option under high-risk conditions. In contrast, under low-risk conditions, controls chose Play more frequently than participants with BPD.

Because choosing Play reflects a mixture of self- and other-regarding motives whose relative weighting depends on partner dependence, group differences averaged across partner conditions cannot be interpreted as differences in cooperation per se. Instead, the observed group x risk effect suggests that specifically under low-risk conditions, controls were more willing than participants with BPD to engage with a potentially beneficial action across social contexts. Thus, behavioral group differences were expressed as differences in action engagement when personal costs were minimal, rather than as a global reduction in cooperative behavior. Importantly, no group differences were observed in response times (H3), suggesting that these selective behavioral differences are unlikely to be driven by decision speed, impulsivity, or effort, but instead reflect differences in motivational engagement when personal costs are low.

This pattern should be interpreted in the context of a heterogeneous literature on risk-related decision-making in BPD. Although heightened risk-taking is considered a core feature of the disorder (American

Psychiatric Association, 2013; Bandelow et al., 2010; Gunderson et al., 2018), empirical findings across laboratory paradigms are mixed and strongly task-dependent. Some studies report increased selection of risky options or reduced sensitivity to potential losses in BPD (Bazanis et al., 2002; Haaland & Landrø, 2007; Paret et al., 2017; Sánchez-Navarro et al., 2014; Svaldi et al., 2012). Other works find no reliable BPD-related group differences in risk-taking propensity (Coffey et al., 2011), and several studies point to a reduced differentiation between risk levels or attenuated use of risk-related performance feedback (Endrass et al., 2016; Hüpen et al., 2020; Schuermann et al., 2011). Further neurophysiological and neuroimaging evidence suggests that behavioral variability may be linked to attenuated processing of outcome and feedback signals in BPD. Studies have reported reduced neural differentiation during reward and loss anticipation (Enzi et al., 2013; Herbort et al., 2016; Völlm et al., 2007), as well as blunted neural responses to negative feedback, including feedback-related negativity and altered prefrontal and orbitofrontal responses (Andreou et al., 2015; Schuermann et al., 2011; Vega et al., 2013). Together, these findings suggest that decisions in BPD may be less strongly shaped by outcome-related signals and their affective or motivational significance, which could impair adaptive adjustment to changing levels of risk or reward across contexts. Accordingly, the present group difference may reflect context-dependent differences in how potential gains, losses, and their subjective salience are translated into action engagement, rather than a uniform shift toward risk seeking or risk aversion.

Notably, the present task differs from most prior risk-taking paradigms in that choices entail mixed consequences for self and other, making the motivational meaning of risk less purely self-directed. Under such conditions, elevated real-world risk-taking tendencies in BPD may not straightforwardly translate into increased action engagement, especially when affective arousal is low and external incentives are modest. Consistent with this view, the present results suggest that individuals with BPD were able to distinguish risk at the level of explicit expectations (see below) but did not show the same low- versus high-risk shift in Play choices as controls. Rather than reflecting reduced cooperation or exaggerated risk aversion, this pattern points to altered integration of risk information with social and motivational context. Low-risk situations – where controls increased engagement – may therefore represent a critical condition under which differences in adaptive risk modulation become visible in BPD.

Altered expectations despite preserved cooperation in BPD

While behavioral group differences were selectively tied to personal risk, group differences in expectations emerged along the partner context dimension. Consistent with our hypothesis (H2), participants with BPD expected their partner to choose Play less often than controls when they themselves depended on the partner's help, whereas no group difference emerged when the partner was independent. Across groups, expectations were also modulated by risk, with participants in both groups expecting less help under high-risk compared to low-risk conditions.

This pattern of the BPD group expecting less help than the control group aligns with extensive evidence of pessimistic social beliefs, reduced trust, and negative interpersonal priors in BPD (Barnow et al., 2009; C. W. Korn et al., 2016; Lévay et al., 2021; Liebke et al., 2018; Masland et al., 2020; Weinbrecht et al., 2018). Notably, expectations were assessed in a role-reversed version of the cooperation task, strengthening the interpretation that these effects reflect altered beliefs about others' behavior rather than differences in task comprehension, structure, or incentives.

Exploratory analyses of response times in the expectation task showed a parallel pattern to the choice task. Participants took longer to make judgments when they depended on their partner, regardless of group or risk, indicating heightened engagement with socially consequential expectations while again showing no evidence for group differences in decision speed.

Together with the choice data, these findings indicate that action selection and expectation formation in BPD are differentially sensitive to contextual features of social decision-making. Although individuals with BPD showed preserved sensitivity to partner need in their own choices, they nonetheless anticipated only half as much help from others when they were socially dependent compared to controls. This dissociation mirrors findings by Lévay et al. (2021), who reported that individuals with BPD compared to controls expected greater selfishness from others despite behaving prosocially themselves in a modified SVO task.

Importantly, individuals with BPD differentiated between low- and high-risk situations to a similar extent as controls when forming expectations about their partner's behavior, expecting less help under high-risk compared to low-risk conditions. This preserved risk sensitivity indicates that group differences in choices in the cooperation task cannot be attributed to a failure to recognize or represent differences in risk. Rather, individuals with BPD appear able to track changes in risk yet show reduced willingness to act on this information when personal costs are low.

Such discrepancies between beliefs and behavior are consistent with accounts suggesting that active cooperation remains intact in BPD, whereas difficulties emerge more readily in reciprocity-based contexts that depend on interpreting partner behavior and adapting to social feedback over time (Hepp et al., 2014; King-Casas et al., 2008; Thielmann et al., 2014). Altered expectations may have particularly strong consequences in real-world interactions, where intentions are ambiguous, contingencies are uncertain, and social feedback unfolds dynamically over time.

Neural correlates of cooperation across groups

At the neural level, we tested preregistered hypotheses concerning the role of mPFC in socially consequential decision-making. We hypothesized that decisions involving a dependent partner would be associated with increased mPFC engagement across groups (H6.1), reflecting heightened social relevance. While social context robustly modulated neural activity during the choice phase, we did not

observe the expected increase in mPFC activity for dependent > independent partner conditions. Instead, we found increased activity in lateral frontal regions including the bilateral IFG (mainly the triangular parts), together with engagement of the precuneus and middle and superior frontal areas. These regions have been implicated in partially distinct processes, including inhibitory and risk-averse control functions in the (right) IFG (Aron et al., 2004, 2014; Christopoulos et al., 2009; Juan & Muggleton, 2012; Menon et al., 2001) and self-referential processing, episodic memory, and theory of mind in the precuneus (Dadario & Sughrue, 2023). Notably, the dependent > independent partner contrast did not show enhanced recruitment of vmPFC, ventral striatum, or ACC subregions that are frequently reported in studies of cooperative choice, vicarious reward, or joint value computation during prosocial choices or mutual cooperation (Decety et al., 2004; Lockwood et al., 2024; Morelli et al., 2015; Rilling et al., 2002; Ruff & Fehr, 2014; Strombach et al., 2015; Zoh et al., 2022).

Interestingly, medial prefrontal engagement emerged in the opposite direction. Stronger engagement of subregions of the mPFC, specifically bilateral medial OFC and medial superior frontal regions extending into left midcingulate cortex and pregenual ACC, was observed when the partner was independent rather than dependent. Additionally, lateral frontal regions as well as memory and salience related regions, including the right insula and bilateral hippocampus, were recruited in this contrast. This pattern suggests that medial prefrontal recruitment in the present task may not primarily reflect the salience of helping per se but may be more strongly engaged when decisions are less socially prescribed and require greater integration of mixed incentives (e.g., balancing self-regarding rewards with potentially weaker or less urgent other-regarding considerations). This interpretation is consistent with accounts linking mPFC regions to integrative valuation and mentalizing processes during social decision-making (Gangopadhyay et al., 2021; Morelli et al., 2015; Schurz et al., 2014; Tusche & Bas, 2021; Zoh et al., 2022). The OFC has been consistently implicated in representing diverse and abstract reward values across both social and nonsocial contexts (Bray et al., 2010; Cutler & Campbell-Meiklejohn, 2019; Diekhof et al., 2012; Hare et al., 2009; O'Doherty et al., 2001; Peters & Büchel, 2010). Notably, (medial) OFC activity has also been reported during cooperative and prosocial choices, where it is thought to reflect the rewarding aspects of cooperation (Cutler & Campbell-Meiklejohn, 2019; Decety et al., 2004). Its stronger recruitment under independent-partner conditions may therefore reflect increased subjective value integration when self- and other-related incentives must be jointly considered in the absence of clear normative solutions. The independent-partner compared to the dependent-partner condition also recruited middle cingulate regions, which have been discussed in the context of monitoring the outcomes of social decisions and integrating motivational signals related to self- and other-oriented rewards (Apps et al., 2013, 2016). Related work further suggests that interactions between the anterior mid-cingulate cortex and anterior insula (AI) may contribute to guiding social decisions based on combined self- and other-related value signals (Shaw et al., 2019) and to predict reciprocal behavior in trust-based interactions (Cáceda et al., 2017).

These neuroimaging results align with the behavioral framework adopted in the present study. When the partner depended on help, the motivational meaning of the Play option was dominated by its helping component, leaving relatively little ambiguity about the socially appropriate choice. By contrast, when the partner was independent, the same choice required greater integration of self-regarding incentives, other-regarding motives, and probabilistic outcomes. Greater engagement of medial orbitofrontal and cingulate regions under these latter conditions may therefore reflect increased valuation, mentalizing, or context integration when social dependence is low and the decision space is less constrained by social requirements.

Neural alterations in BPD in lateral prefrontal and orbitofrontal regions

We further hypothesized that group differences in mPFC engagement would be most pronounced under conditions combining high personal risk and partner dependence (H6.2). This preregistered prediction was not supported: we did not observe BPD-specific alterations in mPFC activation during high-risk helping conditions. Accordingly, the present data provide no evidence for a robust disruption of mPFC recruitment in BPD in socially consequential decisions, and this finding is consistent with the behavioral results showing intact sensitivity to partner need in individuals with BPD. Notably, this null group effect should be interpreted in the context of the absent across-group vmPFC/mPFC increase in dependent- versus independent-partner conditions, which suggests that medial prefrontal signals were not a dominant signature of helping in this task.

Importantly, however, the absence of the predicted interaction does not imply a complete lack of neural group differences. The opposite contrast revealed greater engagement of lateral prefrontal regions, most notably bilateral IFG (mainly the triangular part), alongside a small cluster in bilateral medial orbital regions, in controls compared to individuals with BPD during low-risk conditions involving an independent partner. In other words, the group effect was expressed in lateral prefrontal engagement and a small cluster in the mOFC which is consistent with the independent > dependent valuation pattern described above. This neural pattern mirrors the behavioral finding that group differences in choices emerged selectively under low-risk conditions, which we interpreted as reflecting differences in willingness to engage in a potentially beneficial action when personal costs are minimal.

Reduced IFG engagement in individuals with BPD during low-risk decisions involving an independent partner is notable in light of recent meta-analytic evidence identifying bilateral inferior frontal regions as potential loci of hypoactivation in BPD in social cognition and emotion processing, albeit less consistently than limbic and medial prefrontal regions (Schurz et al., 2024). Additionally, a broader meta-analysis on the neural underpinnings of BPD reported greater IFG convergence in healthy controls compared to individuals with BPD (Degasperi et al., 2021), suggesting that inferior frontal differences may generalize beyond any single task domain. Although the IFG involvement has not been a prominent

target in many social decision-making studies, it has been consistently implicated in inhibitory control, including response inhibition in motor contexts but also in inhibitory mechanisms in the affective domain (Aron et al., 2014; Berkman et al., 2009; Christopoulos et al., 2009; Frank et al., 2014; Juan & Muggleton, 2012). In addition, IFG engagement has been observed when individuals reappraise the intentions of others (Grecucci et al., 2013; Wager et al., 2008), which is conceptually relevant to socially embedded choices in which motives and consequences must be interpreted. Within the present task, reduced IFG recruitment may therefore tentatively reflect diminished engagement of regulatory or evaluative processes when decisions are weakly constrained by risk or a clear social obligation. However, because this inference is based on an exploratory interaction contrast and the cited IFG functions are not specific to the current paradigm, the interpretation remains speculative and should be treated with caution.

To establish a clean baseline of cooperative decision-making, the task was designed to limit affective arousal, social threat, and interpersonal ambiguity. This matters because fronto-limbic dysregulation in BPD is most reliably observed in paradigms that are emotionally evocative or socially ambiguous, often including heightened amygdala and insula reactivity alongside reduced recruitment of regulatory prefrontal regions (Donegan et al., 2003; Herpertz et al., 2001; Koenigsberg, Fan, et al., 2009; Krause-Utz et al., 2014; Schulze et al., 2016). The absence of group differences in these regions therefore likely reflects task demands rather than preserved function across context. Within this framework, the observed effects under low-risk, independent-partner conditions point to a context-specific alteration in prefrontal engagement when external constraints are minimal, rather than a global disruption of the neural mechanisms supporting cooperation.

Taken together, the fMRI findings indicate that social decision-making in individuals with BPD and controls relies on largely overlapping neural systems during explicitly structured cooperative decisions, with no evidence for prefrontal abnormality in high-risk helping in BPD. Neural group differences were subtle and emerged most clearly under low-risk conditions in which social and motivational constraints were weakest. Under these circumstances, differences were expressed primarily in lateral prefrontal engagement – most notably the IFG – alongside a small medial orbital/superior frontal cluster, rather than in core medial prefrontal or limbic mechanisms typically associated with helping or prosocial behavior. This pattern converges with the behavioral findings and suggests that alterations in BPD may be most evident in context-sensitive engagement of evaluative and regulatory processes when decisions are weakly constrained and more mixed-motive, rather than reflecting a fundamental disruption of the neural networks supporting cooperation itself.

Implications for interpersonal dysfunction in BPD

Together, these findings suggest that interpersonal difficulties in BPD are unlikely to arise from a fundamental inability to cooperate. Instead, they point to altered social expectations and context-dependent reductions in engagement with cooperative actions, particularly when personal costs are low and external constraints are weak. Even across repeated interactions with the same partner, individuals with BPD demonstrated preserved sensitivity to another person's need for help and largely intact recruitment of core social decision-making networks.

These results are consistent with accounts proposing that interpersonal dysfunction in BPD arises not from diminished prosocial motivation, but from negatively biased expectations, maladaptive feedback integration, and problems with forgiveness (Hepp, Kieslich, Schmitz, et al., 2021; C. W. Korn et al., 2016; Liebke et al., 2018; Thielmann et al., 2014). Under conditions that provide clear social structure and explicit responsibility, cooperative behavior and its neural implementation appear largely preserved. By contrast, situations characterized by ambiguity or low constraints – common in everyday social life – may be especially prone to reveal alterations in how individuals with BPD engage regulatory processes, with potential consequences for interpersonal functioning.

Limitations, strengths, and future directions

Several limitations warrant consideration.

First, although the task involved repeated interactions, partner behavior was pre-programmed and optimal, limiting opportunities for learning, reciprocity, or expectation updating. As a result, the task captures sensitivity to social context and risk under controlled conditions but does not fully model dynamic contingencies of real social exchange. Prior work has shown that alterations in cooperation in individuals with BPD, particularly in reactive cooperation, often emerge over time and across different phases of interaction in multi-round paradigms with real partners or trust violations (Abramov et al., 2020; King-Casas et al., 2008). Future studies should therefore extend this paradigm to more naturalistic settings and examine how cooperation and expectations, and their neural correlates evolve across more dynamic, reciprocally contingent, and emotionally salient interactions.

Second, this study did not include additional behavioral or questionnaire-based measures of risk preferences or social risk-taking, which limits interpretation of group differences along the risk dimension. At the same time, prior work suggests that different risk measures often show weak or inconsistent associations (Holzmeister & Stefan, 2021; Pedroni et al., 2017), underscoring the need for future studies to adopt multi-method approaches that integrate behavioral, computational, and physiological measures to better capture individual differences in social decision-making.

Third, expectations were assessed behaviorally but not during fMRI acquisition, precluding direct examination of the neural correlates of expectation formation and their relationship with neural processes related to cooperative choices. Ideally, future studies would combine choice and expectation judgments within the same neuroimaging paradigm to directly test how beliefs about others are represented neurally and how they interact with action selection in social decision-making.

Finally, the generalizability of the findings is limited by sample characteristics. We tested only female participants, consistent with much prior BPD research (for an overview of BPD studies using economic games see Jeung et al., 2016), and although the sample was relatively large, statistical power may still have been insufficient to detect subtle group-by-context interactions. This challenge is compounded by high behavioral variability in BPD populations and the typically small effect sizes observed in low-emotion economic tasks.

Nevertheless, several strengths of the present study should also be noted. We employed a novel task design that dissociates partner dependence and personal risk in a controlled setting and within a single paradigm and allows the same structure to be used for both choices and expectations via role reversal. This design can be readily adapted to real interactive settings and is well suited for the application of computational modeling approaches that differentiate strategic components of social behavior. In addition, the study was conducted in a comparatively large and well-characterized clinical sample of BPD patients, matched with controls on age and intelligence, enhancing the reliability of the observed effects. Moreover, all hypotheses and main analyses were preregistered, strengthening the transparency and robustness of the findings. By deliberately avoiding additional experimental manipulations (e.g., emotion induction or variable partner strategies), we aimed to establish a clear baseline of cooperative behavior under controlled conditions before introducing further complexity in future work.

Conclusion

In conclusion, this study demonstrates that cooperation across repeated interactions is strongly shaped by partner dependence and modulated by personal risk, highlighting the fundamentally context-sensitive nature of social decision-making. Individuals with BPD showed largely preserved cooperative behavior and engagement of medial prefrontal decision-making networks when social contingencies were explicit and responsibility for another person's outcome was clear. At the same time, individuals with BPD showed more pessimistic expectations about others' willingness to help. They also engaged less often in options with higher probabilities of positive outcomes when personal risk was low, indicating reduced engagement with potentially beneficial actions in situations with weak external constraints. At the neural level, group differences emerged most clearly in lateral prefrontal engagement, particularly within the IFG, consist with behavioral differences in action engagement rather than reflecting disruptions of core social value or helping-related networks.

Together, these findings underscore the importance of distinguishing between social expectations, overt cooperative behavior, and their neural implementation. Rather than pointing to a global deficit in cooperative capacity, the present results suggest that interpersonal difficulties in BPD may arise from altered beliefs about others and reduced context-sensitive modulation of behavior – particularly when decisions are weakly constrained by risk or explicit social obligation.

Declarations

Ethics approval and consent to participate

The project was approved by the local ethics committee (Faculty of Medicine, Heidelberg University, S-137/2020). All participants were briefed on the study procedure and gave written informed consent before participating.

Availability of data and materials

We will make data and code available on a public repository upon acceptance.

Funding

This project was supported by the Emmy Noether Research Group grant (392443797) from the German Research Foundation (DFG).

Acknowledgements

We thank our collaborators Sabine C. Herpertz and Katja Bertsch. We also thank Clara Beitz, Clara Heilmann, Inga von Kortzfleisch, Sita Sielaff, and Johanna Löloff for help with data acquisition, Raphael Perla for his guidance on statistical methods, and Marie-Luise Otte for advising on the diagnostic interviews.

References

Abramov, G., Miellet, S., Kautz, J., Grenyer, B. F. S., & Deane, F. P. (2020). The paradoxical decline and growth of trust as a function of borderline personality disorder trait count: Using discontinuous growth modelling to examine trust dynamics in response to violation and repair. *PLOS ONE*, *15*(7), e0236170. https://doi.org/10.1371/journal.pone.0236170

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268–277. https://doi.org/10.1038/nrn1884

Andreou, C., Kleinert, J., Steinmann, S., Fuger, U., Leicht, G., & Mulert, C. (2015). Oscillatory responses to reward processing in borderline personality disorder. *The World Journal of Biological Psychiatry*, *16*(8), 575–586. https://doi.org/10.3109/15622975.2015.1054880

Apps, M. A. J., Lockwood, P. L., & Balsters, J. H. (2013). The role of the midcingulate cortex in monitoring others' decisions. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00251

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron*, *90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*(4), 170–177. https://doi.org/10.1016/j.tics.2004.02.010

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*(4), 177–185. https://doi.org/10.1016/j.tics.2013.12.003

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Bach, B., Kerber, A., Aluja, A., Bastiaens, T., Keeley, J. W., Claes, L., Fossati, A., Gutierrez, F., Oliveira, S. E. S., Pires, R., Riegel, K. D., Rolland, J.-P., Roskam, I., Sellbom, M., Somma, A., Spanemberg, L., Strus, W., Thimm, J. C., Wright, A. G. C., & Zimmermann, J. (2020). International Assessment of DSM-5 and ICD-11 Personality Disorder Traits: Toward a Common Nosology in DSM-5.1. *Psychopathology*, *53*(3–4), 179–188. https://doi.org/10.1159/000507589

Bandelow, B., Schmahl, C., Falkai, P., & Wedekind, D. (2010). Borderline personality disorder: A dysregulation of the endogenous opioid system? *Psychological Review*, *117*(2), 623–636. https://doi.org/10.1037/a0018095

Barnow, S., Stopsack, M., Grabe, H. J., Meinke, C., Spitzer, C., Kronmüller, K., & Sieswerda, S. (2009). Interpersonal evaluation bias in borderline personality disorder. *Behaviour Research and Therapy*, *47*(5), 359–365. https://doi.org/10.1016/j.brat.2009.02.003

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Baudson, T. G., & Preckel, F. (2016). mini-q: Intelligenzscreening in drei Minuten. *Diagnostica*, *62*(3), 182–197. https://doi.org/10.1026/0012-1924/a000150

Bazanis, E., Rogers, R. D., Dowson, J. H., Taylor, P., Meux, C., Staley, C., Nevinson-Andrews, D., Taylor, C., Robbins, T. W., & Sahakian, B. J. (2002). Neurocognitive deficits in decision-making and planning of patients with DSM-III-R borderline personality disorder. *Psychological Medicine*, *32*(8), 1395–1405. https://doi.org/10.1017/S0033291702006657

Beck, A. T., Steer, R. A., & Brown, G. K. (2006). *BDI-II. Beck-Depressionsinventar [Beck Depression Inventory II]*. Pearson Verlag.

Beeney, J. E., Hallquist, M. N., Ellison, W. D., & Levy, K. N. (2016). Self–other disturbance in borderline personality disorder: Neural, self-report, and performance-based evidence. *Personality Disorders: Theory, Research, and Treatment*, *7*(1), 28–39. https://doi.org/10.1037/per0000127

Bellucci, G., Camilleri, J. A., Eickhoff, S. B., & Krueger, F. (2020). Neural signatures of prosocial behaviors. *Neuroscience & Biobehavioral Reviews*, *118*, 186–195. https://doi.org/10.1016/j.neubiorev.2020.07.006

Berkman, E. T., Burklund, L., & Lieberman, M. D. (2009). Inhibitory spillover: Intentional motor inhibition produces incidental limbic inhibition via right inferior frontal cortex. *NeuroImage*, *47*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2009.04.084

Bertsch, K., Krauch, M., Stopfer, K., Haeussler, K., Herpertz, S. C., & Gamer, M. (2017). Interpersonal Threat Sensitivity in Borderline Personality Disorder: An Eye-Tracking Study. *Journal of Personality Disorders*, *31*(5), 647–670. https://doi.org/10.1521/pedi_2017_31_273

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: Social information processing in the human striatum. *WIREs Cognitive Science*, *5*(1), 61–73. https://doi.org/10.1002/wcs.1266

Blankenstein, N. E., & Van Duijvenvoorde, A. C. K. (2019). Neural tracking of subjective value under riskand ambiguity in adolescence. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(6), 1364–1378. https://doi.org/10.3758/s13415-019-00749-5

Bode, N. W. F., Miller, J., O'Gorman, R., & Codling, E. A. (2015). Increased costs reduce reciprocal helping behaviour of humans in a virtual evacuation experiment. *Scientific Reports*, *5*(1), 15896. https://doi.org/10.1038/srep15896

Bohus, M., Kleindienst, N., Limberger, M. F., Stieglitz, R.-D., Domsalla, M., Chapman, A. L., Steil, R., Philipsen, A., & Wolf, M. (2009). The Short Version of the Borderline Symptom List (BSL-23): Development and Initial Data on Psychometric Properties. *Psychopathology*, *42*(1), 32–39. https://doi.org/10.1159/000173701

Bouchard, S., & Sabourin, S. (2009). Borderline personality disorder and couple dysfunctions. *Current Psychiatry Reports*, *11*(1), 55–62. https://doi.org/10.1007/s11920-009-0009-x

Bray, S., Shimojo, S., & O'Doherty, J. P. (2010). Human Medial Orbitofrontal Cortex Is Recruited During Experience of Imagined and Real Rewards. *Journal of Neurophysiology*, *103*(5), 2506–2512. https://doi.org/10.1152/jn.01030.2009

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002, June 2). Presented at the 8th International Conferance on Functional Mapping of the Human Brain. https://www.nitrc.org/projects/marsbar/

Cáceda, R., Prendes-Alvarez, S., Hsu, J.-J., Tripathi, S. P., Kilts, C. D., & James, G. A. (2017). The neural correlates of reciprocity are sensitive to prior experience of reciprocity. *Behavioural Brain Research*, *332*, 136–144. https://doi.org/10.1016/j.bbr.2017.05.030

Cavicchioli, M., & Maffei, C. (2020). Rejection sensitivity in borderline personality disorder and the cognitive–affective personality system: A meta-analytic review. *Personality Disorders: Theory, Research, and Treatment*, *11*(1), 1–12. https://doi.org/10.1037/per0000359

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. *Neuron*, *70*(3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056

Christopoulos, G. I., Tobler, P. N., Bossaerts, P., Dolan, R. J., & Schultz, W. (2009). Neural Correlates of Value, Risk, and Risk Aversion Contributing to Decision Making under Risk. *The Journal of Neuroscience*, *29*(40), 12574–12583. https://doi.org/10.1523/JNEUROSCI.2614-09.2009

Coffey, S. F., Schumacher, J. A., Baschnagel, J. S., Hawk, L. W., & Holloman, G. (2011). Impulsivity and risk-taking in borderline personality disorder with and without substance use disorders. *Personality Disorders: Theory, Research, and Treatment*, *2*(2), 128–141. https://doi.org/10.1037/a0020574

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Cutler, J., & Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *NeuroImage*, *184*, 227–241. https://doi.org/10.1016/j.neuroimage.2018.09.009

Dadario, N. B., & Sughrue, M. E. (2023). The functional role of the precuneus. *Brain*, *146*(9), 3598–3607. https://doi.org/10.1093/brain/awad181

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

De Panfilis, C., Riva, P., Preti, E., Cabrino, C., & Marchesi, C. (2015). When social inclusion is not enough: Implicit expectations of extreme inclusion in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *6*(4), 301–309. https://doi.org/10.1037/per0000132

De Panfilis, C., Schito, G., Generali, I., Gozzi, L. A., Ossola, P., Marchesi, C., & Grecucci, A. (2019). Emotions at the border: Increased punishment behavior during fair interpersonal exchanges in borderline personality disorder. *Journal of Abnormal Psychology*, *128*(2), 162–172. https://doi.org/10.1037/abn0000404

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: An fMRI investigation. *NeuroImage*, *23*(2), 744–751. https://doi.org/10.1016/j.neuroimage.2004.05.025

Degasperi, G., Cristea, I. A., Di Rosa, E., Costa, C., & Gentili, C. (2021). Parsing variability in borderline personality disorder: A meta-analysis of neuroimaging studies. *Translational Psychiatry*, *11*(1), 314. https://doi.org/10.1038/s41398-021-01446-z

Diekhof, E. K., Kaps, L., Falkai, P., & Gruber, O. (2012). The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude – An activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and

outcome processing. *Neuropsychologia*, *50*(7), 1252–1266. https://doi.org/10.1016/j.neuropsychologia.2012.02.007

Domes, G., Czieschnek, D., Weidler, F., Berger, C., Fast, K., & Herpertz, S. C. (2008). Recognition of Facial Affect in Borderline Personality Disorder. *Journal of Personality Disorders*, *22*(2), 135–147. https://doi.org/10.1521/pedi.2008.22.2.135

Domsalla, M., Koppe, G., Niedtfeld, I., Vollstädt-Klein, S., Schmahl, C., Bohus, M., & Lis, S. (2014). Cerebral processing of social rejection in patients with borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *9*(11), 1789–1797. https://doi.org/10.1093/scan/nst176

Donegan, N. H., Sanislow, C. A., Blumberg, H. P., Fulbright, R. K., Lacadie, C., Skudlarski, P., Gore, J. C., Olson, I. R., McGlashan, T. H., & Wexler, B. E. (2003). Amygdala hyperreactivity in borderline personality disorder: Implications for emotional dysregulation. *Biological Psychiatry*, *54*(11), 1284–1293. https://doi.org/10.1016/S0006-3223(03)00636-X

Doppelhofer, L. M., Löloff, J., Neukel, C., Herpertz, S. C., & Korn, C. W. (2025). Cooperative decision-making in borderline personality disorder: Insights from a preregistered study using a comprehensive economic task battery. *Borderline Personality Disorder and Emotion Dysregulation*, *12*(1), 24. https://doi.org/10.1186/s40479-025-00295-2

Endrass, T., Schuermann, B., Roepke, S., Kessler-Scheil, S., & Kathmann, N. (2016). Reduced risk avoidance and altered neural correlates of feedback processing in patients with borderline personality disorder. *Psychiatry Research*, *243*, 14–22. https://doi.org/10.1016/j.psychres.2016.06.016

Enzi, B., Doering, S., Faber, C., Hinrichs, J., Bahmer, J., & Northoff, G. (2013). Reduced deactivation in reward circuitry and midline structures during emotion processing in borderline personality disorder. *The World Journal of Biological Psychiatry*, *14*(1), 45–56. https://doi.org/10.3109/15622975.2011.579162

Fehr, E., & Rockenbach, B. (2004). Human altruism: Economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology*, *14*(6), 784–790. https://doi.org/10.1016/j.conb.2004.10.007

Fenske, S., Lis, S., Liebke, L., Niedtfeld, I., Kirsch, P., & Mier, D. (2015). Emotion recognition in borderline personality disorder: Effects of emotional information on negative bias. *Borderline Personality Disorder and Emotion Dysregulation*, *2*(1), 10. https://doi.org/10.1186/s40479-015-0031-z

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage. https://www.john-fox.ca/Companion/

Frank, D. W., Dewitt, M., Hudgens-Haney, M., Schaeffer, D. J., Ball, B. H., Schwarz, N. F., Hussein, A. A., Smart, L. M., & Sabatinelli, D. (2014). Emotion regulation: Quantitative meta-analysis of functional activation and deactivation. *Neuroscience & Biobehavioral Reviews*, *45*, 202–211. https://doi.org/10.1016/j.neubiorev.2014.06.010

Franke, G. H. (2017). *BSCL-53®-S. Brief symptom-checklist–standard–German manual*. Hogrefe Publishing Group.

Frith, C. D., & Frith, U. (2012). *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Frith, C. D., & Singer, T. (2008). The role of social cognition in decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3875–3886. https://doi.org/10.1098/rstb.2008.0156

Gangopadhyay, P., Chawla, M., Dal Monte, O., & Chang, S. W. C. (2021). Prefrontal–amygdala circuits in social decision-making. *Nature Neuroscience*, *24*(1), 5–18. https://doi.org/10.1038/s41593-020-00738-9

Gläscher, J. P., & Gitelmann, D. (2008). *(PDF) Contrast weights in flexible factorial design with multiple groups of subjects*. ResearchGate. https://www.researchgate.net/publication/267779738_Contrast_weights_in_flexible_factorial_design_with_multiple_groups_of_subjects

Graumann, L., Kulakova, E., Cho, A. B., Deuter, C. E., Wolf, O. T., Schell, J., Hellmann-Regen, J., Roepke, S., Otte, C., & Wingenfeld, K. (2025). Elevated testosterone and prosocial behavior in female patients with borderline personality disorder independent of social exclusion. *Psychoneuroendocrinology*, *171*, 107232. https://doi.org/10.1016/j.psyneuen.2024.107232

Grecucci, A., Giorgetta, C., Bonini, N., & Sanfey, A. G. (2013). Reappraising social emotions: The role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00523

Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nature Reviews Disease Primers*, *4*(1), 18029. https://doi.org/10.1038/nrdp.2018.29

Haaland, V. Ø., & Landrø, N. I. (2007). Decision making as measured with the Iowa Gambling Task in patients with borderline personality disorder. *Journal of the International Neuropsychological Society*, *13*(4), 699–703. https://doi.org/10.1017/S1355617707070890

Haber, S. N., & Knutson, B. (2010). The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology*, *35*(1), 4–26. https://doi.org/10.1038/npp.2009.129

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science*, *324*(5927), 646–648. https://doi.org/10.1126/science.1168450

Hepp, J., Hilbig, B. E., Moshagen, M., Zettler, I., Schmahl, C., & Niedtfeld, I. (2014). Active versus reactive cooperativeness in Borderline Psychopathology: A dissection based on the HEXACO model of personality. *Personality and Individual Differences*, *56*, 19–23. https://doi.org/10.1016/j.paid.2013.08.013

Hepp, J., Kieslich, P. J., Schmitz, M., Schmahl, C., & Niedtfeld, I. (2021). Negativity on two sides: Individuals with borderline personality disorder form negative first impressions of others and are perceived negatively by them. *Personality Disorders: Theory, Research, and Treatment*, *12*(6), 514–525. https://doi.org/10.1037/per0000412

Hepp, J., Kieslich, P. J., Wycoff, A. M., Bertsch, K., Schmahl, C., & Niedtfeld, I. (2021). Mouse-tracking reveals cognitive conflict during negative impression formation in women with Borderline Personality Disorder or Social Anxiety Disorder. *PLOS ONE*, *16*(3), e0247955. https://doi.org/10.1371/journal.pone.0247955

Herbort, M. C., Soch, J., Wüstenberg, T., Krauel, K., Pujara, M., Koenigs, M., Gallinat, J., Walter, H., Roepke, S., & Schott, B. H. (2016). A negative relationship between ventral striatal loss anticipation response and impulsivity in borderline personality disorder. *NeuroImage: Clinical*, *12*, 724–736. https://doi.org/10.1016/j.nicl.2016.08.011

Herpertz, S. C., Dietrich, T. M., Wenning, B., Krings, T., Erberich, S. G., Willmes, K., Thron, A., & Sass, H. (2001). Evidence of abnormal amygdala functioning in borderline personality disorder: A functional MRI study. *Biological Psychiatry*, *50*(4), 292–298. https://doi.org/10.1016/S0006-3223(01)01075-7

Hiser, J., & Koenigs, M. (2018). The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. *Biological Psychiatry*, *83*(8), 638–647. https://doi.org/10.1016/j.biopsych.2017.10.030

Holzmeister, F., & Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in)consistency? *Experimental Economics*, *24*(2), 593–616. https://doi.org/10.1007/s10683-020-09674-8

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLOS Computational Biology*, *14*(2), e1005935. https://doi.org/10.1371/journal.pcbi.1005935

Hüpen, P., Wagels, L., Weidler, C., Kable, J. W., Schneider, F., & Habel, U. (2020). Altered psychophysiological correlates of risk-taking in borderline personality disorder. *Psychophysiology*, *57*(5), e13540. https://doi.org/10.1111/psyp.13540

Izurieta Hidalgo, N. A., Oelkers-Ax, R., Nagy, K., Mancke, F., Bohus, M., Herpertz, S. C., & Bertsch, K. (2016). Time course of facial emotion processing in women with borderline personality disorder: An ERP study. *Journal of Psychiatry and Neuroscience*, *41*(1), 16–26. https://doi.org/10.1503/jpn.140215

Jeung, H., Schwieren, C., & Herpertz, S. C. (2016). Rationality and self-interest as economic-exchange strategy in borderline personality disorder: Game theory, social preferences, and interpersonal behavior. *Neuroscience & Biobehavioral Reviews*, *71*, 849–864. https://doi.org/10.1016/j.neubiorev.2016.10.030

Juan, C.-H., & Muggleton, N. G. (2012). Brain stimulation and inhibitory control. *Brain Stimulation*, *5*(2), 63–69. https://doi.org/10.1016/j.brs.2012.03.012

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263. https://doi.org/10.2307/1914185

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The Rupture and Repair of Cooperation in Borderline Personality Disorder. *Science*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

Klein-Flügge, M. C., Bongioanni, A., & Rushworth, M. F. S. (2022). Medial and orbital frontal cortex in decision-making and flexible behavior. *Neuron*, *110*(17), 2743–2770. https://doi.org/10.1016/j.neuron.2022.05.022

Koenigsberg, H. W., Fan, J., Ochsner, K. N., Liu, X., Guise, K. G., Pizzarello, S., Dorantes, C., Guerreri, S., Tecuta, L., Goodman, M., New, A., & Siever, L. J. (2009). Neural Correlates of the Use of Psychological Distancing to Regulate Responses to Negative Social Cues: A Study of Patients with Borderline Personality Disorder. *Biological Psychiatry*, *66*(9), 854–863. https://doi.org/10.1016/j.biopsych.2009.06.010

Koenigsberg, H. W., Siever, L. J., Lee, H., Pizzarello, S., New, A. S., Goodman, M., Cheng, H., Flory, J., & Prohovnik, I. (2009). Neural correlates of emotion processing in borderline personality disorder. *Psychiatry Research: Neuroimaging*, *172*(3), 192–199. https://doi.org/10.1016/j.pscychresns.2008.07.010

Korn, C. W., & Bach, D. R. (2018). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications*, *9*(1), 325. https://doi.org/10.1038/s41467-017-02750-3

Korn, C. W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine*, *46*(3), 575–587. https://doi.org/10.1017/S003329171500207X

Korn, Christopher W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine*, *46*(3), 575–587.

Krause-Utz, A., Winter, D., Niedtfeld, I., & Schmahl, C. (2014). The Latest Neuroimaging Findings in Borderline Personality Disorder. *Current Psychiatry Reports*, *16*(3), 438. https://doi.org/10.1007/s11920-014-0438-z

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lazarus, S. A., Beeney, J. E., Howard, K. P., Strunk, D. R., Pilkonis, Paul. A., & Cheavens, J. S. (2020). Characterization of relationship instability in women with borderline personality disorder: A social network analysis. *Personality Disorders: Theory, Research, and Treatment*, *11*(5), 312–320. https://doi.org/10.1037/per0000380

Lee, V. K., & Harris, L. T. (2013). How social cognition can inform social decision making. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00259

Lenth, R. V. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Lévay, E. E., Bajzát, B., & Unoka, Z. S. (2021). Expectation of Selfishness From Others in Borderline Personality Disorder. *Frontiers in Psychology*, *12*, 702227. https://doi.org/10.3389/fpsyg.2021.702227

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Liebke, L., Bungert, M., Thome, J., Hauschild, S., Gescher, D. M., Schmahl, C., Bohus, M., & Lis, S. (2017). Loneliness, social networks, and social functioning in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *8*(4), 349–356. https://doi.org/10.1037/per0000208

Liebke, L., Koppe, G., Bungert, M., Thome, J., Hauschild, S., Defiebre, N., Izurieta Hidalgo, N. A., Schmahl, C., Bohus, M., & Lis, S. (2018). Difficulties with being socially accepted: An experimental study in borderline personality disorder. *Journal of Abnormal Psychology*, *127*(7), 670–682. https://doi.org/10.1037/abn0000373

Lischke, A., Herpertz, S. C., Berger, C., Domes, G., & Gamer, M. (2017). Divergent effects of oxytocin on (para-)limbic reactivity to emotional and neutral scenes in females with and without borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *12*(11), 1783–1792. https://doi.org/10.1093/scan/nsx107

Liu, C., Xiao, X., Pi, Q., Tan, Q., & Zhan, Y. (2023). Are you more risk-seeking when helping others? Effects of situational urgency and peer presence on prosocial risky behavior. *Frontiers in Psychology*, *14*, 1036624. https://doi.org/10.3389/fpsyg.2023.1036624

Lockwood, P. L., Cutler, J., Drew, D., Abdurahman, A., Jeyaretna, D. S., Apps, M. A. J., Husain, M., & Manohar, S. G. (2024). Human ventromedial prefrontal cortex is necessary for prosocial motivation. *Nature Human Behaviour*, *8*(7), 1403–1416. https://doi.org/10.1038/s41562-024-01899-4

Lockwood, P. L., & Wittmann, M. K. (2018). Ventral anterior cingulate cortex and social decision-making. *Neuroscience & Biobehavioral Reviews*, *92*, 187–191. https://doi.org/10.1016/j.neubiorev.2018.05.030

Loranger, A. W., Sartorius, N., & Andreoli, A. (1994). The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration International Pilot Study of Personality Disorders. *Archives of General Psychiatry*, *51*(3), 215. https://doi.org/10.1001/archpsyc.1994.03950030051005

Loranger, A. W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., Coid, B., Dahl, A. A., Diekstra, R. F. W., & Ferguson, B. (1998). Deutschsprachige Fassung der International personality disorder examination: IPDE. *Genf: WHO*.

Margraf, J., & Cwik, J. C. (2017). *Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen* [PDF]. https://doi.org/10.13154/RUB.102.91

Masland, S. R., & Hooley, J. M. (2020). When Trust Does Not Come Easily: Negative Emotional Information Unduly Influences Trustworthiness Appraisals for Individuals With Borderline Personality Features. *Journal of Personality Disorders*, *34*(3), 394–409. https://doi.org/10.1521/pedi_2019_33_404

Masland, S. R., Schnell, S. E., & Shah, T. V. (2020). Trust Beliefs, Biases, and Behaviors in Borderline Personality Disorder: Empirical Findings and Relevance to Epistemic Trust. *Current Behavioral Neuroscience Reports*, *7*(4), 239–249. https://doi.org/10.1007/s40473-020-00220-7

Mauchnik, J., & Schmahl, C. (2010). The Latest Neuroimaging Findings in Borderline Personality Disorder. *Current Psychiatry Reports*, *12*(1), 46–55. https://doi.org/10.1007/s11920-009-0089-7

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835. https://doi.org/10.1073/pnas.211415698

Menon, V., Adleman, N. E., White, C. D., Glover, G. H., & Reiss, A. L. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, *12*(3), 131–143. https://doi.org/10.1002/1097-0193(200103)12:3%253C131::AID-HBM1010%253E3.0.CO;2-C

Mier, D., Lis, S., Esslinger, C., Sauer, C., Hagenhoff, M., Ulferts, J., Gallhofer, B., & Kirsch, P. (2013). Neuronal correlates of social cognition in borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *8*(5), 531–537. https://doi.org/10.1093/scan/nss028

Minzenberg, M. J., Fan, J., New, A. S., Tang, C. Y., & Siever, L. J. (2007). Fronto-limbic dysfunction in response to facial emotion in borderline personality disorder: An event-related fMRI study. *Psychiatry Research: Neuroimaging*, *155*(3), 231–243. https://doi.org/10.1016/j.pscychresns.2007.03.006

Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage*, *112*, 244–253. https://doi.org/10.1016/j.neuroimage.2014.12.056

Niedtfeld, I., & Kroneisen, M. (2020). Impaired memory for cooperative interaction partners in borderline personality disorder. *Borderline Personality Disorder and Emotion Dysregulation*, *7*(1), 22. https://doi.org/10.1186/s40479-020-00137-3

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, *4*(1), 95–102. https://doi.org/10.1038/82959

Ongur, D. (2000). The Organization of Networks within the Orbital and Medial Prefrontal Cortex of Rats, Monkeys and Humans. *Cerebral Cortex*, *10*(3), 206–219. https://doi.org/10.1093/cercor/10.3.206

Paret, C., Jennen-Steinmetz, C., & Schmahl, C. (2017). Disadvantageous decision-making in borderline personality disorder: Partial support from a meta-analytic review. *Neuroscience & Biobehavioral Reviews*, *72*, 301–309. https://doi.org/10.1016/j.neubiorev.2016.11.019

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*(11), 803–809. https://doi.org/10.1038/s41562-017-0219-x

Peters, J., & Büchel, C. (2010). Neural representations of subjective reward value. *Behavioural Brain Research*, *213*(2), 135–141. https://doi.org/10.1016/j.bbr.2010.04.031

Poggi, A., Richetin, J., & Preti, E. (2019). Trust and Rejection Sensitivity in Personality Disorders. *Current Psychiatry Reports*, *21*(8), 69. https://doi.org/10.1007/s11920-019-1059-3

Preuss, N., Brändle, L. S., Hager, O. M., Haynes, M., Fischbacher, U., & Hasler, G. (2016). Inconsistency and social decision making in patients with Borderline Personality Disorder. *Psychiatry Research*, *243*, 115–122. https://doi.org/10.1016/j.psychres.2016.06.017

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430. https://doi.org/10.1038/nature11467

Renneberg, B., Herm, K., Hahn, A., Staebler, K., Lammers, C., & Roepke, S. (2012). Perception of Social Participation in Borderline Personality Disorder. *Clinical Psychology & Psychotherapy*, *19*(6), 473–480. https://doi.org/10.1002/cpp.772

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron*, *35*(2), 395–405. https://doi.org/10.1016/S0896-6273(02)00755-9

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, *206*, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562. https://doi.org/10.1038/nrn3776

Ruocco, A. C., Medaglia, J. D., Tinker, J. R., Ayaz, H., Forman, E. M., Newman, C. F., Williams, J. M., Hillary, F. G., Platek, S. M., Onaral, B., & Chute, D. L. (2010). Medial prefrontal cortex hyperactivation during social exclusion in borderline personality disorder. *Psychiatry Research: Neuroimaging*, *181*(3), 233–236. https://doi.org/10.1016/j.pscychresns.2009.12.001

Sánchez-Navarro, J. P., Weller, J. A., López-Navarro, J. M., Martínez-Selva, J. M., & Bechara, A. (2014). Suboptimal Decision Making in Borderline Personality Disorder: Effect of Potential Losses. *The Spanish Journal of Psychology*, *17*, E38. https://doi.org/10.1017/sjp.2014.39

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science*, *318*(5850), 598–602. https://doi.org/10.1126/science.1142996

Saunders, K. E. A., Goodwin, G. M., & Rogers, R. D. (2015). Borderline personality disorder, but not euthymic bipolar disorder, is associated with a failure to sustain reciprocal cooperative behaviour: Implications for spectrum models of mood disorders. *Psychological Medicine*, *45*(8), 1591–1600. https://doi.org/10.1017/S0033291714002475

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Scheunemann, J., Jelinek, L., Biedermann, S. V., Lipp, M., Yassari, A. H., Kühn, S., Gallinat, J., & Moritz, S. (2023). Can you trust this source? Advice taking in borderline personality disorder. *European Archives of Psychiatry and Clinical Neuroscience*, *273*(4), 875–885. https://doi.org/10.1007/s00406-022-01539-w

Schuermann, B., Kathmann, N., Stiglmayr, C., Renneberg, B., & Endrass, T. (2011). Impaired decision making and feedback evaluation in borderline personality disorder. *Psychological Medicine*, *41*(9), 1917–1927. https://doi.org/10.1017/S003329171000262X

Schultz, W., O'Neill, M., Tobler, P. N., & Kobayashi, S. (2011). Neuronal signals for reward risk in frontal cortex. *Annals of the New York Academy of Sciences*, *1239*(1), 109–117. https://doi.org/10.1111/j.1749-6632.2011.06256.x

Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C. D., Tobler, P. N., & Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3801–3811. https://doi.org/10.1098/rstb.2008.0152

Schulze, L., Schmahl, C., & Niedtfeld, I. (2016). Neural Correlates of Disturbed Emotion Processing in Borderline Personality Disorder: A Multimodal Meta-Analysis. *Biological Psychiatry*, *79*(2), 97–106. https://doi.org/10.1016/j.biopsych.2015.03.027

Schulze, L., Schulze, A., Renneberg, B., Schmahl, C., & Niedtfeld, I. (2019). Neural Correlates of Affective Disturbances: A Comparative Meta-analysis of Negative Affect Processing in Borderline Personality Disorder, Major Depressive Disorder, and Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(3), 220–232. https://doi.org/10.1016/j.bpsc.2018.11.004

Schurz, M., Berenz, J.-P., Maerz, J., Perla, R., Buchheim, A., & Labek, K. (2024). Brain Activation for Social Cognition and Emotion Processing Tasks in Borderline Personality Disorder: A Meta-Analysis of Neuroimaging Studies. *Brain Sciences*, *14*(4), 395. https://doi.org/10.3390/brainsci14040395

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303

Shaw, D., Czekóová, K., Gajdoš, M., Staněk, R., Špalek, J., & Brázdil, M. (2019). Social decision-making in the brain: Input-state-output modelling reveals patterns of effective connectivity underlying reciprocal choices. *Human Brain Mapping*, *40*(2), 699–712. https://doi.org/10.1002/hbm.24446

Spitzer, C., Müller, S., Kerber, A., Hutsebaut, J., Brähler, E., & Zimmermann, J. (2021). Die deutsche Version der Level of Personality Functioning Scale-Brief Form 2.0 (LPFS-BF): Faktorenstruktur, konvergente Validität und Normwerte in der Allgemeinbevölkerung. *PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, *71*(07), 284–293. https://doi.org/10.1055/a-1343-2396

Spunt, B. (2016). *spunt/bspmview: BSPMVIEW v.20161108* (Version 20161108) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.168074

Staebler, K., Helbing, E., Rosenbach, C., & Renneberg, B. (2011). Rejection sensitivity and borderline personality disorder. *Clinical Psychology & Psychotherapy*, *18*(4), 275–283. https://doi.org/10.1002/cpp.705

Staebler, K., Renneberg, B., Stopsack, M., Fiedler, P., Weiler, M., & Roepke, S. (2011). Facial emotional expression in reaction to social exclusion in borderline personality disorder. *Psychological Medicine*, *41*(9), 1929–1938. https://doi.org/10.1017/S0033291711000080

Stepp, S. D., Pilkonis, P. A., Yaggi, K. E., Morse, J. Q., & Feske, U. (2009). Interpersonal and Emotional Experiences of Social Interactions in Borderline Personality Disorder. *Journal of Nervous & Mental Disease*, *197*(7), 484–491. https://doi.org/10.1097/NMD.0b013e3181aad2e7

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, *112*(5), 1619–1624. https://doi.org/10.1073/pnas.1414715112

Svaldi, J., Philipsen, A., & Matthies, S. (2012). Risky decision-making in borderline personality disorder. *Psychiatry Research*, *197*(1–2), 112–118. https://doi.org/10.1016/j.psychres.2012.01.014

Tebartz Van Elst, L., Hesslinger, B., Thiel, T., Geiger, E., Haegele, K., Lemieux, L., Lieb, K., Bohus, M., Hennig, J., & Ebert, D. (2003). Frontolimbic brain abnormalities in patients with borderline personality disorder. *Biological Psychiatry*, *54*(2), 163–171. https://doi.org/10.1016/S0006-3223(02)01743-2

Telzer, E. H., Masten, C. L., Berkman, E. T., Lieberman, M. D., & Fuligni, A. J. (2011). Neural regions associated with self control and mentalizing are recruited during prosocial behaviors towards the family. *NeuroImage*, *58*(1), 242–249. https://doi.org/10.1016/j.neuroimage.2011.06.013

The MathWorks Inc. (2024). *MATLAB version: 24.1.0. (R2024a)* [Computer software]. The MathWorks Inc. https://www.mathworks.com

Thielmann, I., Hilbig, B. E., & Niedtfeld, I. (2014). Willing to Give but Not to Forgive: Borderline Personality Features and Cooperative Behavior. *Journal of Personality Disorders*, *28*(6), 778–795. https://doi.org/10.1521/pedi_2014_28_135

Thome, J., Liebke, L., Bungert, M., Schmahl, C., Domes, G., Bohus, M., & Lis, S. (2016). Confidence in facial emotion recognition in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *7*(2), 159–168. https://doi.org/10.1037/per0000142

Tomasello, M. (2009). *Why We Cooperate*. The MIT Press. https://doi.org/10.7551/mitpress/8470.001.0001

Trepel, C., Fox, C. R., & Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive Brain Research*, *23*(1), 34–50. https://doi.org/10.1016/j.cogbrainres.2005.01.016

Tusche, A., & Bas, L. M. (2021). Neurocomputational models of altruistic decision-making and social motives: Advances, pitfalls, and future directions. *WIREs Cognitive Science*, *12*(6), e1571. https://doi.org/10.1002/wcs.1571

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*(4), 399–409. https://doi.org/10.1521/pedi.2009.23.4.399

Vega, D., Soto, À., Amengual, J. L., Ribas, J., Torrubia, R., Rodríguez-Fornells, A., & Marco-Pallarés, J. (2013). Negative reward expectations in Borderline Personality Disorder patients: Neurophysiological evidence. *Biological Psychology*, *94*(2), 388–396. https://doi.org/10.1016/j.biopsycho.2013.08.002

Vieira, J. B., Schellhaas, S., Enström, E., & Olsson, A. (2020). Help or flight? Increased threat imminence promotes defensive helping in humans. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1933), 20201473. https://doi.org/10.1098/rspb.2020.1473

Völlm, B., Richardson, P., McKie, S., Elliott, R., Dolan, M., & Deakin, B. (2007). Neuronal correlates of reward and loss in Cluster B personality disorders: A functional magnetic resonance imaging study. *Psychiatry Research: Neuroimaging*, *156*(2), 151–167. https://doi.org/10.1016/j.pscychresns.2007.04.008

Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A., & Ochsner, K. N. (2008). Prefrontal-Subcortical Pathways Mediating Successful Emotion Regulation. *Neuron*, *59*(6), 1037–1050. https://doi.org/10.1016/j.neuron.2008.09.006

Warneken, F., & Tomasello, M. (2006). Altruistic Helping in Human Infants and Young Chimpanzees. *Science*, *311*(5765), 1301–1303. https://doi.org/10.1126/science.1121448

Weekers, L. C., Hutsebaut, J., & Kamphuis, J. H. (2019). The Level of Personality Functioning Scale-Brief Form 2.0: Update of a brief instrument for assessing level of personality functioning: The Level of Personality Functioning Scale - Brief Form 2.0. *Personality and Mental Health*, *13*(1), 3–14. https://doi.org/10.1002/pmh.1434

Weinbrecht, A., Niedeggen, M., Roepke, S., & Renneberg, B. (2018). Feeling excluded no matter what? Bias in the processing of social participation in borderline personality disorder. *NeuroImage: Clinical*, *19*, 343–350. https://doi.org/10.1016/j.nicl.2018.04.031

Wilbanks, B. A., Matheson, M. D., Eubanks, J. L., & DeVietti, T. L. (2005). Effects of Perceived Risk and Strength of Affiliation on Decisions to Help Others: Toward a Convergence of Evolutionary and Social Psychology. *Psychological Reports*, *96*(3_suppl), 977–987. https://doi.org/10.2466/pr0.96.3c.977-987

Wittmann, M. K., Lockwood, P. L., & Rushworth, M. F. S. (2018). Neural Mechanisms of Social Cognition in Primates. *Annual Review of Neuroscience*, *41*(1), 99–118. https://doi.org/10.1146/annurev-neuro-080317-061450

Zimmermann, J., Altenstein, D., Krieger, T., Holtforth, M. G., Pretsch, J., Alexopoulos, J., Spitzer, C., Benecke, C., Krueger, R. F., Markon, K. E., & Leising, D. (2014). The Structure and Correlates of Self-Reported DSM-5 Maladaptive Personality Traits: Findings From Two German-Speaking Samples. *Journal of Personality Disorders*, *28*(4), 518–540. https://doi.org/10.1521/pedi_2014_28_130

Zoh, Y., Chang, S. W. C., & Crockett, M. J. (2022). The prefrontal cortex and (uniquely) human cooperation: A comparative perspective. *Neuropsychopharmacology*, *47*(1), 119–133. https://doi.org/10.1038/s41386-021-01092-5

5 Study III: Fairness preferences and allocation of payoffs

Cooperative decision-making in borderline personality disorder: insights from a preregistered study using a comprehensive economic task battery. *Borderline Personality Disorder and Emotion Dysregulation*, *12*(1), 24 (2025)

Lisa M. Doppelhofer, Johanna Löloff, Corinne Neukel, Sabine C. Herpertz, & Christoph W. Korn

Doppelhofer *et al.*
Borderline Personality Disorder and Emotion Dysregulation (2025) 12:24
https://doi.org/10.1186/s40479-025-00295-2

Borderline Personality Disorder
and Emotion Dysregulation

RESEARCH Open Access

Cooperative decision-making in borderline personality disorder: insights from a preregistered study using a comprehensive economic task battery

L. M. Doppelhofer[1,2*], J. Löloff[1], C. Neukel[1,2], S. C. Herpertz[1,2] and C. W. Korn[1,2]

Abstract

Background Social decision-making poses challenges for individuals with Borderline Personality Disorder (BPD), which can lead to maladaptive interpersonal functioning. Although previous research on cooperative decisions in BPD has produced mixed findings, studies examining different aspects of cooperative behavior suggest that individuals with BPD may specifically show reduced reactive cooperation, while their active cooperation appears similar to that of healthy individuals.

Methods To address these mixed results, we used a comprehensive battery of economic tasks in this preregistered study to assess different aspects of cooperation in 35 women with BPD and 50 healthy women.

Results Consistent with our expectations, there were no significant differences between groups in active cooperation within the Dictator Game. However, contrary to our hypotheses, neither BPD status nor dissociality scores were related to reactive cooperation in the Ultimatum Game. To synthesize findings across studies, a mini meta-analysis was conducted, revealing no significant overall effect of group membership on rejection rates in the Ultimatum Game. Additionally, women with BPD showed similar social preferences, fairness perceptions and inequality aversion as the control group.

Conclusion These results suggest that in the context of anonymous, one-shot economic games, individuals with BPD may not differ from healthy individuals in terms of cooperative behavior. This highlights the importance of using baseline variants of standard economic tasks when considering contextual factors that affect social behavior in BPD.

Keywords Borderline Personality Disorder, Economic Games, Cooperation, Fairness

*Correspondence:
L. M. Doppelhofer
lisamarie.doppelhofer@med.uni-heidelberg.de
[1]Department of General Psychiatry, Center for Psychosocial Medicine, Heidelberg University, Heidelberg, Germany
[2]German Center for Mental Health (DZPG), partner site Mannheim/Heidelberg/Ulm, Germany

© The Author(s) 2025. **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/.

Introduction

Individuals often face situations where their decisions have ripple effects on others. The ability to make appropriate social decisions and cooperate is essential for successful social interactions and relationships. Mental disorders like Borderline Personality Disorder (BPD) can alter how individuals perceive social cues and react in social situations. BPD is characterized by symptoms including intense negative affect, impulsive behavior, unstable relationships, and an unstable self-image [1]. These features are linked to deficits in various processes that are critical for social functioning.

Individuals diagnosed with BPD hold a more inconsistent and negative view of themselves [2–4] and extend this negativity to the social world around them, which they perceive as more negative and malevolent compared to those without the disorder [5, 6]. They also tend to evaluate other people as more negative [3], aggressive [7], less trustworthy [8, 9], and anticipate greater selfishness in others' behavior [10] compared to healthy individuals. Relatedly, individuals with BPD have been shown to exhibit higher rejection sensitivity [11–14] and feel reduced connectedness with others even when being included [15–18]. These challenges impact both the quantity and quality of social relationships. Individuals with BPD tend to maintain fewer social connections, and their relationships are often characterized by greater instability and perceived ambivalence and negativity compared to those of healthy and clinical controls [19–21].

To investigate complex social behavior, economic games provide a useful tool, offering controlled settings to assess cooperation and decision-making. While different economic games have been widely used in BPD research (for reviews see [22, 23]), the findings remain inconsistent.

For example, findings from economic game studies suggest that individuals with BPD struggle to sustain and repair broken cooperation compared to controls, particularly in Trust Game (TG) paradigms [24–26]. Similar patterns have been observed in other paradigms, such as the Prisoner's Dilemma [27] and the Ultimatum Game (UG) among individuals with elevated BPD features [28]. However, these patterns are not consistently replicated: while some studies report no overall group differences in prosocial preferences [10, 28, 29], others note inconsistent trust and punishment behavior in individuals with BPD [30, 31], or point to context-specific impairments, particularly following positive social interactions [17, 32, 33].

A review by Jeung et al. [22] suggested that individuals with BPD tend to act in a more self-interested and economically rational manner in these paradigms. But overall mixed findings suggest the need for a more nuanced framework. One promising direction is the distinction between *reactive cooperation* (e.g., non-retaliation and the ability to forgive after unfair treatment) and *active cooperation* (e.g., non-exploitation and a willingness to share or give) [28, 34, 35]. Research indicates that borderline personality features are associated with deficits in reactive cooperation but not necessarily in active cooperation [28, 34]. This may explain why individuals with BPD typically perform similarly to controls in tasks like the Social Value Orientation (SVO) task or the Dictator Game (DG) but diverge in contexts like the UG or the trustee role in the TG, which require reactive responses.

To better understand the mechanisms underlying reduced cooperation in their TG, King-Casas et al. [25] proposed that individuals with BPD may struggle to recognize or interpret norm violations [17, 25]. Unfair offers may confirm negative expectations, making it less likely that such behaviors are perceived as violations [10, 17, 25]. These results are complemented by studies reporting that individuals with BPD also fail to respond cooperatively after positive interactions [17, 32]. A related account suggests that these impairments stem from a failure to recognize fair or positive interactions [32], which may, in turn, be influenced by exaggerated ideas about fairness and justice [36]. Consistent with this view, Lis et al. [37] found that BPD traits are associated with heightened justice sensitivity—directed both toward oneself and toward others. Justice sensitivity is specifically linked to perceiving others' behavior as unfair [38], and may thus influence fairness perception in economic games. Fairness perception, in turn, is closely related to inequality aversion—a preference against unequal outcomes [39]. We therefore propose that individuals with BPD may also exhibit heightened inequality aversion. However, this hypothesis remains to be systematically tested, as inequality aversion has not yet been directly studied in the context of BPD. Notably, some studies report no differences between individuals with and without BPD in fairness perception or recognition of norm violations [30, 40–42], highlighting the need for further research to clarify these inconsistencies.

In sum, prior research highlights that cooperative behavior in individuals with BPD varies considerably across social exchange paradigms, depending on the specific task structure, the individual's role within the interaction, and the situational context. Adding complexity, specific personality traits may modulate responses to different situational affordances [43, 44]. In particular, Hepp and Niedtfeld [44] propose that the ICD-11 personality trait dissociality—marked by reduced empathy, increased hostility, and a tendency to exploit others [45]–may contribute to reduced prosocial behavior in situations that entail negative reciprocity.

Hypotheses

To address the mixed findings of previous studies on cooperative behavior in individuals with BPD, we used a wide set of economic games with a focus on simple one-shot tasks (Fig. 1) to test a well characterized sample of BPD patients and a healthy control (HC) group on their social preferences, their ability to cooperate in active and reactive situations, as well as their fairness perception and inequality aversion. Based on previous literature, we expected no differences in active cooperation but impairments in reactive cooperation and an altered inequality aversion and fairness perception in individuals with BPD compared to a control group. We tested active cooperation using two different tasks, in which participants actively allocate money between themselves and another anonymous person: the Social Value Orientation (SVO) Slider Measurement [46] and the Dictator Game (DG; [47, 48]). To assess reactive cooperation, participants were engaged in an Ultimatum Game (UG; [49]) in the role of the recipient and had to decide whether to accept or reject an offer from an anonymous person. Taking a dimensional approach, we investigated whether the maladaptive trait of dissociality as assessed by the German version of the Personality Inventory for ICD-11 (PiCD [45, 50]) is associated with reduced reactive cooperation (reciprocal fairness) as suggested by Hepp and Niedtfeld [44]. To examine the reliability of choices in the UG, participants were asked to indicate the minimum amount they would accept from an anonymous proposer. To differentiate between decisions that participants actually

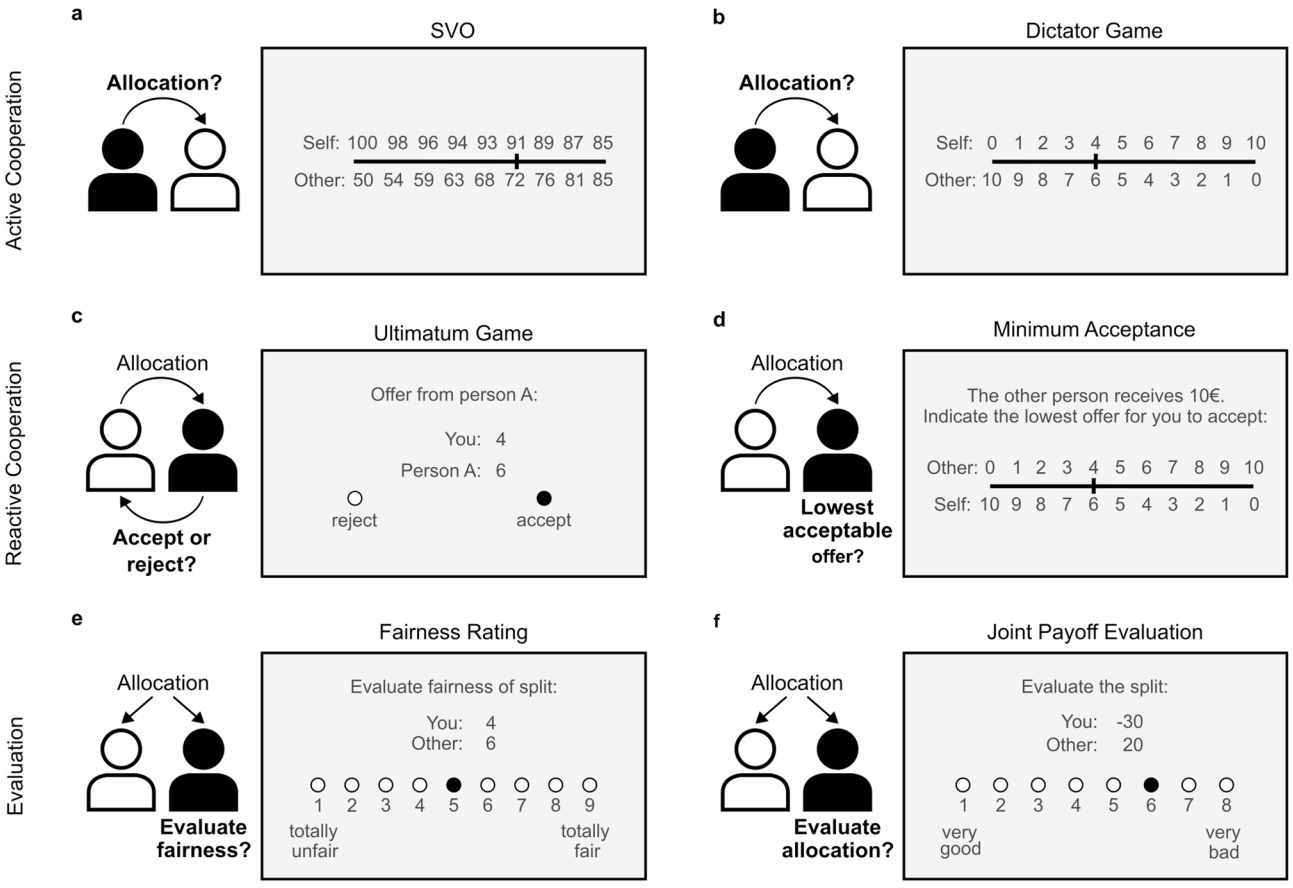


Fig. 1 Overview of the tasks and concepts. The first row depicts active cooperation tasks, the second row depicts reactive cooperation tasks, and the third row depicts evaluations of hypothetical allocations to measure fairness perception and inequality aversion. The black icon represents the participant and the white icon the anonymous partner. The arrows indicate how the resources are being allocated. The bold text near the icons signifies the participant's specific task assignment in the depicted scenario. **a** Example item of the Social Value Orientation (SVO) task to measure general social value orientation and a categorical classification (altruistic, prosocial, individualistic, or competitive). Participants were asked to choose one of nine predefined options to allocate money between themselves and another participant. **b** In the Dictator Game (DG) participants were in the role of the proposer and had to allocate 10€ between themselves and another participant. **c** Example offer in the Ultimatum Game (UG). Participants acting as responders, had to decide whether to accept or reject offers from six anonymous participants. Accepting an offer resulted in both participants receiving the proposed allocation, whereas rejecting it meant neither party received any payout. **d** In the Minimum Acceptance Rating question, participants were asked to indicate the minimum amount they would require the other participant to offer for them to accept the split of 10€. **e** Example item of the Fairness Ratings. Participants had to assess the fairness of 11 hypothetical offers using the same allocation options as in the DG and UG. **f** Example item of the Joint Payoff Evaluation (JPE) task. In this task, participants rated payoff pairs (ranging from − 50 to 50) for themselves and another participant

make and decisions that participants consider fair, we explicitly prompted them to assess the fairness of various hypothetical offers using the same allocation options as in the DG and UG. In addition, we employed a computational modeling approach on an adapted version of a Joint Payoff Evaluation (JPE) task [51], wherein participants evaluated a broader range of allocation pairs. This allowed us to better understand the mechanisms governing social preferences and test potential group differences between specific parameters.

We state our preregistered (https://osf.io/jkvh8) hypotheses in the following numbered list.

H1. We hypothesized no group differences in social value orientation (primary items of SVO task) because we did not expect differences in active cooperation (non-exploitation) between individuals with and without BPD.

H2. We expected individuals with BPD to be more sensitive to inequality and therefore hypothesized a higher inequality aversion (and lower joint gain maximization) in the secondary items of the SVO task compared to the control group.

H3. We expected that our findings would replicate previous research [28, 43] leading to the hypothesis of no group differences in a Dictator Game (active cooperation).

H4. We expected reduced reactive cooperation in individuals with BPD and therefore, hypothesized a higher rate of offer rejections in the Ultimatum Game in the BPD group compared to the control group.

H5. In line with our expectation of impaired reactive cooperation among individuals with BPD, we hypothesized that individuals with BPD would require a higher offer compared to controls to accept an allocation.

H6. We hypothesized that the BPD group would judge fair offers as less fair than individuals in the control group would.

H7.1. We postulated that the Fehr-Schmidt model [39], which offers a mathematical framework that considers own gain, disadvantageous inequality, and advantageous inequality, would provide the best fit for both groups' data in the Joint Payoff Evaluation task.

H7.2. Given our expectation of heightened inequality aversion among individuals with BPD, we hypothesized greater weights on the disadvantageous inequality aversion parameter of the Fehr-Schmidt model in the BPD group compared to the control group.

H8. We hypothesized that higher dissociality scores would correlate with higher rejection rates in the Ultimatum Game (situational affordance of negative reciprocity, see [44]).

Since we could not replicate previous findings on higher rejection rates in the UG in the BPD group compared to the HC group [28, 32], we conducted a mini meta-analysis [52] to clarify the overall evidence and synthesize the results across relevant studies. The meta-analysis was not part of our preregistration plan.

Methods

Ethics and preregistration

This study is part of a larger project on social decision making in women with BPD. All participants also took part in another behavioral experiment and some participants took part in an fMRI experiment, which will be reported elsewhere. We preregistered this study on June 1, 2023, on the Open Science Framework (https://osf.io/jkvh8) after collection but before any analyses of the data. We analyzed the PiCD questionnaire [45, 50], which includes the dissociality scores, as part of other experiments that are preregistered separately (https://osf.io/jwky9 and https://osf.io/sd2tm). We did not observe any of the behavioral data (SVO task, JPE task, DG, UG, Minimum Acceptance Rating, Fairness Ratings) before uploading our preregistration. Changes in the protocol are reported in the Methods section.

Participants

We recruited female participants between 18 and 40 years of age who are fluent in German. Control participants were matched for age during the recruitment process. All participants underwent a prescreening via a telephone interview and were invited to the study if they met our inclusion criteria. They were briefed on the study procedure and gave written informed consent before participating. During the in-person assessment, participants took part in a diagnostic interview, a brief cognitive task, and a set of self-report questionnaires (see Supplementary Table S1 for the full list) via the German survey system SoSci Survey [53]. All participants were debriefed at the end of the study. Refer to the Supplementary Material (section *Supplementary Methods*) for details on our power analysis, the preregistered inclusion and exclusion criteria, recruitment of participants and reimbursement.

Clinical assessments

We assessed Axis I disorders with the short version of the open access Diagnostic Interview for Mental Disorders (Mini-DIPS Open Access; [54]) and BPD diagnostic criteria with the German version of the International Personality Disorder Examination (IPDE; [55, 56]). The clinical interview was conducted by trained and supervised raters and typically lasted approximately 3 h. Additionally,

Table 1 Demographic characteristics of both groups

	BPD (*N* = 35)		HC (*N* = 50)	
	N	%	*N*	%
Education				
Secondary general school (9 years)	1	2.86	0	0
Intermediate school (10 years)	8	22.86	1	2
High school diploma (13 years)	26	74.29	49	98
Professional qualification				
Non, in professional training	9	25.71	28	56
Non, not in professional training	5	14.29	0	0
Apprenticeship	6	17.14	1	2
Vocational school	10	28.57	3	6
Technical school	0	0	1	2
University	5	14.29	17	34
Employment status				
Student	7	20	37	74
In training	6	17.14	0	0
Employed	10	28.57	9	18
Unemployed	5	14.28	2	4
Retired or homemaker	2	5.71	0	0
Other	5	14.29	2	4

Table 2 Comorbid diagnoses and medication intake in the BPD group

	N	%
Comorbid Diagnosis		
Panic disorder	4	11.43
Agoraphobia with panic	1	2.86
Specific phobia	6	17.14
Social anxiety disorder	15	42.86
Major depressive episode	2	5.71
Major depressive disorder, recurrent	17	48.57
Persistent depressive disorder	3	8.57
Obsessive-compulsive disorder	1	2.86
Body dysmorphic disorder	1	2.86
Posttraumatic stress disorder	12	34.29
Bulimia nervosa	1	2.86
Binge eating disorder	2	5.71
Illness anxiety disorder	2	5.71
Substance abuse, alcohol	1	2.86
Substance abuse, cannabis	2	5.71
No Axis I diagnosis	4	11.43
Medication		
Antidepressant, SNRI	5	14.29
Antidepressant, SSRI	10	28.57
Antidepressant, atypical	1	2.86
Atypical antipsychotics	3	8.57
Mood stabilizer (Lamotrigine)	2	5.71
No medication	17	48.57

Comorbid diagnoses were tested with the mini-DIPS diagnostic interview

participants underwent a three-minute speeded reasoning test (mini-q; [57]) to estimate their general cognitive abilities. Furthermore, we assessed depressive symptoms with the Beck Depression Inventory-II (BDI-II; [58]), a General Severity Index (GSI) for self-reported psychopathological symptoms experienced in the previous seven days with the Brief Symptom Checklist (BSCL; [59]), the severity of BPD with the Borderline Symptom List-23 (BSL-23; [60]), impairments of personality functioning with the Level of Personality Functioning Scale-Brief Form (LPFS-BF; [61, 62]), and personality dysfunction with the Personality Inventory for the DSM-5-Brief Form (PID-5-BF; [63, 64]).

Sample description

We decided to analyze all valid datasets (according to our inclusion and exclusion criteria) for the current study and included 35 patients with BPD and 50 controls in our data analysis. Age ranged from 18 to 39 in both groups and did not differ significantly between the groups (M_{BPD} = 25.8, SD_{BPD} = 5.58, M_{HC} = 24.88, SD_{HC} = 5.43, Welch's two-sample t-test $t_{72.035}$ = 0.756, p = 0.452, 95% *CI* [−1.505, 3.345], d = *−0.167*; Wilcoxon rank sum test z = 0.842, p = 0.4, r = 0.092). Groups did not differ in their general cognitive abilities (mini-q; M_{BPD} = 34.03, SD_{BPD} = 12.12, M_{HC} = 35.74, SD_{HC} = 9.91, Welch's two-sample t-test $t_{63.579}$ = −0.689, p = 0.493, 95% *CI* = [−6.671, 3.248], d = −0.155; Wilcoxon rank sum test z = −0.541, p = 0.589, r = 0.059). The most common school-leaving certificate among participants was a high school diploma. However, a higher percentage of participants in the HC group (98%) completed high school compared to the BPD group (74.29%). Employment status also differed between the groups: Most participants in the BPD group were currently employed (28.57% in the BPD group vs. 18% in the HC group) while the majority of participants in the HC group were students (74% in the HC group vs. 20% in the BPD group). For detailed demographic data, refer to Table 1. For the BPD group we included participants who met the threshold of five IPDE criteria and participants with sub-threshold BPD who met at least four IPDE criteria and had a prior BPD diagnosis. Within the BPD group, 32 participants had received a formal BPD diagnosis prior to participation in our study and for three participants, suspicion of a BPD diagnosis had been raised by either their psychiatrist or psychotherapist. In the BPD group 88.57% of participants fulfilled criteria of comorbid diagnoses according to the mini-DIPS interview (Table 2). About 51% of participants in the BPD group were taking some form of psychotropic medication, primarily antidepressants (Table 2). Information on current medication intake was self-reported by the participants. The BPD group and the HC group differed significantly in BDI-II, BSCL, BSL-23, IPDE, LPFS, and PID-5-BF (see Table 3 for details).

Table 3 Clinical characteristics of both groups

Questionnaire	BPD ($N = 35$)		HC ($N = 50$)		Welch's t-test					Wilcoxon rank sum test		
	M	*SD*	*M*	*SD*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*p*	*d*	*z*	*p*	*r*
BDI-II	27.743	12.363	3.720	4.071	11.083	39.201	19.639	28.406	**< 0.001**	2.610	7.443	**< 0.001**	0.807
BSCL	1.584	0.737	0.186	0.261	10.764	40.003	1.136	1.661	**< 0.001**	2.530	7.426	**< 0.001**	0.805
BSL-23	1.836	0.818	0.212	0.299	11.226	40.412	1.332	1.916	**< 0.001**	2.636	7.538	**< 0.001**	0.818
IPDE	13.657	2.114	0.160	0.548	36.921	37.219	12.757	14.238	**< 0.001**	8.742	8.473	**< 0.001**	0.919
LPFS	35.629	5.309	18.840	5.068	14.618	71.135	14.499	19.078	**< 0.001**	3.235	7.523	**< 0.001**	0.816
PID-5-BF	37.571	8.859	13.020	7.789	13.207	67.116	20.841	28.262	**< 0.001**	2.943	7.283	**< 0.001**	0.790

CI 95% Confidence Intervals with lower limits (*LL*) and upper limits (*UL*), *d* Cohen's (effect size for Welch's t-test), *r* effect size *r* for Wilcoxon rank sum test, *BDI-II* Beck Depression Inventory-II (total score), *BSCL* Brief Symptom Checklist (General Severity Index), *BSL-23* Borderline Symptom List-23 (mean score), *IPDE-Dimensional* International Personality Disorder Examination (dimensional score), *LPFS* Level of Personality Functioning Scale-Brief Form (total score), *PID-5-BF* Personality Inventory for DSM-5-Brief Form (total score); significant results in bold font

Tasks

In a set of behavioral tasks (see Fig. 1 for an overview), participants were asked to either allocate payoffs or respond to payoff combinations for themselves and anonymous others. All interactions were one-shot exchanges, meaning participants did not engage with the same partner more than once. They were led to believe that the other people were also participants in the study. Participants were informed that they would receive a bonus based on their decisions and the decisions of other participants from the same study. That is, a few trials were drawn randomly and translated into a bonus payoff of a maximum of 5€. We chose a specific order of the tasks to minimize the influence between the tasks (see Supplementary Material section *Ordering and randomization of the tasks*).

SVO task

All participants performed the SVO Slider Measurement [46] which measures participants' general social value orientation and prosocial motivation. Social value orientation delineates how individuals evaluate the allocation of resources between themselves and others and serves as a measure for active cooperation. In addition to continuous scores, the task provides a categorical classification of participants according to the primary items (altruistic, prosocial, individualistic, and competitive) and secondary items (joint gain maximizing and inequality averse).

The SVO Slider Measurement consists of 15 items (trials): six primary items to measure the general SVO score and nine secondary items to measure the prosocial motivation score. Each of the items consists of nine different prespecified joint payoff options. For example, the primary item 6 ranges from the first option with 100 points for self and 50 points for the other to the ninth option with 85 points for self and 85 points for the other (Fig. 1a or Supplementary Figure S1a). Here, a more prosocial participant would try to avoid inequality and maximize joint gain and choose option 9 (with 85 for self and 85 for the other) or one of the options close to option nine. A more individualistic participant would try to maximize their own gain and choose the first option (with 100 for self and 50 for the other) or an option close to the first option.

The secondary items can be used to differentiate between two different motivations for being prosocial: joint gain maximization and inequality aversion. For example, the secondary item 9 ranges from the first option with 100 points for self and 70 points for the other to the ninth option with 50 points for self and 100 points for the other (Supplementary Fig. 1b). Here, a joint gain maximizing participant would choose one of the options with the highest joint outcome (e.g., option 1 with 100 for self and 70 for the other), while a more inequality averse

participant would choose one of the options that minimize inequality (e.g., option 4 with 81 for self and 81 for the other).

Dictator game

Participants played a single trial of a one-shot Dictator Game (DG; [47, 48]) in the role of the proposer and were asked to allocate 10€ between themselves and another anonymous participant as they like. Similar to the SVO task, the proposer had full control over the allocation. The allocation options ranged from 0 to 10 in steps of 1, with the amount shared serving as an indicator of active cooperation.

Ultimatum game

To assess reactive cooperation, participants were engaged in six one-shot UG [49] trials, each with a different anonymous proposer, in the role of the responder. The UG is similar to the DG with a key difference: The responder can either accept or reject the proposer's offer. If the responder accepts, both players receive the proposed payoff allocation but if the responder rejects neither of them gets any money. In this context, rejection rates serve as an indicator of punishment for perceived unfair behavior and a measure for reactive cooperation [28]. In each trial, we asked participants to either accept or reject a split of 10€ offered by an anonymous participant. The six offers ranged from 1€ to 6€ in steps of 1. In line with a previous study, we considered offers from 1 to 3 as unfair and offers from 4 to 6 as fair [32].

Minimum acceptance rating

To examine the reliability of choices in the UG, participants played one trial of a variant of the UG [28] and were asked to indicate the minimum amount the other anonymous participant would have to offer for the participant to accept a split of 10€. Acceptance options ranged from 0 to 10 in steps of 1.

Fairness ratings

To test potential group differences in fairness perception, we explicitly prompted participants to assess the fairness of various hypothetical offers using the same allocation options as in the DG and UG. Offers from another anonymous participant ranged from 0 to 10€ (in steps of 1, resulting in 11 trials). Perceived fairness was rated on a 9-point Likert scale (1 = totally unfair, 9 = totally fair).

JPE task

All participants completed an adapted version of the Joint Payoff Evaluation (JPE) task [51]. In this task, participants rated pairs of payoffs for themselves and another anonymous participant on an 8-point Likert scale (1 = very good, 8 = very bad). Payoffs ranged from −50 to 50 in steps of 10 (resulting in 11*11 = 121 trials). We employed a computational modeling approach to test nine different economic models, aiming to describe participants' motives when evaluating payoff allocations.

Questionnaire

Dissociality

Participants completed the German version of the Personality Inventory for ICD-11 (PiCD [45, 50]) to assess the five ICD-11 maladaptive personality domains. The PiCD is a self-report measure that consists of 60 items, with 12 items for each of the five domains: Negative Affectivity, Detachment, Dissociality, Disinhibition, and Anankastia. Each item is rated on a 5-point Likert scale (from 1 = strongly disagree to 5 = strongly agree).

Analyses

We use the standard alpha level of $p < 0.05$ to determine whether results from the GLMM and t-tests differed significantly from those expected under the null hypothesis.

We report results from Welch's two-sample t-test, which is considered more robust than the Student's t-test [65]. To ensure robustness of our findings, we also conducted the Wilcoxon rank sum tests as a non-parametric alternative, which rely on ranking of observations and do not assume a specific distribution. For the experimental tasks and the dissociality factor of the PiCD questionnaire, t-tests were performed using MATLAB's (2024) *ttest2* and *ranksum* functions. For clinical assessment, analyses were performed in R [66] using the *stats* [66] and the *coin* [67] packages. The effect sizes *Cohens'd* for Welch's two-sample t-tests and Wilcoxon effect size *r* for Wilcoxon rank sum tests were calculated in R using the package *rstatix* [68]. For data visualization, we used R and the packages *ggplot2* [69] and *gghalves* [70].

We calculated Bayes Factors (BFs) for the tasks in which we expected no group differences to provide evidence for the null hypothesis versus the alternative hypothesis (Bayes Factor approach to equivalence testing). Bayes Factors were calculated in R [66] using the package *BayesFactor* [71] with default non-informative priors and interpreted by using the classification scheme by Jeffreys [72]. We report the BF01, which provides the posterior odds for the null hypothesis versus the alternative hypothesis and equals 1/BF10 [73].

SVO task

We computed the SVO score from the six primary items following the procedure described in Murphy et al. [46].

In addition, we computed the prosocial motivation score from the nine secondary items for all participants who showed consistent prosocial choices in both primary and secondary items (see Supplementary Material section *Supplementary Analyses* for details). The secondary

items can be used to identify the underlying motivation for being prosocial, namely being inequality averse or aiming to maximize the joint gain.

We compared the primary SVO score between the BPD group and the healthy control group using two-sample t-tests. In addition, we calculated BFs to provide evidence for the null hypothesis versus the alternative hypothesis. For participants with consistent primary and secondary prosocial choices, we compared the secondary SVO scores between the groups using two-sample t-tests. For consistency and completeness, we provide the BFs.

Dictator game

We compared the amount of money allocated to the anonymous other between the groups using two-sample t-tests. In addition, we calculated BFs to provide evidence for the null hypothesis versus the alternative hypothesis.

Ultimatum game

We set up a Generalized Linear Mixed Model (GLMM) to test the association between rejection of offer (yes/no) as the dependent variable as well as group (BPD/HC), different offer amounts (1€, 2€, 3€, 4€, 5€, 6€), and their interaction as fixed effects and participants (subject ID) as random effects. Offer amounts were mean centered and we used the HC group as the reference category.

In addition, we set up a second GLMM to test if higher dissociality scores (PiCD [45, 50]) are related to higher rejection rates in the UG. We tested a logistic mixed model with rejection of offer (yes/no) as the dependent variable, dissociality and different offer amounts (1€, 2€, 3€, 4€, 5€, 6€) as fixed effects, and participants (subject ID) as random effect. We also tested interactions between the fixed effects variables. Offer amounts and dissociality scores were mean centered. Note that the models differ from our preregistered models due to incorrect specifications in the preregistration.

The analysis was performed with the MATLAB function *fitglme.* We specified our models with a binomial response distribution and a logit link function and used MATLAB's maximum pseudo likelihood (MPL) estimation to fit the models. We calculated Odds Ratios (OR) by exponentiating the coefficients.

Minimum acceptance rating

We tested the smallest accepted offer between the groups using two-sample t-tests.

Fairness ratings

We set up a GLMM to test the association between perceived fairness (ratings from 1 = totally unfair, 9 = totally fair) as the dependent variable, group (BPD/HC), different offer amounts (0€ to 10€), and their interaction as fixed effects, and participants (subject ID) as random effect. Offer amounts were mean centered and included as a quadratic term. The analysis was performed with the MATLAB function *fitglme.* We specified our models with a normal response distribution, an identity link function and used MATLAB's MPL estimation to fit the models with the HC group as the reference category.

JPE task

We compared nine different economic models to test participants' social motives in evaluating the payoff pairs. Each participant's evaluation during the JPE task was linearly regressed with different predictors of the specified economic models. See Supplementary Material section Supplementary *Analyses* for model descriptions.

To assess model evidence, we conducted both fixed-effects analyses, using log-group Bayes Factors, and random-effects analyses, using protected exceedance probabilities (see Supplementary Material section *Supplementary Analyses* for details).

We postulated that the Fehr-Schmidt model, which considers own gain, disadvantageous inequality, and advantageous inequality, would provide the best fit for both group's data and that individuals from the BPD group would show more inequality aversion. Therefore, we tested if the BPD group shows higher weights on the disadvantageous inequality aversion parameter compared to the control group. We conducted two-sample t-tests to compare model parameter weights between the groups.

Mini meta-analysis on ultimatum games

The meta-analysis was conducted in R [66] using the *meta* package [74], following the guidelines outlined by Harrer et al. [75]. For each study, between-group effect sizes (standardized mean differences) were calculated as Cohen's d. For the study reporting BPD feature scores, participants were divided into groups based on established cut-off criteria [76]. If studies reported outcomes separately for additional conditions, the data were aggregated across these conditions before calculating the effect sizes. Detailed information about the study selection, analyses, and characteristics of the included studies are provided in the Supplementary Material section *Mini Meta-Analysis on Ultimatum Games*.

Results

SVO task

General social value orientation (H1)

We employed the SVO measurement to test hypothesis H1 that individuals with BPD and controls do not differ in their general social value orientation and willingness to actively cooperate in the primary items of the SVO task.

The majority of participants in both groups was categorized as prosocial (BPD = 89.66%, HC = 98%), with

the remaining few categorized as individualistic (BPD = 10.34, %, HC = 2%).

When analyzing the continuous SVO scores, we found a significant difference between the BPD group and the control group with the control group demonstrating higher, and therefore more prosocial scores (Table 4; Fig. 2a). The Bayes Factor analysis revealed a BF01 of 0.47, indicating only anecdotal evidence for the absence of an effect [72, 77]. While continuous SVO scores indicated a potential group difference, the Bayesian analysis did not provide strong evidence for neither the null hypothesis nor the alternative hypothesis. Given that the vast majority of participants in both groups were categorized as prosocial, we cannot draw definitive conclusions regarding differences in active cooperation between individuals with BPD and healthy controls.

Prosocial motivation (H2)

To differentiate between two different motivations for being prosocial (joint gain maximization and inequality aversion) and test hypothesis H2 that individuals with BPD are more inequality averse than healthy controls, we analyzed the secondary items of the SVO measurement.

Among the prosocial participants, all in the BPD group were categorized as inequality averse, as were the majority in the control group (inequality averse N_{Hc} = 88.64%, joint gain maximizing N_{Hc} = 11.36%; Supplementary Figure S2). We did not find a significant difference in prosocial motivation scores between the groups (Table 4). For consistency, we calculated Bayes Factors. The analysis revealed a BF01 of 3.49, indicating that the data were approximately 3.5 times more likely under the null hypothesis than under the alternative hypothesis (BF10 = 0.29). This provides moderate evidence for the absence of an effect [72]. These results suggest no difference in inequality aversion (versus joint gain maximization) in individuals with BPD compared to healthy controls.

Dictator game (H3)

To test Hypothesis H3—that the groups do not differ in active cooperation in the DG—we compared the amount of money allocated to another person in the DG. The results showed that participants with BPD shared a similar amount to participants from the control group, with no significant difference observed (Table 4; Fig. 2b). We conclude that participants from both groups exhibited equally cooperative behavior in the DG. The Bayes Factor analysis revealed a BF01 of 0.92, demonstrating only anecdotal evidence in favor of the null hypothesis [72, 77].

Ultimatum game (H4)

We used the UG to test H4 that individuals with BPD, compared to individuals without BPD show impairments

Table 4 Summary of statistical results comparing the BPD and HC group

Task	Hypothesis	BPD			HC			Welch's t-test						Wilcoxon rank sum test		
		M	*SD*	*N*	*M*	*SD*	*N*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*p*	*d*	*z*	*p*	*r*
SVO (G)	H1	32.421	7.127	29	35.579	5.203	50	−2.085	45.497	−6.207	−0.109	**0.043**	−0.506	−1.989	**0.047**	0.224
SVO (PM)	H2	0.264	0.101	24	0.287	0.211	44	−0.592	65.242	−0.098	0.053	0.556	−0.136	0.225	0.822	0.028
DG	H3	4.457	0.980	35	4.780	0.582	50	−1.745	50.683	−0.694	0.049	0.087	−0.400	−1.456	0.145	0.159
UG	H4	36.190	23.391	35	40.476	26.571	49	−0.782	78.389	−0.152	0.066	0.437	−0.717	−0.746	0.455	0.082
MA	H5	4.057	2.222	35	4.280	1.852	50	−0.487	64.520	−1.138	0.692	0.628	−0.109	−0.279	0.780	0.031
JPE	H7.2	0.023	0.029	35	0.025	0.026	50	−0.367	69.104	−0.015	0.010	0.715	−0.082	−0.246	0.806	0.027

This table reports results from task-based comparisons between the BPD and HC groups using Welch's t-tests and Wilcoxon rank sum tests. Task: abbreviated name of the experimental task (and measure); Hypothesis: preregistered hypothesis; *CI* 95% Confidence Intervals with lower limits (*LL*) and upper limits (*UL*), *d* Cohen's (effect size for Welch's t-test), *r* effect size *r* for Wilcoxon rank sum test; Task abbreviations and dependent variables measured: *SVO* Social Value Orientation (G = general SVO score measured with the primary items, *PM* Prosocial Motivation score measured with the secondary items), *DG* Dictator Game (allocation to the other), *UG* Ultimatum Game (rejection rates), *MA* Minimum Acceptance rating (minimum offer required to accept the split), *JPE* Joint Payoff Evaluation task (disadvantageous inequality aversion parameter weights of the Fehr-Schmidt model); significant results in bold font

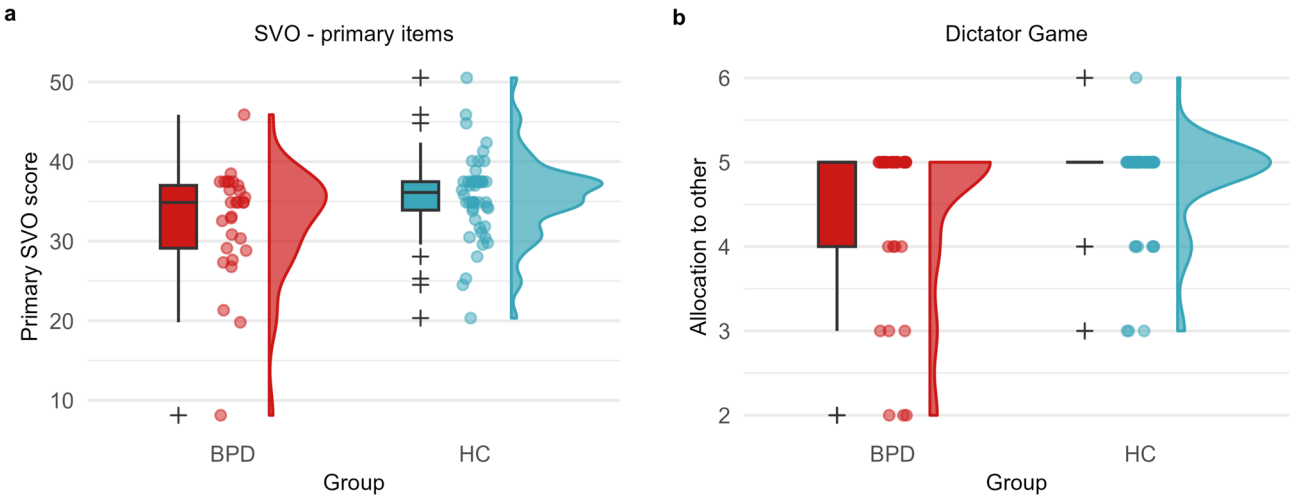


Fig. 2 Active cooperation task results. Raincloud plots show the distribution of (**a**) SVO scores (primary items) and (**b**) the distribution of allocations of 10€ to the anonymous partner in the BPD group and the HC group. The plot includes individual scores for each participant, a box plot, and a probability density plot. The boxplot visualizes the median, the 25th and 75th percentiles, two whiskers (extending from the upper and lower hinges to the largest and lowest value within 1.5 times the inter-quartile range, and all outliers individually. The density plot was generated using Gaussian kernel density estimation

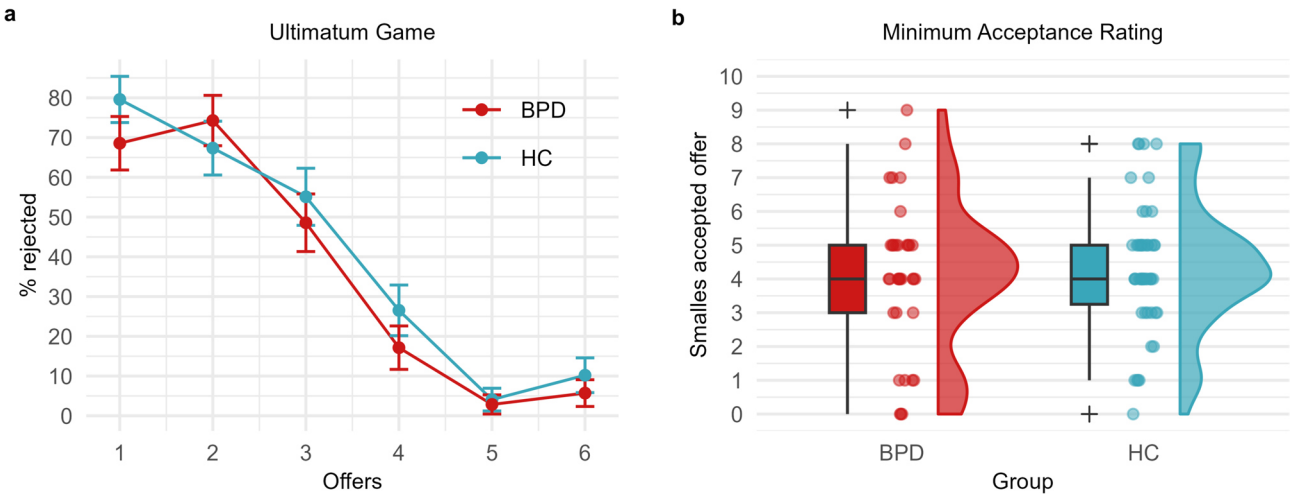


Fig. 3 Reactive cooperation task results. **a** Mean rejection rates for each possible offer in the UG in the BPD group and the HC group. Error bars show the standard error of the mean (SEM). **b** Raincloud plots show the distribution of the smallest accepted offers in the Minimum Acceptance Rating in the BPD group and the HC group

in reactive cooperation and therefore higher rejection rates. Across all possible offers, the mean rejection rates did not differ significantly between the groups (Table 4).

We set up a GLMM to test the association between rejection of offer (yes/no) and group (BPD/HC), different offer amounts (1€, 2€, 3€, 4€, 5€, 6€), and the interaction between these factors. We included the data of 35 BPD participants and 49 HC participants in this analysis. The results indicated that for the HC group (reference) the *Offer* amounts significantly predicted rejection rates (b = −1.14, SE = 0.127, $t_{(500)}$ = −8,971, p < 0.001, 95% CI [−1.389, −0.89], OR = 0.32). The *Group* (b = −0.335, SE = 0.456, $t_{(500)}$ = −0.736, p = 0.462, 95% CI [−1.231, 0.56], OR = 0.715) and *Group x Offer* interaction (b = −0.007, SE = 0.2, $t_{(500)}$ = −0.037, p = 0.971, 95% CI [−0.4, 0.386], OR = 0.993) were not significant predictors. These results suggest that the rejection rates seem to depend solely on the offer amounts and that there is no reduction in reactive cooperation in the BPD group compared to the control group (see Fig. 3a).

Dissociality (H8)

Taking a dimensional approach, we investigated whether the maladaptive trait of dissociality is associated with reduced reactive cooperation as specified in hypothesis H8. We specified a second GLMM to test the association

between rejection of offer (yes/no) and different offer amounts (1€, 2€, 3€, 4€, 5€, 6€), dissociality scores, and the interaction between these factors. We included the data of 35 BPD participants and 49 HC participants in this analysis. Again, the results indicated that *Offer* significantly predicted rejection (b = −1.142, SE = 0.098, $t_{(500)}$ = −11.644, $p < 0.001$, 95% CI [−1.335, −0.949], *OR* = 0.319). The effect of *Dissociality* (b = 0.002, SE = 0.027, $t_{(500)}$ = 0.089, $p = 0.929$, 95% CI [−0.051, 0.056]. *OR* = 1.002) and *Offer* x *Dissociality* interaction (b = 0.003, SE = 0.012, $t_{(500)}$ = 0.25, $p = 0.803$, 95% CI [−0.02, 0.026], *OR* = 1.003) were not significant predictors. We conclude that higher dissociality scores are not related to higher rejection rates in the UG. We report group mean comparisons for dissociality scores and descriptive statistics in the Supplementary Material section *Supplementary Results* and Supplementary Figure S3).

Minimum acceptance rating (H5)

To test H5—that individuals with BPD show impaired reactive cooperation and therefore require higher offers to accept an allocation compared to controls—and to examine the reliability of choices in the UG, we tested for differences in the smallest accepted offers between the groups. The results showed that participants with BPD did not require a higher offer compared to participants from the control group (Table 4; Fig. 3b) to accept.

Fairness ratings (H6)

To test H6—that the BPD group would judge fair offers as less fair than the HC group—we set up a GLMM with linear and quadratic terms to examine the association between perceived fairness and group, different offer amounts, and their interaction (see Table 5 for the full model summary). For the HC group (reference), there was a significant positive linear effect of *Offer* and a significant negative quadratic effect, suggesting an inverted U-shaped relationship between offer amounts and fairness ratings. A significant linear *Group* x *Offer* interaction revealed that the BPD group exhibited a steeper slope at the mean offer level (offer = 5) compared to the HC group. Importantly, the linear interaction reflects group differences in the slopes at the center of the offer range, not a consistent group difference across all offer levels, as the slopes change across the offer range due to the quadratic component. The *Group* x quadratic *Offer* interaction was not significant, suggesting that the quadratic relationship between offer and fairness rating was similar across groups. Together, these results indicate that offers are perceived as increasingly fair up to a certain value, after which higher offers are perceived as less fair by both groups (see Fig. 4 for group-wise fairness ratings across offers). We report descriptive statistics for disadvantageous offers (offer 0 to 3), fair offers (4 to 6), and advantageous offers (7 to 10) for both groups in the Supplementary Table S2.

JPE task (H7)

To explore participants' motives when evaluating payoff allocations, we tested how well nine different economic models fit their evaluations of allocation pairs. We included the data of 35 BPD participants and 49 HC participants in this analysis. As hypothesized in H7.1, the fixed-effects analysis based on log-group Bayes Factors revealed that the Fehr-Schmidt model, which accounts for payoffs for self, disadvantageous, and advantageous inequality, was the best fitting model in both groups. We performed an additional random-effects analysis using protected exceedance probabilities and found that the Inequality & Joint-Gain model was the best-fitting model in both groups (Supplementary Figure S4). The inequality & Joint-Gain model is also the second-best model in both groups according to the log-group Bayes Factors. This implies that the Inequality & Joint-Gain model performed better in more participants in both groups. For those participants, for whom the Fehr-Schmidt model performed better, it did so by a larger margin. To test H7.2—that individuals with BPD show heightened inequality aversion compared to controls—we compared disadvantageous inequality parameter weights of the Fehr-Schmidt model. We found no significant difference in the disadvantageous inequality parameter weights between the groups (Table 4). In conclusion, inequality and gain magnitude are important factors for participants in both groups when evaluating joint payoffs, and there is no evidence of heightened inequality aversion in the BPD group compared to the control group.

Table 5 Model summary of the GLMM on the linear and quadratic effects of group, offer, and their interaction on fairness ratings

Predictors	b	SE	t	DF	*p*	CI (LL)	CI (UL)	OR
(Intercept)	5.550	0.151	36.801	929	**< 0.001**	5.254	5.846	257.172
Group [BPD]	−0.114	0.235	−0.487	929	0.626	−0.576	0.347	0.892
Offer	0.055	0.026	2.100	929	**0.036**	0.004	0.106	1.056
Group [BPD] x Offer	0.094	0.040	2.323	929	**0.020**	0.015	0.173	1.099
Offer^2	−0.192	0.009	−20.613	929	**< 0.001**	−0.210	−0.173	0.826
Group [BPD] x Offer^2	0.009	0.014	0.618	929	0.537	−0.019	0.037	1.009

CI 95% Confidence Interval *LL* Lower Limit, *UL* Upper Limit; significant results in bold font; Sample size: $N_{BPD} = 35$, $N_{HC} = 50$

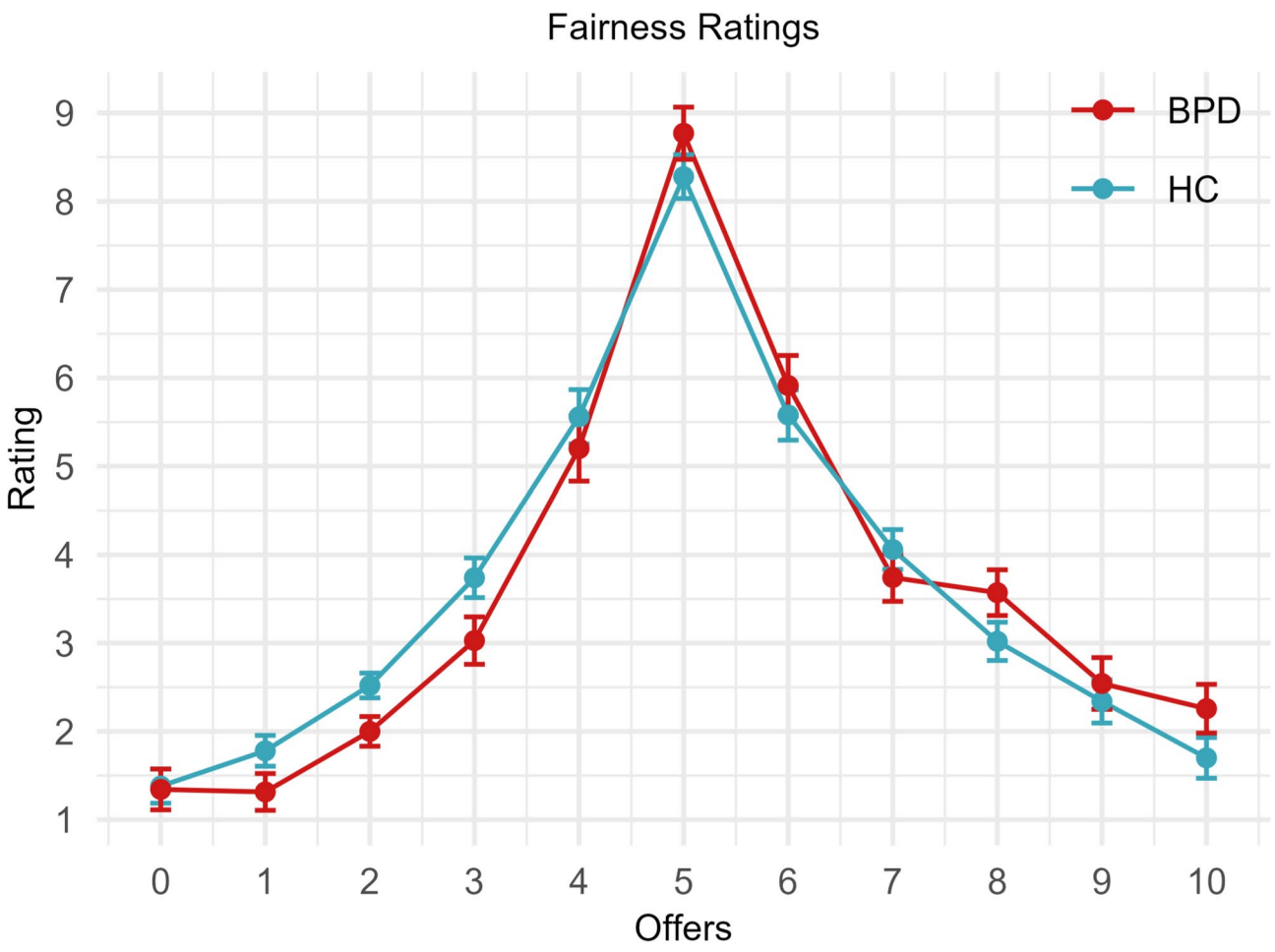


Fig. 4 Mean ratings of fairness for each possible split of 10€ (with steps of 1€) in the BPD group and the HC group. Error bars show the standard error of the mean (SEM). In both groups, offers were perceived as increasingly fair up to the offer of 5€ and perceived as less fair for offers higher than 5 (significant quadratic effect in GLMM). We found no significant group differences in the quadratic effect of offers on fairness perception

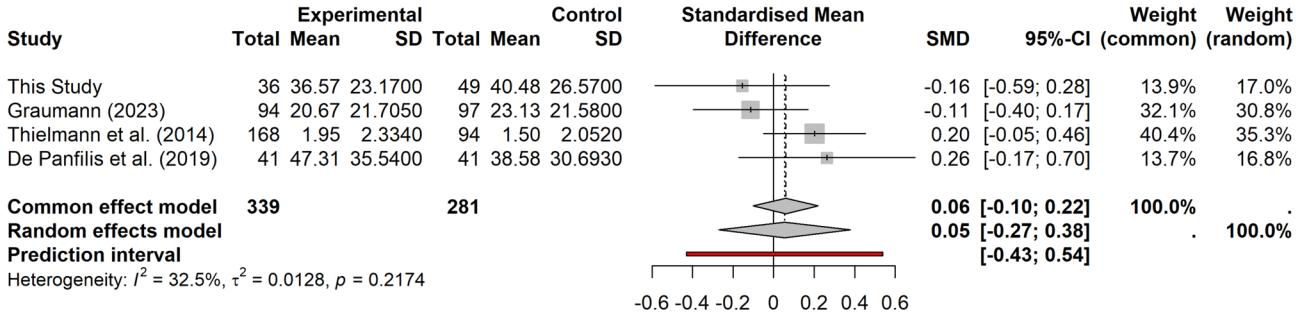


Fig. 5 Mini meta-analysis. Forest plot comparison of studies evaluating rejection rates in the Ultimatum Game in individuals with BPD (features) and controls. The standardized mean difference (SMD) for each study is displayed with the line representing the confidence interval and the square representing the weight in the analysis. The diamond symbolizes the average effect, its length represents the confidence interval. Experimental = BPD group, Control = control group, CI = confidence interval

Mini meta-analysis on ultimatum games

We included four studies that met our inclusion criteria (see Supplementary Material section *Mini Meta-Analysis on Ultimatum Games* for details and Supplementary Figure S5 and Supplementary Table S3) in our meta-analysis and found an overall effect size of $d = 0.054$ (95% CI [− 0.271, 0.379]), indicating a small, non-significant effect ($p = 0.633$). Heterogeneity across studies was low (I^2 = 32.5%, 95% *CI* [0.0, 76]; $\tau^2 = 0.013$, 95% *CI* [0.00, 0.599]), suggesting minimal variability in effect sizes (see forest plot, Fig. 5). Therefore, no significant overall effect of group membership on rejection rates in the Ultimatum

Game was found. A funnel plot (see Supplementary Figure S6) revealed no indication of publication bias.

Discussion

In this study, we aimed to explore the mechanisms underlying social dysfunctions in individuals with BPD by examining their social preferences and behaviors in a comprehensive set of standard economic games. While we hypothesized no differences in active cooperation, we expected lower reactive cooperation and altered fairness perception and sensitivity to inequality in the BPD group. However, when tested in anonymous, one-shot economic games individuals with BPD did not differ from healthy controls in any of the tasks. Our study was strengthened by using multiple tasks within the same well-characterized sample, a preregistered design, and the application of robust and elaborate statistical analyses to ensure rigorous examination of the data. In the following, we discuss active, reactive cooperation, fairness perception and inequality aversion in turn.

Active cooperation

With respect to active cooperation as measured by social preferences, we expected no differences between participants with and without BPD. We found a small but significant difference in continuous SVO scores, namely participants in the control group demonstrated higher SVO scores. However, this result should be interpreted with caution. The average SVO score for the BPD group was close to the SVO scores of the BPD and control samples in a previous study that showed no group differences [10]. Our control group exhibited higher SVO scores than both the BPD and the control group in Lévay et al. [10] and also higher scores than we found in three healthy samples in our previous work [78]. Despite the significant difference between the groups, the average group scores of both groups still fall within the prosocial category. This suggests that the significant difference between the groups may be driven by exceptionally high prosocial preferences in the control group, rather than a low prosocial preference in the BPD group. The Bayes Factor analysis revealed that no clear distinction between the null and the alternative hypothesis is possible. Thus, even though we found significantly lower SVO scores in the BPD group compared to the control group, we cannot conclude that people with BPD are impaired in active cooperation.

With respect to active cooperation as measured by the DG, we found no group differences. Participants in the BPD group shared a similar amount of money compared to the control group. This result is consistent with our hypothesis and previous studies that reported no differences between BPD and HC groups in investment in a DG with real incentives and anonymous partners [29, 31, 43]. Similarly, Thielmann et al. [28] reported that they found no association between BP features and hypothetical offers in a DG in a large community sample ($N = 556$), suggesting that high levels of BP features are not related to impairments in active cooperation. Our results are also in line with a study reporting no group differences in the first round of investment in a TG between a BPD and a healthy control group [17].

Taken together, our results replicate previous studies and suggest that people with BPD show intact prosocial preferences and no impairment in active cooperation (i.e., non-exploitation) [28].

Reactive cooperation

Next, we tested reactive cooperation to determine whether previous findings on impaired reactive cooperation (i.e., non-retaliation) in people with high levels of BP features would also extend to patients diagnosed with BPD [28]. We therefore implemented an UG in which participants played the role of the receiver and were able to retaliate if they found the proposer's offer unfair. We tested if BPD (effect of group) or the maladaptive personality trait dissociality significantly affected rejection rates in the UG. Participants in both groups accepted higher offers at a higher rate and lower offers at a lower rate and there was no significant difference between the groups. Therefore, neither BPD diagnosis nor dissociality scores had an effect on reactive cooperation. To test the robustness of the UG results, we asked participants to indicate the lowest offer they would be willing to accept if a proposer were to share 10€. We found no significant differences between the BPD and control groups regarding minimum acceptance ratings, which suggests no indication of less reactive cooperation in the BPD group. This finding is consistent with our UG results and indicates that the threshold for acceptable offers does not significantly differ between the groups.

Previous studies on BPD and the UG have reported mixed findings. Polgár et al. [79] tested 47 BPD patients and 43 healthy controls in a 40-trial multi-round UG. BPD participants accepted more unfair offers than controls, and while positive facial expressions increased unfair offer acceptance among controls, they had no effect on those with BPD. De Panfilis et al. [32] tested 41 BPD patients and 41 controls in a UG with emotion regulation strategies. BPD participants rejected fair offers more frequently (independent of the emotion regulation strategy used) and perceived them as less fair, suggesting a bias toward underestimating positive feedback. Thielmann et al. [28] examined 559 participants with varying levels of BP features. In a one-shot strategy variant, those with higher BP features demanded higher offers, while in a game variant, they rejected more offers after a breakdown in cooperation.

Our results do not align with previous findings of either lower rejection rates [79] or higher rejection rates in individuals with BPD [32] or elevated BP features [28]. Differences in study design, such as the inclusion of emotion regulation strategies [32], the manipulation of facial expressions [79], or the use of BPD features instead of clinical diagnosis as criterion [28], could have contributed to the variability in outcomes. To assess these differences in the literature, we conducted a mini meta-analysis as part of the present study. The results showed no significant group differences, with small heterogeneity of effect sizes and no indication of publication bias. These findings suggest that BPD-related differences in the UG may not generalize across variations in task design.

Overall, our findings indicate that individuals with BPD are just as capable of forgiving uncooperative behavior as healthy controls. While our results diverge from previous findings using UG paradigms, they are consistent with studies employing multi-round TG paradigms, which have demonstrated that individuals with BPD can show forgiveness in response to unfair behavior [17, 30]. For future research, we recommend the use of more ecologically valid tasks involving real social interactions [3, 43] that are relevant to BPD symptomatology.

Fairness perception and inequality aversion

Regarding fairness perception and sensitivity to inequality, prior research indicates that individuals with BPD often feel excluded, even when included or overincluded [15, 18, 80], perceive fair exchanges as less fair and underestimate positive feedback [25, 32], and exhibit high victim sensitivity [37]. They also tend to punish fair behavior more frequently [15] and struggle to rebuild cooperation after being treated unfairly [25]. Based on these findings, we expected individuals with BPD to rate fair offers as less fair and show greater inequality aversion than controls.

Contrary to our hypothesis, we found no significant group differences in fairness perception in our standard task. These results are consistent with participants' UG behavior and minimum acceptance ratings, suggesting that individuals with BPD do value equality and fairness and do not differ significantly in related behaviors from healthy controls in these economic games. Our results also align with those of Jeung et al. [41], who found that fairness perception remains intact in individuals with BPD. Notably, half of the BPD group in their study chose an unfair interaction partner, which the authors suggest may be linked to BPD patients' negative and instable self-image, contributing to dysfunctional social interactions and relationships [41].

We compared several economic models on participants' evaluations of a broader range of payoff pairs to better understand the mechanisms governing social preferences. Here, we used a large payoff space including negative payoffs. We found that two related models performed well across both samples: the Fehr-Schmidt model and the Inequality & Joint-Gain model, which is in line with the results from previous work in healthy samples [78]. Both models include metrics for inequality and joint- or self-gain. When looking at disadvantageous inequality model parameter weights (Fehr-Schmidt model), we found no differences between the groups. This suggests that for both groups inequality and gain magnitude play an important role when evaluating joint payoffs and there seems to be no heightened inequality aversion in individuals with BPD compared to individuals in the control group. Additionally, we found that all participants in the BPD group and the majority of the HC group were categorized as inequality averse based on secondary SVO scores. This indicates that in both groups high sensitivity to inequality may be a central driver for prosocial behavior.

In summary, our findings indicate that both individuals with BPD and healthy controls exhibit inequality aversion and value fairness. Although these results were unexpected, they are consistent with prior research suggesting that individuals with clinically relevant BPD features may display heightened justice concerns not only for themselves but also for others [37]. Supporting this, Wischniewski and Brüne [42] found that while individuals with BPD punished unfair behavior in a modified DG to the same extent as controls, their motivation differed: whereas control participants were primarily driven by self-interest, the BPD group appeared motivated by angry retaliation, likely due to identifying with the victim's perspective. Our findings also align with those of previous studies reporting no differences between individuals with BPD and controls in fairness appraisals of trustees in TG paradigms [30, 40].

Strengths, limitations, and outlooks

In this study, we used a set of standard economic tasks to assess baseline levels of social preferences, active and reactive cooperation, fairness perception and inequality aversion in women with and without BPD. By deliberately avoiding additional experimental conditions, we sought to minimize confounding factors, which, however, entails limitations. We here discuss four specific limitations and make suggestions for future studies.

First, the lack of direct social interactions and variations in social contexts in our design may have limited the possibility to dissect more nuanced group differences. Participants made hypothetical choices or believed their partner was an anonymous participant. While anonymity in economic games can reveal underlying preferences by removing external influences like concerns about reputation or consequences [81], the absence of

direct interaction may have reduced the tasks' relevance, particularly for BPD-related behaviors. Including contextual factors like facial expressions or emotion regulation strategies could have made these tasks more salient. Nonetheless, we emphasize the importance of using standard economic games to establish baseline behavior before introducing task modifications to test specific questions [81]. Ideally, studies should start with baseline assessments using standard economic tasks, followed by specific alterations combined with self-reports and real social interactions. Future research should incorporate direct social interactions in repeated game paradigms to better simulate real-world social dynamics [3, 82] or test participants in naturalistic approaches with close others [83]. Moreover, the use of emotion induction procedures, as suggested by Jeung et al. [22], may help clarify how emotional states influence social decision-making in BPD. For instance, Masland and Hooley [84] found that negative emotional primes specifically influenced trustworthiness appraisals in individuals with higher levels of BPD symptoms. An especially important direction for future studies is the consideration of situational context. Prior research suggests that impairments in social behavior among individuals with BPD may be particularly pronounced following unexpected positive interactions or feedback [17, 30, 32, 33]. Differentiating responses to positive versus negative social contexts may be key to understanding the variability and triggers of social dysfunction in BPD. We specifically recommend that future studies adopt multi-method approaches—integrating self-report, behavioral, and physiological measures—to assess baseline behavior across diverse social contexts. For instance, the same economic game could be administered with different types of social partners—such as virtual agents, unfamiliar strangers, or close others—to investigate how relationship context shapes the association between maladaptive personality traits and reactive cooperation in individuals with BPD. Additionally, the impact of situational context could be explored by comparing behavior in these games following positive versus negative interaction scenarios [17].

Second, the use of one-shot games may not fully capture the complexities of repeated social interactions. Multi-round games in previous research have repeatedly revealed alterations in individuals with BPD [25, 26, 79], especially in reactive cooperation, which may emerge over time. For example, Abramov et al. [30] emphasized the importance of capturing the dynamic nature of social behavior by using a multi-round TG. Their findings revealed paradoxical patterns among individuals with higher BPD trait counts, whose behavior shifted across different phases of the game. These results highlight the need to investigate how trust and cooperation develop over time—an approach that may be particularly crucial for understanding reactive forms of social behavior.

Third, our study did not incorporate personality measures based on the HEXACO model [85], particularly the honesty-humility and agreeableness dimensions, which have been shown to play distinct roles in active and reactive cooperation [28]. Active cooperation—which involves non-exploitation and a willingness to share or give, is tied to honesty-humility, while reactive cooperation—which refers to non-retaliation and the capacity to forgive after unfair treatment—is associated with HEXACO agreeableness [35, 86]. Borderline personality features are related to deficits in reactive cooperation due to low levels of HEXACO agreeableness, yet these features do not appear to impair active cooperation [28, 34]. Although the HEXACO and Big Five models share some conceptual overlap, they differ in key aspects: HEXACO honesty-humility captures traits such as fairness, sincerity, and greed-avoidance, whereas HEXACO agreeableness emphasizes patience, forgiveness, and non-retaliation [85]. In comparison, Big Five agreeableness reflects trust, altruism, and compliance [87]. Given these distinctions, they do not represent the same underlying personality constructs and cannot be used interchangeably [88]. Future research should therefore incorporate HEXACO personality assessments when examining cooperative behavior of individuals with BPD under varying situational affordances [44]. This line of research further underscores the value of adopting dimensional approaches for understanding psychopathology. In our study, we aimed to address this by including a measure of maladaptive personality traits. We advocate for the continued integration of dimensional approaches alongside traditional categorical models, as this combination allows for a more comprehensive understanding of impairments and behavioral patterns while remaining applicable in clinical settings—particularly in the context of personality disorders [89]. Dimensional frameworks, such as the Research Domain Criteria (RDoC; [90]) initiative, advocate for investigating mental health through core functional domains rather than relying solely on categorical diagnostic labels. This shift enables a more nuanced understanding of individual differences and the underlying mechanisms that contribute to social dysfunction in BPD.

Lastly, our sample is limited in terms of gender and size. We only tested female participants as many studies on BPD [22] and our sample size may have reduced the statistical power to detect subtle group differences. Nonetheless, we want to stress that our study has a relatively large and well-characterized sample of BPD patients, matched with controls on age and intelligence. Achieving adequate sample sizes remains a broader challenge in psychological and clinical research, where sample

sizes are often determined by convention rather than power analyses [91]. This issue is further complicated by the high behavioral variability in BPD populations and the potentially small effect sizes observed in low-emotion economic tasks. We underscore the importance of reporting effect sizes, justifying sample sizes [92], and leveraging existing datasets. In our study, we addressed this challenge by integrating existing data and performing a mini meta-analysis for the UG. We advocate for the use of similar strategies—drawing on available datasets and using meta-analytic procedures to summarize comparable studies [52].

Despite these limitations, our study provides valuable insights and underscores the importance of further research with robust and transparent methodologies to better understand social behavior in individuals with BPD. Replications with larger, more diverse samples will be essential to confirm and extend our findings. Looking ahead, preregistration of hypotheses—as encouraged by open science platforms such as the Open Science Framework [93] or AsPredicted [94]–can help promoting more reliable results. In line with these recommendations, we preregistered our study and have made our behavioral data and analysis code openly available. Reporting null results and sharing data and code are critical steps toward enhancing transparency and reproducibility. These practices not only strengthen the credibility of individual studies but also contribute to addressing the broader replication crisis in clinical psychology and psychiatry [95]. Broader adoption of open science initiatives and active participation in collaborative consortia—such as the German Reproducibility Network [96], or the German Center for Mental Health (DZPG; [97])—can facilitate open large-scale, multi-site studies that enhance the generalizability and impact of clinical research.

Summary/conclusion

Our study did not find significant differences in active or reactive cooperation between individuals with BPD and healthy controls. These results suggest that, at least in the context of anonymous one-shot games, cooperation seems not to differ between those with BPD and healthy individuals. This underscores the importance of considering fine-grained contextual factors of social interactions for understanding cooperative decision-making in BPD.

Understanding the mechanisms of social dysfunction in BPD is crucial for developing therapeutic interventions that can help individuals form stable and successful relationships. By exploring the specific contexts and underlying reasons for these dysfunctional interactions, we can better tailor interventions to address these challenges, ultimately improving social functioning and quality of life for those with BPD.

In conclusion, our study provides valuable insights into the social decision-making processes of individuals with BPD. We emphasize the role of context and the need for replication. Further research is needed to refine these findings and to develop targeted interventions that can help individuals with BPD improve their social interactions and relationships.

Abbreviations

BFs	Bayes Factors
BP	Borderline Personality
BPD	Borderline Personality Disorder
DG	Dictator Game
GLMM	Generalized Linear Mixed Model
HC	Healthy Control
IPDE	International Personality Disorder Examination
JPE	Joint Payoff Evaluation
Mini-DIPS	short version of the Diagnostic Interview for Mental Disorders
Mini-q	three-minute speeded reasoning test
MPL	Maximum Pseudo Likelihood
OR	Odds Ratios
PiCD	Personality Inventory for ICD-11
SVO	Social Value Orientation
TG	Trust Game
UG	Ultimatum Game

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40479-025-00295-2.

Supplementary Material 1.

Acknowledgements

We thank Clara Beitz, Clara Heilmann, Inga von Kortzfleisch, and Sita Sielaff for help with data acquisition, Marie-Luise Otte for advising on the diagnostic interviews, and Benjamin Kuper-Smith for helpful discussions. For the publication fee we acknowledge financial support by Heidelberg University.

Authors' contributions

LMD and CWK designed the study and wrote the original draft. LMD collected the data and performed the data analyses. CWK supervised the project, realized the funding for the project, and helped with data analyses. JL collected the data and wrote parts of the original draft. CN and SCH helped with the study design, provided essential support with the clinical sample, and revised the manuscript. All authors reviewed and edited the manuscript and gave final approval for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL.
This project was supported by the Emmy Noether Research Group grant (392443797) from the German Research Foundation (DFG). Furthermore, this project was endorsed by the German Center for Mental Health (DZPG).

Data availability

The data and code for this study are publicly available on Github (https://github.com/dnhi-lab/CooperativeDecisionsBPD_2025) and Zenodo (https://zenodo.org/records/15322897).

Declarations

Ethics approval and consent to participate

The project was approved by the local ethics committee (Faculty of Medicine, Heidelberg University, S-137/2020). All participants were briefed on the study procedure and gave written informed consent before participating.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 January 2025 / Accepted: 22 April 2025
Published online: 17 June 2025

References

1. Gunderson JG, Herpertz SC, Skodol AE, Torgersen S, Zanarini MC. Borderline personality disorder. Nat Rev Dis Primer. 2018;4(1):18029. https://doi.org/10.1038/nrdp.2018.29.
2. Klein MH, Wonderlich SA, Crosby R. Self-Concept correlates of the personality disorders. J Personal Disord. 2001;15(2):150–6. https://doi.org/10.1521/pedi.15.2.150.19214.
3. Korn CW, La Rosée L, Heekeren HR, Roepke S. Social feedback processing in borderline personality disorder. Psychol Med. 2016;46(3):575–87. https://doi.org/10.1017/S003329171500207X.
4. Vater A, Schröder-Abé M, Weißgerber S, Roepke S, Schütz A. Self-concept structure and borderline personality disorder: evidence for negative compartmentalization. J Behav Ther Exp Psychiatry. 2015;46:50–8. https://doi.org/10.1016/j.jbtep.2014.08.003.
5. Baer RA, Peters JR, Eisenlohr-Moul TA, Geiger PJ, Sauer SE. Emotion-related cognitive processes in borderline personality disorder: a review of the empirical literature. Clin Psychol Rev. 2012;32(5):359–69. https://doi.org/10.1016/j.cpr.2012.03.002.
6. Giesen-Bloo J, Arntz A. World assumptions and the role of trauma in borderline personality disorder. J Behav Ther Exp Psychiatry. 2005;36(3):197–208. https://doi.org/10.1016/j.jbtep.2005.05.003.
7. 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(5):359–65. https://doi.org/10.1016/j.brat.2009.02.003.
8. Masland SR, Schnell SE, Shah TV. Trust beliefs, biases, and behaviors in borderline personality disorder: empirical findings and relevance to epistemic trust. Curr Behav Neurosci Rep. 2020;7(4):239–49. https://doi.org/10.1007/s40473-020-00220-7.
9. Preti E, Richetin J, Poggi A, Fertuck E. A model of trust processes in borderline personality disorder: a systematic review. Curr Psychiatry Rep. 2023;25(11):555–67. https://doi.org/10.1007/s11920-023-01468-y.
10. Lévay EE, Bajzát B, Unoka ZS. Expectation of selfishness from others in borderline personality disorder. Front Psychol. 2021;12:702227. https://doi.org/10.3389/fpsyg.2021.702227.
11. Cavicchioli M, Maffei C. Rejection sensitivity in borderline personality disorder and the cognitive–affective personality system: a meta-analytic review. Personal Disord Theory Res Treat. 2020;11(1):1–12. https://doi.org/10.1037/per0000359.
12. Poggi A, Richetin J, Preti E. Trust and rejection sensitivity in personality disorders. Curr Psychiatry Rep. 2019;21(8):69. https://doi.org/10.1007/s11920-019-1059-3.
13. Renneberg B, Herm K, Hahn A, Staebler K, Lammers C, Roepke S. Perception of social participation in borderline personality disorder. Clin Psychol Psychother. 2012;19(6):473–80. https://doi.org/10.1002/cpp.772.
14. Staebler K, Helbing E, Rosenbach C, Renneberg B. Rejection sensitivity and borderline personality disorder. Clin Psychol Psychother. 2011;18(4):275–83. https://doi.org/10.1002/cpp.705.
15. De Panfilis C, Riva P, Preti E, Cabrino C, Marchesi C. When social inclusion is not enough: implicit expectations of extreme inclusion in borderline personality disorder. Personal Disord Theory Res Treat. 2015;6(4):301–9. https://doi.org/10.1037/per0000132.
16. Domsalla M, Koppe G, Niedtfeld I, Vollstädt-Klein S, Schmahl C, Bohus M, et al. Cerebral processing of social rejection in patients with borderline personality disorder. Soc Cogn Affect Neurosci. 2014;9(11):1789–97.https://doi.org/10.1093/scan/nst176.
17. Liebke L, Koppe G, Bungert M, Thome J, Hauschild S, Defiebre N, et al. Difficulties with being socially accepted: an experimental study in borderline personality disorder. J Abnorm Psychol. 2018;127(7):670–82. https://doi.org/10.1037/abn0000373.
18. Staebler K, Renneberg B, Stopsack M, Fiedler P, Weiler M, Roepke S. Facial emotional expression in reaction to social exclusion in borderline personality disorder. Psychol Med. 2011;41(9):1929–38. https://doi.org/10.1017/S0033291711000080.
19. Bouchard S, Sabourin S. Borderline personality disorder and couple dysfunctions. Curr Psychiatry Rep. 2009;11(1):55–62. https://doi.org/10.1007/s11920-009-0009-x.
20. Lazarus SA, Cheavens JS, Festa F, Zachary Rosenthal M. Interpersonal functioning in borderline personality disorder: a systematic review of behavioral and laboratory-based assessments. Clin Psychol Rev. 2014;34(3):193–205. https://doi.org/10.1016/j.cpr.2014.01.007.
21. Stepp SD, Pilkonis PA, Yaggi KE, Morse JQ, Feske U. Interpersonal and emotional experiences of social interactions in borderline personality disorder. J Nerv Ment Dis. 2009;197(7):484–91. https://doi.org/10.1097/NMD.0b013e3181aad2e7.
22. Jeung H, Schwieren C, Herpertz SC. Rationality and self-interest as economic-exchange strategy in borderline personality disorder: game theory, social preferences, and interpersonal behavior. Neurosci Biobehav Rev. 2016;71:849–64. https://doi.org/10.1016/j.neubiorev.2016.10.030.
23. Lis S, Kirsch P. Neuroeconomic Approaches in Mental Disorders. In: Reuter M, Montag C, editor. Neuroeconomics. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p 311–30. (Studies in Neuroscience, Psychology and Behavioral Economics). [cited 2025 Apr 2]. Available from: https://doi.org/10.1007/978-3-642-35923-1_16.
24. 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(2):e1005935. https://doi.org/10.1371/journal.pcbi.1005935.
25. 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(5890):806–10. https://doi.org/10.1126/science.1156902.
26. Unoka Z, Seres I, Áspán N, Bódi N, Kéri S. Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. J Personal Disord. 2009;23(4):399–409. https://doi.org/10.1521/pedi.2009.23.4.399.
27. Saunders KEA, Goodwin GM, Rogers RD. Borderline personality disorder, but not euthymic bipolar disorder, is associated with a failure to sustain reciprocal cooperative behaviour: implications for spectrum models of mood disorders. Psychol Med. 2015;45(8):1591–600. https://doi.org/10.1017/S0033291714002475.
28. Thielmann I, Hilbig BE, Niedtfeld I. Willing to give but not to forgive: borderline personality features and cooperative behavior. J Personal Disord. 2014;28(6):778–95. https://doi.org/10.1521/pedi_2014_28_135.
29. Graumann L, Kulakova E, Cho AB, Deuter CE, Wolf OT, Schell J, et al. Elevated testosterone and prosocial behavior in female patients with borderline personality disorder independent of social exclusion. Psychoneuroendocrinology. 2025;171:107232. https://doi.org/10.1016/j.psyneuen.2024.107232.
30. Abramov G, Miellet S, Kautz J, Grenyer BFS, Deane FP. The Paradoxical decline and growth of trust as a function of borderline personality disorder trait count: using discontinuous growth modelling to examine trust dynamics in response to violation and repair. PLoS ONE. 2020:15(7);e0236170. https://doi.org/10.1371/journal.pone.0236170.
31. Preuss N, Brändle LS, Hager OM, Haynes M, Fischbacher U, Hasler G. Inconsistency and social decision making in patients with borderline personality disorder. Psychiatry Res. 2016;243:115–22. https://doi.org/10.1016/j.psychres.2016.06.017.
32. De Panfilis C, Schito G, Generali I, Gozzi LA, Ossola P, Marchesi C, et al. Emotions at the border: increased punishment behavior during fair interpersonal exchanges in borderline personality disorder. J Abnorm Psychol. 2019;128(2):162–72. https://doi.org/10.1037/abn0000404.
33. Niedtfeld I, Kroneisen M. Impaired memory for cooperative interaction partners in borderline personality disorder. Borderline Personal Disord Emot Dysregulation. 2020;7(1):22. https://doi.org/10.1186/s40479-020-00137-3.
34. Hepp J, Hilbig BE, Moshagen M, Zettler I, Schmahl C, Niedtfeld I. Active versus reactive cooperativeness in borderline psychopathology: a dissection based on the HEXACO model of personality. Personal Individ Differ. 2014;56:19–23. https://doi.org/10.1016/j.paid.2013.08.013.
35. Hilbig BE, Zettler I, Leist F, Heydasch T. It takes two: Honesty–Humility and agreeableness differentially predict active versus reactive Cooperation. Personal Individ Differ. 2013;54(5):598–603. https://doi.org/10.1016/j.paid.2012.11.008.
36. Bateman AW, Krawitz R. Borderline Personality Disorder (DRAFT): An evidence-based guide for generalist mental health professionals. Oxford University Press; 2013. [cited 2025 Apr 2]. https://doi.org/10.1093/med:psych/9780199644209.001.0001. Available from: https://academic.oup.com/book/1354.

37. Lis S, Schaedler A, Liebke L, Hauschild S, Thome J, Schmahl C, et al. Borderline personality disorder features and sensitivity to injustice. J Personal Disord. 2018;32(2):192–206. https://doi.org/10.1521/pedi_2017_31_292.
38. Bondü R, Richter P. Linking forms and functions of aggression in adults to justice and rejection sensitivity. Psychol Violence. 2016;6(2):292–302. https://doi.org/10.1037/a0039200.
39. Fehr E, Schmidt KM. A theory of fairness, competition, and Cooperation. Q J Econ. 1999;114(3):817–68. https://doi.org/10.1162/003355399556151.
40. Franzen N, Hagenhoff M, Baer N, Schmidt A, Mier D, Sammer G, et al. Superior 'theory of Mind' in borderline personality disorder: an analysis of interaction behavior in a virtual trust game. Psychiatry Res. 2011;187(1–2):224–33. https://doi.org/10.1016/j.psychres.2010.11.012.
41. Jeung H, Vollmann M, Herpertz SC, Schwieren C. Consider others better than yourself: social decision-making and partner preference in borderline personality disorder. J Behav Ther Exp Psychiatry. 2020;67:101436. https://doi.org/10.1016/j.jbtep.2018.11.004.
42. Wischniewski J, Brüne M. How do people with borderline personality disorder respond to norm violations? Impact of personality factors on economic Decision-Making. J Personal Disord. 2013;27(4):531–46. https://doi.org/10.1521/pedi_2012_26_036.
43. Hepp J, Kieslich PJ, Schmitz M, Schmahl C, Niedtfeld I. Negativity on two sides: individuals with borderline personality disorder form negative first impressions of others and are perceived negatively by them. Personal Disord Theory Res Treat. 2021;12(6):514–25. https://doi.org/10.1037/per0000412.
44. Hepp J, Niedtfeld I. Prosociality in personality disorders: status quo and research agenda. Curr Opin Psychol. 2022;44:208–14. https://doi.org/10.1016/j.copsyc.2021.09.013.
45. Oltmanns JR, Widiger TA. A self-report measure for the ICD-11 dimensional trait model proposal: the personality inventory for ICD-11. Psychol Assess. 2018;30(2):154–69. https://doi.org/10.1037/pas0000459.
46. Murphy RO, Ackermann KA, Handgraaf M. Measuring Social Value Orientation. SSRN Electron J. 2011. [cited 2019 Jul 23]. https://doi.org/10.2139/ssrn.1804189. Available from: http://www.ssrn.com/abstract=1804189.
47. Forsythe R, Horowitz JL, Savin NE, Sefton M. Fairness in simple bargaining experiments. Games Econ Behav. 1994;6(3):347–69. https://doi.org/10.1006/game.1994.1021.
48. Kahneman D, Knetsch JL, Thaler RH. Fairness and the assumptions of economics. J Bus. 1986;59(4):S285–300.
49. Güth W, Schmittberger R, Schwarze B. An experimental analysis of ultimatum bargaining. J Econ Behav Organ. 1982;3(4):367–88. https://doi.org/10.1016/0167-2681(82)90011-7.
50. Damovsky F, Zettl M, Zimmermann J, Herbold W, Curtius T, Buecker S et al. Das Persönlichkeitsinventar für ICD-11: Reliabilität und Validität der deutschen Version in einer klinischen und nicht-klinischen Stichprobe. Open Science Framework; 2021. [cited 2022 May 17]. https://doi.org/10.31219/osf.io/cm3qv. Available from: https://osf.io/cm3qv.
51. Haruno M, Frith CD. Activity in the amygdala elicited by unfair divisions predicts social value orientation. Nat Neurosci. 2010;13(2):160–1. https://doi.org/10.1038/nn.2468.
52. Goh JX, Hall JA, Rosenthal R. Mini Meta-Analysis of your own studies: some arguments on why and a primer on how. Soc Personal Psychol Compass. 2016;10(10):535–49. https://doi.org/10.1111/spc3.12267.
53. Leiner DJ. SoSci Survey (Version 3.1.06). 2019. Available from: https://www.soscisurvey.de.
54. Margraf J, Cwik JC. Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen. 2017. [cited 2022 May 17]. https://doi.org/10.13154/RUB.102.91. Available from: https://omp.ub.rub.de/index.php/RUB/catalog/book/102.
55. Loranger AW. The international personality disorder examination: the world health organization/alcohol, drug abuse, and mental health administration international pilot study of personality disorders. Arch Gen Psychiatry. 1994;51(3):215. https://doi.org/10.1001/archpsyc.1994.03950030051005.
56. Loranger AW, Sartorius N, Andreoli A, Berger P, Buchheim P, Channabasavanna SM, Coid B, Dahl AA, Diekstra RFW, Ferguson B. Deutschsprachige Fassung der International personality disorder examination. IPDE. Genf: WHO; 1998.
57. Baudson TG, Preckel F. mini-q: intelligenzscreening in Intelligenzscreening in drei Minuten. Diagnostica. 2016;62(3):182–97. https://doi.org/10.1026/0012-1924/a000150.
58. Beck AT, Steer RA, Brown GK. BDI-II. Beck-Depressionsinventar [Beck Depression Inventory II]. Frankfurt: Pearson; 2006.
59. Franke GH. BSCL-53®-S. Brief symptom-checklist–standard–German manual. Göttingen: Hogrefe Publishing Group; 2017.
60. Bohus M, Kleindienst N, Limberger MF, Stieglitz RD, Domsalla M, Chapman AL, et al. The short version of the borderline symptom list (BSL-23): development and initial data on psychometric properties. Psychopathology. 2009;42(1):32–9. https://doi.org/10.1159/000173701.
61. Spitzer C, Müller S, Kerber A, Hutsebaut J, Brähler E, Zimmermann J. Die Deutsche version der level of personality functioning Scale-Brief form 2.0 (LPFS-BF): Faktorenstruktur, konvergente validität und Normwerte in der allgemeinbevölkerung. PPmP Psychother Psychosom Med Psychol. 2021;71(07):284–93. https://doi.org/10.1055/a-1343-2396.
62. Weekers LC, Hutsebaut J, Kamphuis JH. The level of personality functioning Scale-Brief form 2.0: update of a brief instrument for assessing level of personality functioning: the level of personality functioning Scale - Brief form 2.0. Personal Ment Health. 2019;13(1):3–14. https://doi.org/10.1002/pmh.1434.
63. Bach B, Kerber A, Aluja A, Bastiaens T, Keeley JW, Claes L, et al. International assessment of DSM-5 and ICD-11 personality disorder traits: toward a common nosology in DSM-5.1. Psychopathology. 2020;53(3-4):179–88. https://doi.org/10.1159/000507589.
64. Zimmermann J, Altenstein D, Krieger T, Holtforth MG, Pretsch J, Alexopoulos J, et al. The structure and correlates of self-reported DSM-5 maladaptive personality traits: findings from two German-Speaking samples. J Personal Disord. 2014;28(4):518–40. https://doi.org/10.1521/pedi_2014_28_130.
65. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-test instead of student's t-test. Int Rev Soc Psychol. 2017;30(1):92–101. https://doi.org/10.5334/irsp.82.
66. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/.
67. Hothorn T, Hornik K, Wiel MAVD, Zeileis A. Implementing a Class of Permutation Tests: The coin Package. J Stat Softw. 2008;28(8). [cited 2025 Jan 20]. https://doi.org/10.18637/jss.v028.i08. Available from: http://www.jstatsoft.org/v28/i08/.
68. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. 2019;0.7.2. [cited 2025 Jan 20]. Available from: https://CRAN.R-project.org/package=rstatix.
69. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org.
70. Tiedemann F. gghalves: Compose Half-Half Plots Using Your Favourite Geoms. 2019;0.1.4. [cited 2025 Jan 20]. Available from: https://CRAN.R-project.org/package=gghalves.
71. Morey RD, Rouder JN, BayesFactor. Computation of Bayes Factors for Common Designs. 2018. Available from: https://CRAN.R-project.org/package=BayesFactor.
72. Jeffreys H. The theory of probability. Oxford: OuP; 1998.
73. Keysers C, Gazzola V, Wagenmakers EJ. Using Bayes factor hypothesis testing in neuroscience to Establish evidence of absence. Nat Neurosci. 2020;23(7):788–99. https://doi.org/10.1038/s41593-020-0660-4.
74. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R. Cham: Springer International Publishing; 2015 [cited 2024 Oct 8]. (Use R!). Available from: https://link.springer.com/https://doi.org/10.1007/978-3-319-21416-0.
75. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-On Guide. 1st ed. Boca Raton: Chapman and Hall/CRC; 2021 [cited 2024 Oct 8]. Available from: https://www.taylorfrancis.com/books/9781003107347.
76. Hepp J, Hilbig BE, Kieslich PJ, Herzog J, Lis S, Schmahl C, et al. Borderline personality and the detection of angry faces. PLoS One. 2016;11:e0152947. https://doi.org/10.1371/journal.pone.0152947.
77. Wetzels R, Wagenmakers EJ. A default bayesian hypothesis test for correlations and partial correlations. Psychon Bull Rev. 2012;19(6):1057–64. https://doi.org/10.3758/s13423-012-0295-x.
78. Doppelhofer LM, Hurlemann R, Bach DR, Korn CW. Social motives in a patient with bilateral selective amygdala lesions: shift in prosocial motivation but not in social value orientation. Neuropsychologia. 2021;162:108016. https://doi.org/10.1016/j.neuropsychologia.2021.108016.
79. Polgár P, Fogd D, Unoka Z, Király E, Csukly G. Altered social decision making in borderline personality disorder: an ultimatum game study. J Personal Disord. 2014;28(6):841–52. https://doi.org/10.1521/pedi_2014_28_142.
80. Weinbrecht A, Niedeggen M, Roepke S, Renneberg B. Feeling excluded no matter what? Bias in the processing of social participation in borderline personality disorder. NeuroImage Clin. 2018;19:343–50. https://doi.org/10.1016/j.nicl.2018.04.031.

81. Pisor AC, Gervais MM, Purzycki BG, Ross CT. Preferences and constraints: the value of economic games for studying human behaviour. R Soc Open Sci. 2020;7(6):192090. https://doi.org/10.1098/rsos.192090.
82. Ruocco AC, Medaglia JD, Tinker JR, Ayaz H, Forman EM, Newman CF, et al. Medial prefrontal cortex hyperactivation during social exclusion in borderline personality disorder. Psychiatry Res Neuroimaging. 2010;181(3):233–6. https://doi.org/10.1016/j.pscychresns.2009.12.001.
83. Miano A, Fertuck EA, Roepke S, Dziobek I. Romantic relationship dysfunction in borderline personality disorder—a naturalistic approach to trustworthiness perception. Personal Disord Theory Res Treat. 2017;8(3):281–6. https://doi.org/10.1037/per0000196.
84. Masland SR, Hooley JM. When trust does not come easily: negative emotional information unduly influences trustworthiness appraisals for individuals with borderline personality features. J Personal Disord. 2020;34(3):394–409. https://doi.org/10.1521/pedi_2019_33_404.
85. Ashton MC, Lee K. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personal Soc Psychol Rev. 2007;11(2):150–66. https://doi.org/10.1177/1088868306294907.
86. Hilbig BE, Glöckner A, Zettler I. Personality and prosocial behavior: linking basic traits and social value orientations. J Pers Soc Psychol. 2014;107(3):529–39. https://doi.org/10.1037/a0036074.
87. Goldberg LR. An alternative "description of personality": the Big-Five factor structure. J Pers Soc Psychol. 1990;59(6):1216–29. https://doi.org/10.1037/0022-3514.59.6.1216.
88. Thielmann I, Moshagen M, Hilbig BE, Zettler I. On the comparability of basic personality models: Meta-Analytic correspondence, scope, and orthogonality of the big five and HEXACO dimensions. Eur J Personal. 2022;36(6):870–900. https://doi.org/10.1177/08902070211026793.
89. Herpertz SC, Huprich SK, Bohus M, Chanen A, Goodman M, Mehlum L, et al. The challenge of transforming the diagnostic system of personality disorders. J Personal Disord. 2017;31(5):577–89. https://doi.org/10.1521/pedi_2017_31_338.
90. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010;167(7):748–51. https://doi.org/10.1176/appi.ajp.2010.09091379.
91. Bakker M, Hartgerink CHJ, Wicherts JM, Van Der Maas HLJ. Researchers' intuitions about power in psychological research. Psychol Sci. 2016;27(8):1069–77. https://doi.org/10.1177/0956797616647519.
92. Lakens D. Sample size justification. Collabra Psychol. 2022;8(1):33267. https://doi.org/10.1525/collabra.33267.
93. Center for Open Science. OSF Home [Internet]. Charlottesville (VA): Center for Open Science; [cited 2025 Apr 4]. Available from: https://osf.io/.
94. Wharton Credibility Lab. AsPredicted [Internet]. Philadelphia (PA): University of Pennsylvania; [cited 2025 Apr 4]. Available from: https://aspredicted.org/.
95. Tackett JL, Brandes CM, King KM, Markon KE. Psychology's replication crisis and clinical psychological science. Annu Rev Clin Psychol. 2019;15(1):579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710.
96. German Reproducibility Network. German Reproducibility Network (GRN) [Internet]. Munich: German Reproducibility Network; [cited 2025 Apr 4]. Available from: https://reproducibilitynetwork.de/.
97. Meyer-Lindenberg A, Falkai P, Fallgatter AJ, Hannig R, Lipinski S, Schneider S, et al. The future German Center for Mental Health (Deutsches Zentrum für Psychische Gesundheit): a model for the co-creation of a national translational research structure. 1(3):153–6. Available from: https://doi.org/10.1038/s44220-023-00026-y.

Publisher's Note

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

6 Study IV: Shift in prosocial motivation but not in social value orientation

Social motives in a patient with bilateral selective amygdala lesions: Shift in prosocial motivation but not in social value orientation. *Neuropsychologia*, *162*, 108016 (2021)

Lisa M. Doppelhofer, René Hurlemann, Dominik R. Bach, Christoph W. Korn

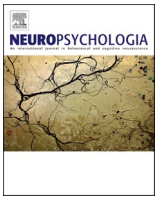
Neuropsychologia 162 (2021) 108016



Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia





Social motives in a patient with bilateral selective amygdala lesions: Shift in prosocial motivation but not in social value orientation

Lisa M. Doppelhofer [a,b,*], René Hurlemann [c,d], Dominik R. Bach [e,f,g], Christoph W. Korn [a,b,c,d,**]

[a] Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, 20246, Hamburg, Germany
[b] Department of General Adult Psychiatry, Heidelberg University, 69115, Heidelberg, Germany
[c] Department of Psychiatry and Division of Medical Psychology, University of Bonn, 53012, Bonn, Germany
[d] Department of Psychiatry, University of Oldenburg Medical Campus, 26160, Bad Zwischenahn, Germany
[e] Department of Psychiatry, Psychotherapy, and Psychosomatics, University of Zurich, Zurich, Switzerland
[f] Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, WC1N 3BG, United Kingdom
[g] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, WC1B 5EH, United Kingdom

ARTICLE INFO

Keywords:
Amygdala lesion
Prosociality
Allocation preferences
Social cognition
Decision-making

ABSTRACT

Humans hold social motives that are expressed in social preferences and influence how they evaluate and share payoffs. Established models in psychology and economics quantify social preferences such as general social value orientation, which captures people's tendency to be prosocial or individualistic. Prosocials further differ by how much they maximize joint gains or minimize inequality. Functional neuroimaging studies have linked increased amygdala activity in prosocials to payoff inequality between self and other. However, it is unclear whether amygdala lesions alter social motives. We used two tasks to test a patient with selective bilateral amygdala lesions and three healthy samples (a priori matched control sample $N = 20$, online sample $N = 603$, student sample $N = 40$), which allowed us to assess and model social motives across a relatively large number of participants. In a social value orientation task, the patient was categorized as prosocial and her social value orientation score did not differ from healthy participants. Importantly, the patient differed in prosocial motivation by maximizing joint gains rather than minimizing payoff inequality. In a joint payoff evaluation task, Bayesian model comparisons revealed that participants' evaluations were best described by models, which link participants' evaluations to the payoff magnitude and to inequality. Overall, amygdala lesions did not seem to alter general social value orientation but shifted prosocial motivation toward maximizing joint gains.

1. Introduction

The functioning of social groups builds upon individuals cooperating and sharing payoffs with each other. Humans hold social motives that are expressed in their social preferences and influence how they evaluate and share payoffs with each other (MacCrimmon and Messick, 1976). Psychological and economic models assume that individuals' preferences for payoff allocations depend—in addition to self-gains—on several social motives such as maximizing joint gains and minimizing inequality (Fehr and Schmidt, 1999; Murphy et al., 2011). Functional neuroimaging studies have linked individual differences in amygdala responses to social motives during the evaluation of allocations (Haruno and Frith, 2010; Liu et al., 2019). Here, we investigated whether amygdala lesions have an impact on social preferences and motives in the context of sharing monetary payoffs between oneself and an anonymous other by testing a patient with selective bilateral amygdala lesions and comparing her to an a priori matched control sample.

Social Value Orientation (SVO) refers to how individuals evaluate and allocate payoffs for themselves and others (Messick and McClintock, 1968). General SVO can be measured on a continuous scale that captures the amount of a common payoff a person is willing to sacrifice for another person's payoff (Murphy et al., 2011). To facilitate the interpretability of the continuous SVO scores, people can be classified into at least four different SVO categories that partition the possible range of SVO scores. Empirically, most people are prosocials (maximizing own and others' payoffs and minimizing inequality) and many are individualists (maximizing own payoffs). Fewer people are categorized as competitive (maximizing the difference between own and others'

* Corresponding author. Department of General Adult Psychiatry, Heidelberg University, 69115, Heidelberg, Germany.
** Corresponding author. Department of General Adult Psychiatry, Heidelberg University, 69115, Heidelberg, Germany.
E-mail addresses: LisaMarie.Doppelhofer@med.uni-heidelberg.de (L.M. Doppelhofer), Christoph.Korn@med.uni-heidelberg.de (C.W. Korn).

https://doi.org/10.1016/j.neuropsychologia.2021.108016
Received 13 November 2020; Received in revised form 4 August 2021; Accepted 26 August 2021
Available online 7 September 2021
0028-3932/© 2021 Elsevier Ltd. All rights reserved.

payoffs) or altruistic (maximizing others' payoffs) (Murphy et al., 2011). With additional SVO items, prosocials can be further positioned on two continua: One for joint gain maximization (maximizing the sum of one's own and another person's payoffs) and another for inequality aversion (minimizing the difference between one's own and another person's payoffs). These two continua are typically compiled into one so-called prosocial motivation score.

Distinct SVO categories have repeatedly been associated with differences in cooperative behavior. In laboratory experiments, prosocials contribute more to group interests than individualistic and competitive persons (e.g. De Cremer and Van Lange, 2001). In real life, prosocials engage more in behavior that facilitates long-term collective interests, such as pro-environmental actions (Joireman et al., 2001) or volunteering (Van Lange et al., 2011).

Several functional neuroimaging studies in humans suggest that individual differences in SVO relate to differences in amygdala activity when allocation options are presented (Haruno et al., 2014; Haruno and Frith, 2010; Liu et al., 2019; Tanaka et al., 2017). Specifically, prosocials and individualists differed in their dorsal amygdala activity related to the absolute difference between payoffs for the self and for another person—a measure of overall inequality (Haruno and Frith, 2010). That is, while amygdala activity was positively correlated with absolute payoff differences in prosocials during payoff presentation, it was slightly negatively correlated with absolute payoff differences in individualists (Haruno and Frith, 2010). A recent functional neuroimaging study refined this picture (Liu et al., 2019): Greater deviations of the currently presented allocation from the individually preferred allocation elicited greater amygdala activity on a trial-by-trial basis. In addition, amygdala activity related to deviations from the preferred allocation positively correlated with inequality aversion in prosocials but not individualists (Liu et al., 2019).

Animal studies provide further evidence for the importance of the amygdala for prosocial preferences. Electrophysiological measurements in rhesus macaques revealed enhanced synchronization between the basolateral amygdala and the rostral anterior cingulate gyrus for prosocial decisions in contrast to suppressed synchronizations in antisocial

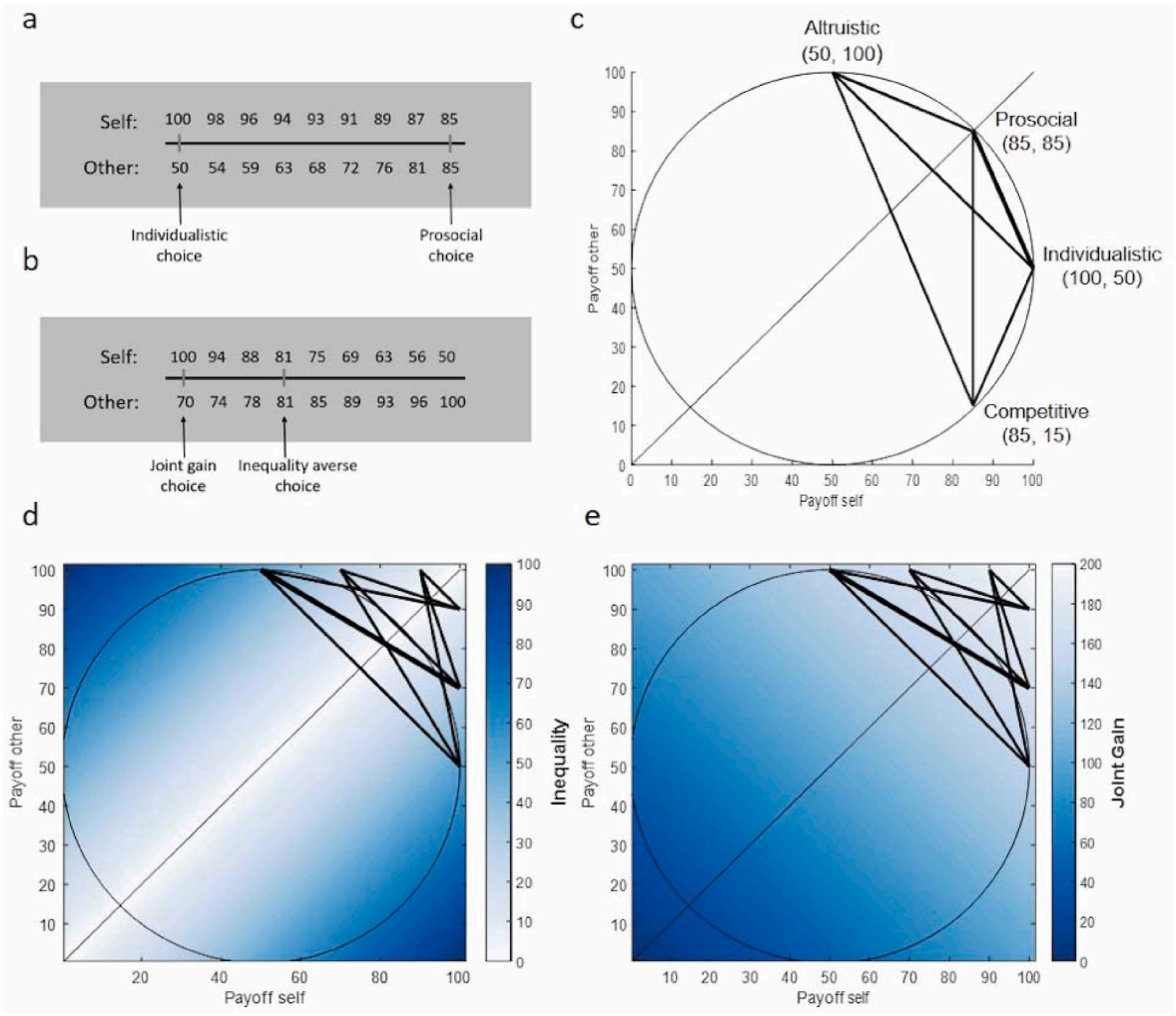


Fig. 1. SVO task and allocation plane for primary and secondary items. **a.** Example of one of the six primary SVO slider items. **b.** Example of one of the nine secondary SVO slider items. **c.** Allocation plane for the six primary items with payoffs for the self on the x-axis and payoffs for the other on the y-axis. The circle delimits the endpoints of the primary items. Black lines indicate payoffs for the nine options of each primary item. The item depicted in (a) corresponds to the bold line ranging from (100, 50) to (85, 85). The locations of the four idealized SVO categories (altruistic, prosocial, individualistic, and competitive) are mapped. **d. & e.** Allocation plane for the nine secondary items with payoffs for the self on the x-axis and payoffs for the other on the y-axis. Black lines indicate payoffs for the nine options of each secondary item. The item depicted in (b) corresponds to the bold line ranging from (100, 70) to (50, 100). **d.** The color map illustrates the gradient of inequality between payoffs for self and other across the allocation plane. Absolute inequality is lowest on the identity line. **e.** The same nine secondary items as in (d) are depicted but here the color map illustrates the gradient of joint gain (joint payoffs) across the allocation plane. Joint gain is highest at the rightmost and uppermost point of the allocation plane. See Supplementary Fig. 1 for numerical details on an exemplary item. See Supplementary Fig. 2 for the depiction of the SVO and JPE items in terms of (absolute) inequality and joint gain. See Supplementary Figs. 3 & 4 for choice distributions across the SVO items.

decisions (Dal Monte et al., 2020). Bilateral lesions in the basolateral amygdala in rats resulted in lower preferences for mutual rewards, indicating that the integrity of the basolateral amygdala facilitates prosocial preferences (Hernandez-Lallement et al., 2016).

Thus, human and animal studies point toward a role of the amygdala for social motives but it is unclear if amygdala lesions influence prosocial preferences in humans. Here, our primary aim was to investigate the impact of selective bilateral amygdala lesions on allocation decisions and preferences by comparing a patient with amygdala lesions to a group of a priori age- and gender-matched healthy individuals. In secondary, exploratory analyses, we compared the patient to a larger group of participants that was matched post hoc. Based on the neuroimaging findings that show relationships between amygdala activity and prosocial preferences (Haruno and Frith, 2010; Liu et al., 2019), we hypothesized a less prosocial SVO (i.e., less prosocial payoff allocations) in a patient with Urbach-Wiethe (UW) syndrome (a disease characterized by amygdala calcification) than in controls. In addition, we tested how well different social motives—such as inequality aversion—describe participants' evaluations of given payoff allocations, by comparing eight different economic models.

We adopted the same tasks as Haruno and Frith (2010) and used two complementary approaches to assess social motives. First, the SVO task provides an individual characterization of social preferences by having participants choose between different payoff options for the self and another person. The SVO task is similar to so-called "dictator games" (Engel, 2011) in the sense that one person single-handedly determines how allocations should be split between the self and another person. This other person has no say in the decision. Primary items (Fig. 1a & d) in the SVO task assess general SVO and secondary items (Fig. 1b & d-e) specify the type of prosocial motivation. Second, in the joint payoff evaluation (JPE) task, participants evaluate the desirability for a series of given joint payoff pairs. We extended the modeling approach of Haruno and Frith (2010) and compared how well eight different economic models describe participants' evaluations in the JPE task. We assessed three healthy samples to provide distributions of SVO and to identify the model that best describes participants' evaluations of monetary allocations. Specifically, we tested one control sample for the UW patient. We also tested two additional samples to characterize the behavior of a wider population. That is, our primary aim was to compare the UW patient with an a priori matched control group. Our secondary aims were to compare the patient to a larger post hoc matched sample and to provide data on social motives of a relatively large number of participants.

2. Methods

2.1. Participants

2.1.1. UW patient

One of two female monozygotic twins (age 43 years) with Urbach-Wiethe (UW) syndrome—a congenital disorder leading to amygdala calcification (Newton et al., 1971)—was tested at the University of Bonn. Both patients have been characterized neurologically and behaviorally in earlier studies (Hurlemann et al., 2007; Talmi et al., 2010). In this study, the patient previously called BG or patient 2 was tested. Her twin sister was not tested because she suffered from a psychiatric condition (Scheele et al., 2019). Below, we list several earlier findings obtained from the same lesion patient. The patient has an average intelligence but exhibits impairments in phonemic fluency and short-term concentration (Talmi et al., 2010). She shows impairments in the recognition of fearful facial expressions (Becker et al., 2012; Hurlemann et al., 2007) and in the prioritization of threatening facial expressions (Bach et al., 2015). She has aberrations in startle responses during fear-eliciting scenes, in threat processing (Becker et al., 2012). In approach-avoidance conflict tasks, she shows reduced loss adaptation (Korn et al., 2017) and decreased vigor of escape from threat (Bach et al., 2019). Furthermore, she has a smaller and less complex social network (Becker et al., 2012).

2.1.2. Healthy samples

In total, we acquired three separate healthy samples (for which we specify details in the next sections): First, an age- and gender-matched control sample was tested in the laboratory. Second, we acquired an online sample that allowed us to compare the UW patient to a rather large subgroup of this online sample that was matched post hoc for secondary, exploratory analyses. Third, a student sample was tested in the laboratory. For interested readers, we report the behavioral results from the full online sample and the student sample, but we did not compare the UW patient to these two samples. Inclusion criteria for all healthy samples were age between 18 and 60 and German mother tongue. Additional exclusion criteria for the laboratory samples were self-reported psychiatric or neurological diagnoses or drug abuse. In all samples and for all analyses, we excluded participants with intransitive answers on the SVO measure (see the paragraph on SVO scores below). Following specific methodological conventions and considerations, further participants needed to be excluded for specific analyses (see below and Supplementary Table 1 for details).

(1) An age- and gender-matched control sample (original $N = 23$; after exclusion for intransitive answers $N = 20$, age: $M = 41.60$, $SD = 3.91$) was recruited from the general population and tested at the Psychiatric University Hospital Zurich.
(2) A more diverse online sample (original $N = 628$; after exclusion for intransitive answers $N = 603$, 377 female, 4 participants chose not to disclose their gender; age: $M = 37.70$, $SD = 11.64$) was recruited via SoSci Panel, which is associated to a German online survey system called SoSci Survey (Leiner, 2019). The administrators of the SoSci Panel offered the opportunity to invite participants who had signed up for the SoSci panel out of interest and receive no monetary reimbursement. Participants of the online sample were tested via SoSci Survey and completed a series of unrelated tasks (described in Korn et al., 2018). This allowed us to compare the UW patient to a rather large matched subgroup of this online sample by choosing—post hoc—female participants within the age range of 36–49 years (to match the age range of the control sample) ($N = 91$, age: $M = 41.99$, $SD = 3.60$). We label this subsample "online subgroup."
(3) To provide more evidence that the general findings in the two tasks hold, we additionally tested a student sample (original $N = 41$; after exclusion for intransitive answers $N = 40$, 26 female; age: $M = 24.35$, $SD = 3.81$) at the University Medical Center Hamburg-Eppendorf.

2.1.3. Self-reported income range

The UW patient did not differ from the control sample and the matched subgroup of the online sample with respect to self-reported income ranges. Participants were asked to indicate their monthly income range on a scale with the lowest option (no personal income) and the highest option (above 6500 CHF or above 4000 €; in the online sample, participants also had the option to not indicate their income range). Income ranges for the two currencies were converted using current information about Swiss and German purchasing power parities acquired from the Federal Statistical Office of Switzerland (BFS, 2018). A Bayesian single-case analysis (for details see methods section) revealed that an estimated 88.77 % (95 % CI [74.28, 97.23], $p = .225$, two-tailed) of the control sample and 63.23 % (95 % CI [54.90, 71.14], $p = .735$, two-tailed) of the online subgroup would have a lower income than the UW patient, indicating no significant income differences between the UW patient and the healthy controls. An exploratory analysis revealed no association between income level and primary SVO scores (Spearman's $r = -.05$, $p = .836$, $N = 20$) in the control group. Furthermore, income level and prosocial motivation scores were not

significantly correlated (Spearman's $r = -0.09$, $p = .758$, $N = 15$) in the prosocial participants of the control group (secondary items are only interpretable in prosocial participants; see methods section below for details). We found no significant correlation between income level and primary SVO scores (Spearman's $r = -.02$, $p = .857$, $N = 91$) in prosocial participants of the matched subgroup of the online sample. Likewise, we found no association between income level and the prosocial motivation scores (Spearman's $r = .05$, $p = .699$, $N = 66$) in prosocial participants of the online subgroup.

2.1.4. Reimbursement and ethics

The laboratory samples (i.e., the UW patient, the control, and student samples) received a show-up fee plus a variable amount depending on their own and on another participants' decisions in the SVO task. Participants in the online sample participated out of interest and did not receive monetary reimbursement. Therefore, participants of the online sample made hypothetical choices. Participants in the laboratory gave written informed consent prior to data collection and the study (including the form of taking consent) was approved by the local ethics committees. As mentioned above, the online sample was recruited via the SoSci Panel. These participants had voluntarily signed up to receive invitations to online studies and chose freely to take part in our study. They were informed about the study similar to the participants in the laboratory and could freely stop taking part in the online study. Due to the online format, participants did not give explicit informed consent in written form. Before conducting the online experiment, we consulted the ethics committee in Zurich: At the time point of testing, online studies were deemed exempt if they used this type of questionnaire stimuli and if they only collected data from participants who remained anonymous (i.e., we collected no contact information at all and had never intended to do so). We only included participants who completed the relevant sections (i.e. demographic questions, SVO and JPE task) of the online study.

2.2. Tasks

All participants performed the SVO Slider Measurement (Murphy et al., 2011) and a joint payoff evaluation (JPE) task (Haruno and Frith, 2010). Moreover, the UW patient and the control and student samples completed a series of unrelated tasks (Bach et al., 2019; the remaining tasks will be reported elsewhere). Data of the laboratory samples were acquired using the MATLAB (Mathworks, Natick, MA) toolbox Cogent (http://www.vislab.ucl.ac.uk/cogent.php). Data of the online sample were acquired via SoSci Survey (Leiner, 2019).

In both tasks, participants were asked to respond to payoff combinations for themselves and another anonymous person. The anonymous other person was introduced as another participant in the same study. It was made clear that the other person would be reimbursed based on the participant's decisions in the SVO task, just as the participant was reimbursed based on another person's decisions in the SVO (except for the participants of the online sample who made hypothetical choices).

2.2.1. SVO Slider Measurement

We adopted a commonly used task (Murphy et al., 2011) to identify participants' SVO using the six primary items for the general SVO score and the nine secondary items for the prosocial motivation score. Fig. 1 provides an outline and Supplementary Fig. 1 visualizes a specific example item in detail. Participants were asked to indicate their preferred allocation of payoffs for themselves and another anonymous person sequentially for each item. Specifically, each SVO item consists of a continuum of 9 different options for joint payoffs for self and another person.

Primary items (used for calculating the SVO scores) are delimited by a circle (centered at 50 points with a diameter of 50 points; Fig. 1c). Intermediate payoff options lie on a line between the endpoints on the circle. For example, item 6 ranges from the first option with 100 points for self and 50 points for other to the ninth option with 85 points for self and 85 points for other (Fig. 1a & c). A more prosocial participant would choose one of the options close to 85 for self and 85 for other, while a more individualistic participant would choose one of the options close to 100 for self and 50 for other.

Secondary items (used for calculating the prosocial motivation scores) allow a differentiation between two different motivations for being prosocial: joint gain maximization and inequality aversion. For example, item 9 ranges from the first option with 100 points for self and 70 points for other to the ninth option with 50 points for self and 100 points for other (Fig. 1b & d-e; Supplementary Fig. 1). An inequality averse participant would choose one of the options close to the identity line, while a joint-gain-maximizing participant would choose one of the options closer to the upper right corner of the item space. Supplementary Fig. 2 visualizes (absolute) inequality and join gain of all secondary items.

See the section on Data analysis below for details on the calculation of the SVO score from primary items and of the prosocial motivation score from secondary items. One decision was randomly chosen and participants were paid accordingly (they received the payoff for self from their choice and the payoff for other from another participant's choice).

2.2.2. Joint Payoff Evaluation (JPE) task

All participants subsequently performed a task established by Haruno and Frith (2010). Pairs of payoffs for the participant and an anonymous other were presented sequentially in a randomized order. For example, the payoff combination for item 1 is 23 points for self and 177 points for other (see Supplementary Fig. 2 & Supplementary Table 2 for all 36 items). As in the SVO task, payoff combinations were defined by two orthogonal dimensions: outcomes for self and outcome for other. Payoffs were uniformly sampled from a circle (centered at 100 points with a diameter of 100 points). That is, all payoff options lie on the circle. In contrast to the SVO task, participants were not asked to choose among different options. Instead, in the JPE task participants were asked to rate each single pair of payoffs on a 4-point Likert scale (1 = least preferable, 4 = most preferable). The social context was controlled in both tasks since participants did not meet the other person and did not receive feedback about the other person's choices. Participants' evaluations were used as dependent variables in economic models (see below). The laboratory samples completed all 36 items of the task. The online sample only completed 12 of the 36 items (every third item) due to time constraints in the online assessment, which also included unrelated tasks (described in Korn et al., 2018).

2.3. Data analysis

All statistical analyses were done using MATLAB (Mathworks, Natick, MA) and all statistical tests are two-tailed. Bar charts were produced in R (R Core Team, 2019) using the package "ggplot2" (Wickham, 2016) and the raincloud plots of Fig. 2 were produced in R using the packages "ggplot2" and "RainCloudPlots" (Allen et al., 2018, 2019).

2.3.1. Use of samples

Our primary results were derived from data acquired in the laboratory from a matched control sample: We compared the results of the UW patient to the control sample in the SVO task and the JPE task. In addition, as secondary, exploratory analyses, we provide a comparison with the age- and gender-matched subgroup of the online sample, which allowed us to compare the UW patient to a larger sample that was tested in a similar but less comparable design (i.e., tasks were done online and participants answered only one-third of the items in the JPE task). For interested readers, we also provide the behavioral results of the full online sample and the student sample.

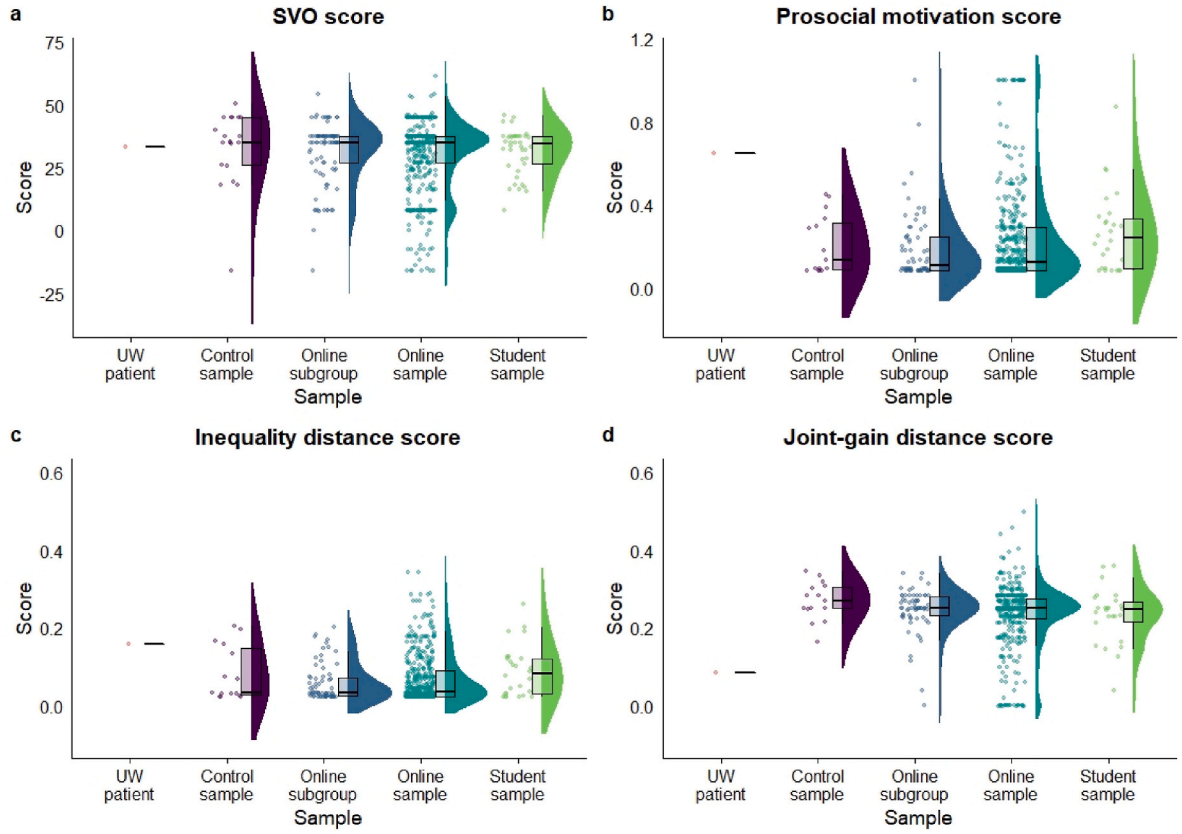


Fig. 2. SVO scores and prosocial motivation scores. Raincloud plots showing the distribution of (**a**) SVO scores, (**b**) prosocial motivation scores, (**c**) inequality distance scores, and (**d**) joint-gain distance scores. We compared the UW patient to the age- and gender-matched control sample and the online subgroup. The UW patient did not differ from these healthy individuals with respect to SVO but showed a higher prosocial motivation score—and specifically a lower joint-gain distance score—which indicates relatively higher joint gain maximization. Data from the online sample and the student sample are depicted for interested readers but were not directly compared to the UW patient. Prosocial motivation scores, inequality distance scores, and joint-gain distance scores are depicted for participants categorized as prosocial according to their SVO scores (i.e., for participants with SVO scores between 22.45 and 57.15). The prosocial motivation score is calculated as the inequality distance score divided by the sum of the inequality distance score and the joint gain distance score. Therefore, the prosocial motivation scores range from 0 (perfect inequality aversion) to 1 (perfect joint gain maximization). The central mark indicates the median, boxes depict 25- and 75-percentiles. The upper (and lower) whiskers extend from the upper (and lower) edges of the box to the largest (and smallest) value within a range of maximal 1.5 times the interquartile range. All data points including extreme data points are shown. For visualization, we also plot estimated distributions. Note that the kernel density estimation for the density plot generates data that fall outside the bounds of the original data. Therefore, the values on the y-axis in (b) extend to 1.2 although the maximum score is bounded to 1. See Supplementary Figs. 3 & 4 for choice distributions on all SVO items.

2.3.2. *SVO scores*

The SVO score was computed from the six primary items. The SVO score is the polar angle in a Cartesian coordinate system with payoffs to self on the x-axis and payoffs to the other person on the y-axis. This angle captures how much participants were willing to sacrifice their own payoff for the other person's payoff. The smaller the angle, the more a participant cares for self and the closer the angle to 90°, the more a participant cares for the other person (Fig. 1c). Continuous SVO scores were used for analyses and are depicted in Fig. 2a. Based on these continuous SVO scores and on the bins established by Murphy et al. (2011), participants were assigned to one of the four categories (altruistic, prosocial, individualistic, and competitive). For example, per definition, an ideal individualist would choose the option that maximizes the payoff for self on each of the items. This principle leads to specific idealized boundaries between the four SVO types. Category boundaries were set symmetrically between the idealized SVO type scores (for details see Murphy et al., 2011).

2.3.3. *Prosocial motivation scores*

To identify the underlying motivation of prosocial preferences, we followed the procedure described in Murphy et al. (2011) for analyzing the nine secondary items of the SVO task. (The code provided by Murphy and colleagues and our publicly available code provide details).

First, an "inequality distance score" quantifies how distant the chosen payoff options are from the options that ideally minimize absolute inequality for the secondary items of the SVO task. Essentially, the payoff ranges of the nine secondary items intersect the diagonal identity line (Fig. 1d & e). The closer a chosen payoff option is to the identity line, the more equal are the payoffs for self and other (Fig. 1d; Supplementary Figs. 1 & 2).

Second, a "joint-gain distance score" quantifies how distant the chosen payoff options are from the options that ideally maximize joint gain for the secondary items of the SVO task. The more to the upper right corner of the coordinate system a chosen option is the higher is the joint gain (Fig. 1e; Supplementary Figs. 1 & 2).

These two scores are combined into participants' "prosocial motivation score," which is a single index ranging from 0 (perfect inequality aversion) to 1 (perfect joint gain maximization). That is, this prosocial motivation score is calculated as the inequality distance score divided by the sum of the inequality distance score and the joint-gain distance score. Based on this prosocial motivation score, each participant with consistent prosocial choices in primary and secondary items can be categorized as inequality averse (prosocial motivation score ≤ 0.5) or joint gain maximizing (> 0.5). In addition to participants' SVO scores

and categories, the rank order of participants' social preferences can be assessed to check if participants make prosocial choices on the secondary items. We followed the recommendations by Murphy et al. (2011) and only analyzed prosocial motivation scores from participants who were categorized as prosocials on both primary and secondary items (see Supplementary Table 1).

2.3.4. Comparison of SVO and prosocial motivation scores

To investigate whether the UW patient differs in her social preferences from the healthy population, we compared her SVO score—and because she was prosocial also her prosocial motivation scores (see results)—to the control sample and a matched subgroup of the online sample.

We conducted a Bayesian single-case analysis using the SingleBayes software (Crawford and Garthwaite, 2007), which applies Bayesian Monte Carlo methods to test the null hypothesis that a patient's score is an observation from the control population. Bayesian single-case analyses are considered robust against deviations from normal distribution (Crawford and Garthwaite, 2006), and provide Bayesian p values, and point (percentile) and interval estimates (Bayesian credibility intervals [CI]) for the percentages of the control groups that would obtain a lower score than the UW patient.

2.3.5. Model-based analyses of the JPE task

We compared eight different economic models that tested for different social motives. These economic models took the form of linear regressions on each individual participant's evaluations of the desirability of a joint payoff (on a scale from 1 to 4) in the JPE task. In line with Haruno and Frith (2010) and Liu et al. (2019), the models assume participants' utility of the given payoff pairs to be linear functions of different predictors (plus an intercept term).

3. Results

3.1. SVO task: comparison of SVO scores between the UW patient and controls

The UW patient was categorized as prosocial, as was the majority of the control sample (80 %) and the other healthy samples (79–81 %; see Table 1 for distributions of SVO categories across all samples). Testing for our primary hypothesis, we found no evidence for a difference in SVO scores between the UW patient (*SVO score* = 33.43) and the healthy control sample (M = 32.78, SD = 15.02) (see Fig. 2a for overall SVO scores and Supplementary Fig. 3 for choice distributions across all primary items). That is, using a Bayesian single-case method, we found that an estimated 51.64 % of the control sample (95 % CI [34.57, 68.58], p = .967) would have a lower SVO score than the UW patient, indicating no difference between the patient and the healthy population. Additionally, in exploratory analyses, the UW patient did not differ from the matched subgroup of the online sample (online subgroup: M = 31.24, SD = 12.4; percentile estimated = 56.95, 95 % CI [48.76, 64.9], p = .861). In other words, the UW patient was quite close to the mean of the tested healthy samples in terms of her general SVO score.

Table 1
Distribution of SVO categories across all samples.

	UW patient	Control sample	Online subgroup	Online sample	Student sample
Altruistic	0	0	0	1 (<1)	0
Prosocial	1	16 (80)	74 (81)	474 (79)	32 (80)
Individualistic	0	3 (15)	16 (18)	120 (20)	8 (20)
Competitive	0	1 (5)	1 (1)	8 (1)	0
Total N	1	20	91	603	40

Participants in the categories according to SVO are given as N (%). This table lists all participants who are categorized as prosocial according to the primary items. See Supplementary Table 1 for exclusion criteria.

3.2. SVO task: comparison of prosocial motivation scores between the UW patient and controls

Since the UW patient was categorized as prosocial, we could look into the secondary items to identify her underlying prosocial motivation (as the secondary items are only applicable if a participant made prosocial choices in the primary items). The UW patient differed from the majority of prosocials in her prosocial motivation score. While 100 % of the prosocials in our control sample (and 88–94 % of the participants in the other healthy samples) were motivated to minimize inequality, the UW patient was motivated to maximize joint gains (see Fig. 2b, Table 2, and Supplementary Fig. 4 for distributions of prosocial motivation scores, categories, and choice patterns, respectively). The Bayesian single-case analysis showed that an estimated 99.5 % of the healthy population would have a lower prosocial motivation score, i.e., would be more inequality averse, than the UW patient (UW patient: *score* = 0.651; control sample: M = 0.207, SD = 0.142, percentile estimate = 99.54, 95 % CI [96.87, 99.99], p = .009). The UW patient did also differ from the matched subgroup of the online sample (online subgroup: M = 0.193, SD = 0.171, percentile estimate = 99.5, 95 % CI [98.43, 99.93], p = .009). Thus, even at the Bonferroni corrected alpha level of $p < .025$ the difference in prosocial motivation scores of the UW patient and the control group was robust.

The proscocial motivation score combines inequality aversion and joint gain maximization. Therefore, we analyzed the respective scores separately. The UW patient did not differ from controls in her inequality distance score (UW patient: *score* = 0.16; control sample: M = 0.08, SD = 0.07, percentile estimate = 85.62, 95 % CI [68.06, 96.3], p = .288; online subgroup: M = 0.06, SD = 0.05, percentile estimate = 97.43, 95 % CI [94.26, 99.21], p = .051; Fig. 2c). But she differed in her joint-gain distance score (UW patient: *score* = 0.09; control sample: M = 0.27, SD = 0.05, percentile estimate = 0.18, 95 % CI [0.00, 1.46], p = .004; online subgroup M = 0.25, SD = 0.06, percentile estimate = 0.51, 95 % CI [0.07, 1.6], p = .01; Fig. 2d).

The prosocial motivation score of the UW patient is very unlikely to have arisen by chance, i.e., by the patient making random choices. We simulated 10,000 random answers for the nine secondary items according to a uniform distribution (i.e., equal probabilities for choosing each of the nine options for each item). As expected due to random answers, only a subset of 6.2 % (i.e., 620) of these 10,000 simulated "participants" were prosocial on both primary and secondary items. The score of the UW patient (or score at least as extreme in the joint gain direction) was only obtained in 0.005 % of these 620 simulated participants. Taken together, this speaks for a difference in the underlying motivation for prosocial preferences in the UW patient compared to the healthy population.

3.3. JPE task: comparison of economic models for allocation evaluations

For allocation evaluations in the JPE task, the Inequality & Joint-gain model performed best in the UW patient (i.e., had the lowest

Table 2
Distribution of prosocial motivation across all samples.

	UW patient	Control sample	Online subgroup	Online sample	Student sample
Prosocial – IA	0	15 (100)	62 (94)	382 (88)	23 (88)
Prosocial – JG	1	0	4 (6)	50 (12)	3 (12)
Total N	1	15	66	432	26

Participants are given as N (%). Prosocial motivation scores are given for all participants who are prosocial according to primary and secondary items (which is why the total numbers can be lower than the number of prosocials listed in Table 1). IA = Inequality aversion, JG = Joint gain maximization. See Supplementary Table 1 for exclusion criteria.

Bayes factor; Fig. 3). The Fehr-Schmidt model was a close runner-up and performed second best. The Fehr-Schmidt model considers the payoff for self, disadvantageous inequality (self gets less than other), and advantageous inequality (other gets less than self). Mathematically, the Fehr-Schmidt model for the JPE task is equivalent to the model formulation used by Haruno and Frith (2010) and to a formulation that replaces payoffs for self by joint gains (see Appendix in the Supplementary Material).

In all healthy samples, the Inequality & Joint-gain model fitted second best and the Fehr-Schmidt model fitted best (using Bayesian model comparisons according to fixed-effects analyses based on log-group Bayes factors; Fig. 3). These fixed-effects analyses neglect variability between participants. We therefore, performed random-effects analyses by calculating protected exceedance probabilities. We found that the Inequality & Joint-gain model had higher protected exceedance probabilities than the Fehr-Schmidt model in all healthy samples (Supplementary Table 4). This implies that the Inequality & Joint-gain model performed better (i.e. had a lower BIC) in more participants than the Fehr-Schmidt model (Supplementary Table 5)—but for those participants, for whom the Fehr-Schmidt model was better, it tended to be better by a larger margin.

Taken together, two related models performed well in the JPE task across samples. Both models consider inequality and gain magnitudes. However, the employed version of the JPE task did not allow a good distinction between the fit of the Inequality & Joint-gain model and the Fehr-Schmidt model. We therefore cannot conclude that a different economic model better explained the evaluations of the UW patient.

3.4. JPE task: comparison of parameter estimates from the Inequality & Join-gain model between the UW patient and controls

We tested whether parameter estimates from the Inequality & Joint-gain model differed between the UW patient and healthy participants (both in the control sample and in the online subgroup). Using Bayesian single-case analysis, we found no evidence for differences in the two parameter estimates of interest: 1) inequality (patient: *estimate* = −0.011; control sample: $M = -0.011$, $SD = 0.004$, percentile estimate = 50.01 %, 95 % CI [32.66, 67.37], $p = .999$; online subgroup: $M = -0.007$, $SD = 0.005$, percentile estimate = 19.48 %, 95 % CI [13.23, 26.8], $p = .39$) and (2) joint-gain (patient: *estimate* = 0.005; control sample: $M = 0.0036$, $SD = 0.002$, percentile estimate = 71.2 %, 95 % CI [53.63, 85.59], $p = .576$; online subgroup: $M = 0.004$, $SD = 0.002$, percentile estimate = 73.02 %, 95 % CI [65.05, 80.18], $p = .54$). Parameter estimates are numerically small because the predictors are in a range from 0 to 200 and the dependent variables are ratings in a range from 1 to 4.

In sum, we found no evidence for differences between the UW patient and controls in their evaluations of the payoff allocations included in the used JPE task.

4. Discussion

This case study investigated the impact of amygdala lesions on allocation decisions and preferences. In addition, we provide rather large datasets of healthy participants to characterize various social motives in a wider population. Based on previous neuroimaging findings and evidence in nonhuman primates and rodents, we hypothesized that the UW patient would have less prosocial preferences due to the amygdala lesions and that this would be reflected in her SVO score and possibly the parameters of the economic model describing allocation preferences. We did not find evidence for a less prosocial SVO in the UW patient compared to the healthy population. Since the UW patient was categorized as prosocial, we looked into the prosocial motivation score, which describes the underlying motivations for prosocial SVO. We found that the UW patient's motivation was to maximize joint gains rather than to minimize inequality. All three of our healthy samples showed similar distributions for prosocial motivation scores. Strikingly, a Bayesian single-case analysis revealed that an estimated 99.5 % of the healthy control sample would have a lower prosocial motivation score than the UW patient. The distinctiveness of the UW patient's prosocial motivation score was driven by higher joint gain maximization rather than lower inequality aversion. The prosocial motivation score of the UW patient is unlikely to have arisen by chance (as determined by simulations).

In addition, we compared several economic models on participants' evaluations of joint payoffs in the JPE task and found that the Inequality & Joint-gain model and the Fehr-Schmidt model were the two best performing models—both of which include metrics for inequality and joint- or self-gain. We found no differences in parameter estimates of the UW patient and control groups. Apart from questions of power and specificity with regard to the used version of the JPE task, it could be that a choice context is needed to detect an effect that would correspond to the difference between the UW patient and controls in prosocial motivation as revealed by the SVO task: In the JPE task, participants

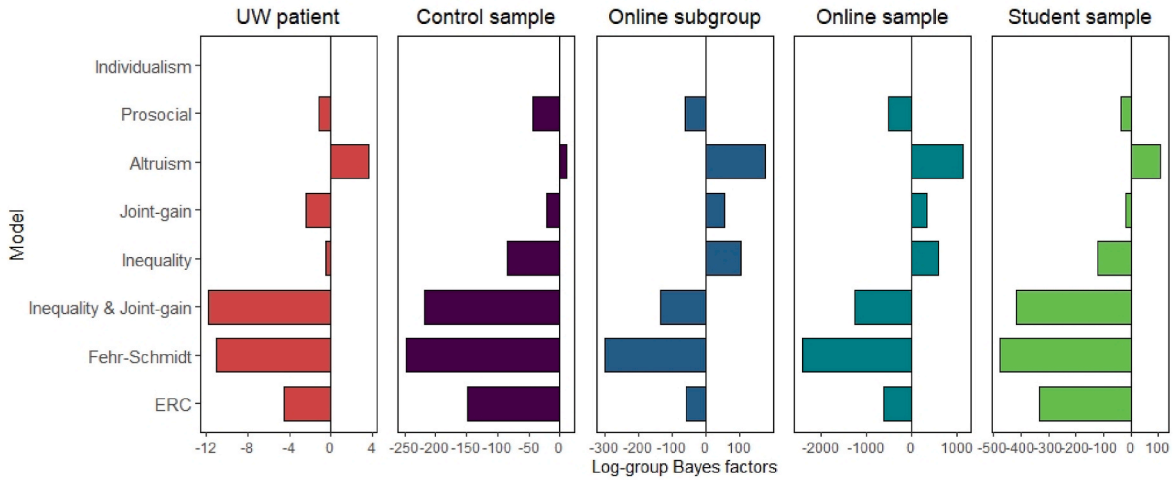


Fig. 3. Model comparisons in the JPE task according to fixed-effects analyses. Log-group Bayes factors relative to the first model (Individualism). The presented fixed-analyses according to log-group Bayes factors assume that every participant uses the same model. The Inequality & Joint-gain model performed best in the UW patient. The Fehr-Schmidt model performed best in all healthy samples. Lower numbers indicate better model fit. See Supplementary Tables 4 and 5 for random-effects analyses, which assume that different participants may use different models. ERC, Equity, Reciprocity, Competition model.

make judgments indicating their evaluation of joint gain and inequality. In contrast, the secondary items of the SVO task force participants to make a decision between these competing motives.

This study was motivated by a neuroimaging finding that the amygdala was the only brain region in which prosocials and individualists differed with respect to tracking the absolute inequality between payoffs for self and other (Haruno and Frith, 2010). A recent neuroimaging study suggested that the amygdala encodes a social-value-distance signal in prosocials (Liu, 2019). Our results speak against a generally less prosocial SVO in a patient with amygdala lesions but are complementary to the neuroimaging findings that showed intra- and inter-individual relationships of the amygdala with metrics of inequality (Haruno and Frith, 2010; Liu et al., 2019). Thus, the amygdala might be functionally and causally involved in setting the specific motivation for prosocial allocations. To determine whether this specific motivation is encoded in the amygdala further studies combining a computational approach with neuroimaging will be needed (Charpentier and O'Doherty, 2018). Mathematically speaking, the UW patient maximized the sum of the payoffs to self and other rather than minimized the differences between the two payoffs. This makes specific predictions for future neuroimaging studies that could pit these (and potentially other) prosocial motivations against each other. Neuroimaging studies could also elucidate the interplay of the amygdala with other brain regions such as the (dorso-)medial prefrontal cortex (Banks et al., 2007) or the insula (Bebko et al., 2015).

Our findings suggest the intriguing possibility that the amygdala might have a rather specific role in the processing of social motives with respect to trading off the maximization of joint gains versus the minimization of inequality. Finer-grained underlying prosocial motivations could also be captured in mathematically defined economic models specified for tasks that separate these prosocial motivations (e.g., by generating allocation tasks that sample the whole allocation plane for self and other). In this sense, our findings could help inform novel modeling and theorizing in psychology and behavioral economics.

We note the general limitation of case studies in terms of sample size. Furthermore, UW syndrome leads to amygdala calcification early in life, hence compensatory mechanisms might have promoted the development of a prosocial preference. A conceptual replication in patients who acquired amygdala lesions later in life would be desirable (Bach et al., 2019). Our findings cannot disentangle the role of different subregions of the amygdala (such as the basolateral amygdala or the central-medial amygdala) since the UW patient tested here shows calcifications in the entire bilateral amygdala (Hurlemann et al., 2007). Human lesion studies commonly consider the amygdala as a unified structure; with the exception of studies on lesions patients with selective basolateral amygdala damage (de Gelder et al., 2014; Klumpers et al., 2015; Koen et al., 2016; Rosenberger et al., 2019; van Honk et al., 2013). The available neuroimaging studies on social motives did not try to separate amygdala nuclei which would require high resolution and possibly an independent method to identify them on an individual basis (e.g. Abivardi and Bach, 2017; Bach et al., 2011). Thus, testing prosocial preferences in patients with more specific lesions would be interesting.

We deem it generally interesting to extend our approach to more vivid tasks with live social interactions and non-monetary incentives because the amygdala has been related to multiple social learning processes (Hurlemann et al., 2010; Olsson and Phelps, 2007). Specifically, an intriguing recent study showed that participants with selective basolateral amygdala damage did not learn to adjust their behavior to the trustworthiness of their interaction partners in a repeated trust game (Rosenberger et al., 2019). It would be worthwhile to see whether the prosocial motivation to maximize joint gain, which might be linked to the amygdala, leads to impaired adaptation during repeated interactive decisions (such as in dictator games, which correspond to the SVO task, trust games, or prisoners' dilemma). Relatedly, live social interactions often entail emotional facial expressions. Since UW patients show aberrations in processing fearful or threatening facial expressions (Bach et al., 2015; de Gelder et al., 2014), it would be interesting to test if social motives would change if the interaction partners enact emotional facial expressions.

To summarize, while our results do not support a generally less prosocial SVO in the UW patient, they indicate a shift toward the motivation for maximizing joint gain in the UW patient with respect to healthy controls.

Author contributions

CWK, DRB, and RH planned the experiments. CWK, DRB and, RH collected the data. LMD and, CWK analyzed the data. LMD, RH, DRB, and, CWK wrote the manuscript.

Data availability

The behavioral data that support the findings of this study are available on GitHub (https://github.com/LisaDoppelhofer/SocialMotives_2021).

Code availability

The MATLAB code for this study is available on GitHub (https://github.com/LisaDoppelhofer/SocialMotives_2021).

Declaration of competing interest

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Acknowledgments

This work was supported by the Emmy Noether Research Group grant (392443797) from the German Research Foundation (DFG).

We thank Rosa Bohlender and Dirk Scheele for help with data acquisition and Benjamin Kuper-Smith, Koen Frolichs, and Lisa Anna Rosenberger for comments on earlier drafts of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2021.108016.

References

Abivardi, A., Bach, D.R., 2017. Deconstructing white matter connectivity of human amygdala nuclei with thalamus and cortex subdivisions in vivo: structural Amygdala-Thalamus and Amygdala-Cortex Connectivity. Hum. Brain Mapp. 38 (8), 3927–3940. https://doi.org/10.1002/hbm.23639.

Allen, M., Whitaker, K., Marshall, T.R., Kievit, R., 2018. RainCloudPlots tutorials and codebase. In: Zenodo. https://doi.org/10.5281/zenodo.3368186, Version v1.1.

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 4, 63. https://doi.org/10.12688/wellcomeopenres.15191.1.

Bach, D.R., Behrens, T.E., Garrido, L., Weiskopf, N., Dolan, R.J., 2011. Deep and superficial amygdala nuclei projections revealed in vivo by probabilistic tractography. J. Neurosci. 31 (2), 618–623. https://doi.org/10.1523/JNEUROSCI.2744-10.2011.

Bach, D.R., Hoffmann, M., Finke, C., Hurlemann, R., Ploner, C.J., 2019. Disentangling Hippocampal and amygdala contribution to human anxiety-like behavior. J. Neurosci. 39 (43), 8517–8526. https://doi.org/10.1523/JNEUROSCI.0412-19.2019.

Bach, D.R., Hurlemann, R., Dolan, R.J., 2015. Impaired threat prioritisation after selective bilateral amygdala lesions. Cortex 63, 206–213. https://doi.org/10.1016/j.cortex.2014.08.017.

Banks, S.J., Eddy, K.T., Angstadt, M., Nathan, P.J., Phan, K.L., 2007. Amygdala–frontal connectivity during emotion regulation. Soc. Cognit. Affect Neurosci. 2 (4), 303–312. https://doi.org/10.1093/scan/nsm029.

Bebko, G., Bertocci, M., Chase, H., Dwojak, A., Bonar, L., Almeida, J., Perlman, S.B., Versace, A., Schirda, C., Travis, M., Gill, M.K., Demeter, C., Diwadkar, V., Sunshine, J., Holland, S., Kowatch, R., Birmaher, B., Axelson, D., Horwitz, S., Phillips, M.L., 2015. Decreased amygdala–insula resting state connectivity in

behaviorally and emotionally dysregulated youth. Psychiatr. Res. Neuroimaging 231 (1), 77–86. https://doi.org/10.1016/j.pscychresns.2014.10.015.
Becker, B., Mihov, Y., Scheele, D., Kendrick, K.M., Feinstein, J.S., Matusch, A., Aydin, M., Reich, H., Urbach, H., Oros-Peusquens, A., Shah, N.J., Kunz, W.S., Schlaepfer, T.E., Zilles, K., Maier, W., Hurlemann, R., 2012. Fear Processing and Social Networking in the Absence of a Functional Amygdala. Biological Psychiatry 72 (1), 70–77. https://doi.org/10.1016/j.biopsych.2011.11.024.
BFS, 2018. https://www.bfs.admin.ch/bfs/de/home/statistiken/preise/kaufkraftparitaeten.html. (Accessed 25 March 2019).
Charpentier, C.J., O'Doherty, J.P., 2018. The application of computational models to social neuroscience: promises and pitfalls. Soc. Neurosci. 13 (6), 637–647. https://doi.org/10.1080/17470919.2018.1518834.
Crawford, J.R., Garthwaite, P.H., 2006. Methods of testing for a deficit in single-case studies: evaluation of statistical power by Monte Carlo simulation. Cogn. Neuropsychol. 23 (6), 877–904. https://doi.org/10.1080/02643290500538372.
Crawford, J.R., Garthwaite, P.H., 2007. Comparison of a single case to a control or normative sample in neuropsychology: development of a Bayesian approach. Cogn. Neuropsychol. 24 (4), 343–372. https://doi.org/10.1080/02643290701290146.
Dal Monte, O., Chu, C.C.J., Fagan, N.A., Chang, S.W.C., 2020. Specialized medial prefrontal–amygdala coordination in other-regarding decision preference. Nat. Neurosci. 23 (4), 565–574. https://doi.org/10.1038/s41593-020-0593-y.
De Cremer, D., Van Lange, P.A.M., 2001. Why prosocials exhibit greater cooperation than proselfs: the roles of social responsibility and reciprocity. Eur. J. Pers. 15 (S1), S5–S18. https://doi.org/10.1002/per.418.
de Gelder, B., Terburg, D., Morgan, B., Hortensius, R., Stein, D.J., van Honk, J., 2014. The role of human basolateral amygdala in ambiguous social threat perception. Cortex 52, 28–34. https://doi.org/10.1016/j.cortex.2013.12.010.
Engel, C., 2011. Dictator games: a meta study. Exp. Econ. 14 (4), 583–610. https://doi.org/10.1007/s10683-011-9283-7.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868. https://doi.org/10.1162/003355399556151.
Haruno, M., Frith, C.D., 2010. Activity in the amygdala elicited by unfair divisions predicts social value orientation. Nat. Neurosci. 13 (2), 160–161. https://doi.org/10.1038/nn.2468.
Haruno, M., Kimura, M., Frith, C.D., 2014. Activity in the nucleus accumbens and amygdala underlies individual differences in prosocial and individualistic economic choices. J. Cognit. Neurosci. 26 (8), 1861–1870. https://doi.org/10.1162/jocn_a_00589.
Hernandez-Lallement, J., van Wingerden, M., Schäble, S., Kalenscher, T., 2016. Basolateral amygdala lesions abolish mutual reward preferences in rats. Neurobiol. Learn. Mem. 127, 1–9. https://doi.org/10.1016/j.nlm.2015.11.004.
Hurlemann, R., Patin, A., Onur, O.A., Cohen, M.X., Baumgartner, T., Metzler, S., Dziobek, I., Gallinat, J., Wagner, M., Maier, W., Kendrick, K.M., 2010. Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. J. Neurosci. 30 (14), 4999–5007. https://doi.org/10.1523/JNEUROSCI.5538-09.2010.
Hurlemann, René, Wagner, M., Hawellek, B., Reich, H., Pieperhoff, P., Amunts, K., Oros-Peusquens, A.-M., Shah, N.J., Maier, W., Dolan, R.J., 2007. Amygdala control of emotion-induced forgetting and remembering: evidence from Urbach-Wiethe disease. Neuropsychologia 45 (5), 877–884. https://doi.org/10.1016/j.neuropsychologia.2006.08.027.
Joireman, J.A., Lasane, T.P., Bennett, J., Richards, D., Solaimani, S., 2001. Integrating social value orientation and the consideration of future consequences within the extended norm activation model of proenvironmental behaviour. Br. J. Soc. Psychol. 40 (1), 133–155. https://doi.org/10.1348/014466601164731.
Klumpers, F., Morgan, B., Terburg, D., Stein, D.J., van Honk, J., 2015. Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. Soc. Cognit. Affect Neurosci. 10 (9), 1161–1168. https://doi.org/10.1093/scan/nsu164.
Koen, N., Fourie, J., Terburg, D., Stoop, R., Morgan, B., Stein, D.J., van Honk, J., 2016. Translational neuroscience of basolateral amygdala lesions: studies of urbach-wiethe disease: BLA Lesions and Urbach-Wiethe Disease. J. Neurosci. Res. 94 (6), 504–512. https://doi.org/10.1002/jnr.23731.
Korn, C.W., Ries, J., Schalk, L., Oganian, Y., Saalbach, H., 2018. A hard-to-read font reduces the framing effect in a large sample. Psychonomic Bull. Rev. 25 (2), 696–703. https://doi.org/10.3758/s13423-017-1395-4.
Korn, C.W., Vunder, J., Miró, J., Fuentemilla, L., Hurlemann, R., Bach, D.R., 2017. Amygdala lesions reduce anxiety-like behavior in a human Benzodiazepine-sensitive approach–avoidance conflict test. Biol. Psychiatr. 82 (7), 522–531. https://doi.org/10.1016/j.biopsych.2017.01.018.
Leiner, D.J., 2019. SoSci survey. https://www.soscisurvey.de. Version 3.1.06.
Liu, Y., Li, S., Lin, W., Li, W., Yan, X., Wang, X., Pan, X., Rutledge, R.B., Ma, Y., 2019. Oxytocin modulates social value representations in the amygdala. Nat. Neurosci. 22 (4), 633–641. https://doi.org/10.1038/s41593-019-0351-1.
MacCrimmon, K.R., Messick, D.M., 1976. A framework for social motives. Behav. Sci. 21 (2), 86–100. https://doi.org/10.1002/bs.3830210203.
Messick, D.M., McClintock, C.G., 1968. Motivational bases of choice in experimental games. J. Exp. Soc. Psychol. 4 (1), 1–25. https://doi.org/10.1016/0022-1031(68)90046-2.
Murphy, R.O., Ackermann, K.A., Handgraaf, M., 2011. Measuring social value orientation. SSRN Electr. J. https://doi.org/10.2139/ssrn.1804189.
Newton, F.H., Rosenberg, R.N., Lampert, P.W., O'Brien, J.S., 1971. Neurologic involvement in Urbach-Wiethe's disease (lipoid proteinosis). A clinical, ultrastructural, and chemical study. Neurology 21 (12), 1205–1213. https://doi.org/10.1212/wnl.21.12.1205.
Olsson, A., Phelps, E.A., 2007. Social learning of fear. Nat. Neurosci. 10 (9), 1095–1102. https://doi.org/10.1038/nn1968.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org.
Rosenberger, L.A., Eisenegger, C., Naef, M., Terburg, D., Fourie, J., Stein, D.J., van Honk, J., 2019. The human basolateral amygdala is indispensable for social experiential learning. Curr. Biol. 29 (20), 3532–3537. https://doi.org/10.1016/j.cub.2019.08.078 e3.
Scheele, D., Zimbal, S., Feinstein, J.S., Delis, A., Neumann, C., Mielacher, C., Philipsen, A., Hurlemann, R., 2019. Treatment-resistant depression and ketamine response in a patient with bilateral amygdala damage. Am. J. Psychiatr. 176 (12), 982–986. https://doi.org/10.1176/appi.ajp.2019.18101219.
Talmi, D., Hurlemann, R., Patin, A., Dolan, R.J., 2010. Framing effect following bilateral amygdala lesion. Neuropsychologia 48 (6), 1823–1827. https://doi.org/10.1016/j.neuropsychologia.2010.03.005.
Tanaka, T., Yamamoto, T., Haruno, M., 2017. Brain response patterns to economic inequity predict present and future depression indices. Nat. Human Behav. 1 (10), 748–756. https://doi.org/10.1038/s41562-017-0207-1.
van Honk, J., Eisenegger, C., Terburg, D., Stein, D.J., Morgan, B., 2013. Generous economic investments after basolateral amygdala damage. Proc. Natl. Acad. Sci. Unit. States Am. 110 (7), 2506–2510. https://doi.org/10.1073/pnas.1217316110.
Van Lange, P.A.M., Schippers, M., Balliet, D., 2011. Who volunteers in psychology experiments? An empirical review of prosocial motivation in volunteering. Pers. Indiv. Differ. 51 (3), 279–284. https://doi.org/10.1016/j.paid.2010.05.038.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

7 General discussion

In this thesis, I examined cooperation as a multi-staged process comprising (i) the evaluation and updating of partner representations, (ii) cooperative choices under interdependence and uncertainty, and (iii) allocation decisions that determine how rewards are distributed between self and other. The main contribution of this thesis is therefore an account that looks at different components of cooperation and specifies which components of cooperation appear preserved in individuals with borderline personality disorder (BPD) and which components are systematically biased under specific contexts, offering insights into why interpersonal functioning can be unstable even when cooperative preferences are not markedly atypical. This is specifically important given the mixed findings in the previous literature.

Across projects, the central pattern that emerges is *not* that individuals with BPD are unwilling to cooperate. Instead, interpersonal dysfunction in BPD is best captured as a set of potentially linked biases that unfold across stages: negatively biased self-evaluations and partner representations, pessimistic expectations under dependence, and context-sensitive differences in action engagement when personal costs are low. In the following, I summarize what each project contributes to this account, how these contributions fit together, and what they imply for theory, neural models, and future research.

7.1 What is altered, what is preserved?

Across Projects 1-3, the most consistent finding is that individuals with BPD do *not* lack cooperative motivation. Rather, the convergent pattern supports four mechanistic claims that help reconcile heterogeneous findings in the literature. Project 4 adds a mechanistic account by showing that different motives can underlie prosocial motivation and that this can be related to amygdala pathology.

7.1.1 Altered social priors with preserved learning mechanisms

Project 1 suggests that social learning in BPD is shaped more by negatively biased trait evaluations than by a clearly dysfunctional updating process. This was assessed in a personality-based social learning task in which participants provided self-ratings on a set of trait words and repeatedly predicted traits of personality profiles, corresponding to BPD-like and control-like profiles, while receiving trial-by-trial feedback. Importantly, the clinical status of these profiles was not disclosed to participants.

Women with BPD showed markedly more negative self-ratings across the trait set of the learning task, a pattern that was confirmed in established personality measures (e.g., higher Neuroticism and lower Extraversion, Agreeableness, and Conscientiousness on the NEO-FFI; Borkenau & Ostendorf, 2007; Costa & McCrae, 1992; and higher maladaptive trait scores on the PID-5-BF; Krueger et al., 2012; Zimmermann et al., 2014). In the learning task, group differences emerged most clearly in initial evaluations: when evaluating others, participants in the control group rated control-like personality profiles more positively than participants with BPD, whereas ratings of BPD-like personality profiles

were comparably negative in both groups, consistent with the trait characteristics inherent to these profiles rather than any explicit diagnostic knowledge. At the same time, learning-related changes in accuracy were modest and depended primarily on the profile type (i.e., was it a BPD-like or control-like profile). Across groups, prediction errors (PEs) decreased across trials for control profiles but not for BPD profiles, indicating that trait inference was generally more difficult for BPD-like personalities. This interpretation is reinforced by model comparisons. In both groups, performance was best captured by fine granularity models. Fine granularity here denotes updating based on specific similarities between individual traits, rather than broad generalization across higher-order personality dimensions, suggesting that both groups use complex knowledge structures for updating (Frolichs et al., 2022; Rosenblau et al., 2023).

Taken together, these findings support the distinction that BPD-related differences may often reflect more negative priors (i.e., where social inference begins) rather than impaired learning mechanisms per se. This aligns with prior work emphasizing negative self-concept and reduced self-related positivity biases in BPD (Korn et al., 2016; Vater et al., 2015; Winter et al., 2017). Finally, the fact that BPD profiles were difficult to infer for both controls and individuals with BPD converges with evidence that people with BPD tend to elicit more negative impressions even in the absence of diagnostic labels (Daros et al., 2016; Hepp et al., 2018; Hepp, Kieslich, Schmitz, et al., 2021; Oltmanns et al., 2004). The combination of more negative partner evaluations and of being evaluated more negatively by others may thus promote a self-reinforcing interpersonal dynamic in which positive interactions are less likely to be initiated, noticed, or maintained.

7.1.2 Preserved cooperative motivation with altered expectations

Project 2 shows that cooperative behavior in BPD is not globally reduced but is strongly shaped by social context (i.e., whether the partner's survival in the task was contingent on the participant's choice) and personal risk, highlighting the context-sensitive nature of social decision-making. In the interactive cooperation task, the critical choice was whether to select Play rather than Wait. Conceptually, Play always increased the probability of a positive outcome for both players, but its helping component varied across social contexts, such that it constituted a more other-regarding helping choice in dependent-partner trials and a more mixed-motive option in independent-partner trials.

When the partner needed help, both groups chose Play more often than when the partner was independent, indicating preserved sensitivity to another person's need under explicit interdependence. Group differences were subtle and emerged primarily under conditions with low compared to high personal risk and across social context conditions. Controls chose Play more often than participants with BPD when personal costs were minimal, whereas no group difference was observed under high risk. Because the exact meaning of Play shifts with partner dependence (i.e., becomes more other-regarding or more mixed-motive), this pattern is best interpreted as a reduced action engagement when costs are

minimal, rather than a broad deficit in cooperative capacity. Crucially, when participants estimated their partner's behavior in a role-reversed version of the task, individuals with BPD expected less help when they themselves depended on the partner.

This combination of a preserved need-sensitivity in one's own choices but pessimistic expectations about the partner's cooperative behavior under dependence fits with evidence for negative interpersonal priors and reduced trust in BPD (Barnow et al., 2009; Korn et al., 2016; Masland et al., 2020; Weinbrecht et al., 2018). At the same time, it is consistent with accounts suggesting that active cooperation capacity (i.e., willingness to give) may remain intact in BPD, while difficulties emerge in contexts that depend on predicting others' intentions and interpreting others' behavior (Abramov et al., 2020; Hepp et al., 2014; King-Casas et al., 2008; Lévay et al., 2021; Thielmann et al., 2014). Together, these patterns provide a plausible bridge to real-world instability: individuals with BPD may behave cooperatively when the situation clearly calls for it, yet still expect others to be unreliable when support is needed. Under everyday ambiguity, these pessimistic expectations can bias how partner behavior is interpreted and can make negative or malevolent intent seem more likely over time.

7.1.3 Preserved allocation behavior in non-interactive settings

Project 3 shows that allocation behavior and fairness-related decisions in BPD are largely preserved under baseline, non-interactive conditions. This project targeted the third stage of the thesis framework – deciding how payoffs are distributed between self and other – using a battery of one-shot economic tasks with anonymous partners. These tasks assessed active cooperation, reactive cooperation, social preferences, and explicit fairness judgements in settings without ongoing interactions.

Under these controlled conditions, women with BPD behaved largely similar to controls across measures of active cooperation and reactive cooperation, as well as across explicit fairness judgements. For active cooperation, both the Social Value Orientation Measurement and the Dictator Game (tasks in which participants actively allocate money between themselves and another anonymous person) yielded no clear evidence for reduced sharing in BPD, and both groups were categorized as prosocial. Importantly, reactive cooperation was also preserved: in the Ultimatum Game (recipient role) both groups showed the expected sensitivity to offer size, with no indication that the BPD diagnosis or maladaptive trait dissociality systematically altered rejection behavior. This finding indicates that reactive responses to (un)fairness are intact when interactions are anonymous and one-shot. This interpretation is further supported by a mini-meta-analysis conducted within Project 3, which integrated the present Ultimatum Game data with previously published Ultimatum Game studies in BPD and revealed no consistent evidence for group differences across task variants. To dissociate behavior from normative evaluation, participants additionally provided fairness ratings for hypothetical offers, again without reliable differences between groups.

Importantly, Project 3 also examined whether apparently intact prosocial allocation behavior in BPD reflects different underlying motives. Among participants classified as prosocial, secondary Social Value Orientation items indicated that all prosocial individuals in the BPD group – and the majority of prosocial controls – were primarily motivated by inequality aversion rather than joint-gain maximization. Finally, an adapted Joint Payoff Evaluation task sampled a broader range of allocation pairs and enabled computational tests of underlying motives. Here, the same model families captured behavior across groups indicating that both groups show prosocial motives such as inequality aversion.

Together, these findings indicate that interpersonal dysfunction in BPD does not arise from globally atypical valuation of payoff distributions or an inability to engage in cooperation per se. Rather, they indicate that both active and reactive cooperation are in principle intact under anonymous, one-shot conditions. Clinically relevant differences are more likely to emerge when allocation and fairness processes are embedded in richer social contexts – where intentions are ambiguous, feedback is socially meaningful, and cooperation must be maintained or repaired over time (Abramov et al., 2020; King-Casas et al., 2008; Liebke et al., 2018; Masland & Hooley, 2020).

7.1.4 Preserved prosocial behavior and altered underlying motives

Project 4 extends the allocation stage of the thesis framework by providing a mechanistic perspective on social motives, demonstrating that preserved prosocial behavior can coexist with markedly altered underlying motive composition. Using the same Social Value Orientation and Joint Payoff Evaluation tasks as Project 3, this project examined whether selective bilateral amygdala lesions are associated with shifts in prosocial orientation and underlying motivation.

In the case study with a patient with Urbach-Wiethe syndrome, the amygdala pathology did not seem to be related to reduced prosocial orientation: the patient's Social Value Orientation score fell well within the prosocial range observed in the matched-control sample and the wider population samples. However, when prosocial choices were decomposed, the patient showed a highly distinctive profile characterized by strong joint-gain maximization rather than inequality aversion. Bayesian single-case analyses suggested that this motive profile was extreme relative to the control distribution, and simulations indicated that it was unlikely to reflect chance variation. Notably, this dissociation was not mirrored in model-based analyses of Joint Payoff Evaluation judgments, consistent with the notion that motive differences may be more readily expressed in choice contexts that require explicit trade-offs between competing motives than in evaluating payoff distributions.

Mechanistically, this pattern aligns with prior neuroimaging work linking the amygdala to social motive signals during allocation evaluation (Haruno & Frith, 2010; Liu et al., 2019) and provides rare lesion-based evidence that the amygdala may causally contribute to weighting different prosocial motives rather than prosocial orientation per se. Within the thesis framework, Project 4 therefore serves as a

targeted mechanistic complement to the clinical findings of Project 3. While no motive differences were observed in the BPD group, the Urbach-Wiethe case illustrates how apparently intact prosocial behavior can arise from qualitatively different valuation processes in other neurobiological contexts.

Together, Projects 3 and 4 demonstrate that similarity in allocation behavior or fairness judgments does not guarantee similarity in the motivational computations that give rise to these behaviors. Whereas participants in Project 3 (both controls and the BPD group) were predominantly characterized by inequality aversion, the Urbach-Wiethe patient in Project 4 showed a pronounced tendency toward joint-gain maximization despite remaining prosocial overall. This dissociation underscores the value of motive decomposition for understanding interpersonal functioning and suggests that differences in social stability may arise not only from whether individuals behave prosocially, but from how social value is computed when outcomes must be shared.

7.2 An integrative account: negative priors and preserved cooperative capacity

Taken together, the findings of this thesis support an account in which interpersonal dysfunctions in BPD are best understood as a combination of negatively biased priors and expectation-driven interpersonal dynamics, rather than as generalized unwillingness or inability to cooperate. Across projects, two convergent patterns are particularly robust.

First, the results point to negative priors as systematic input into social inference. Individuals with BPD showed more negative self-evaluations and more negative evaluations of other people's personalities. In addition, they also tended to hold pessimistic expectations about others' cooperative behavior, especially when their own outcomes depend on the partner. In this framework, the critical alteration is *not* that individuals with BPD cannot learn from social information, but that social inference often begins with a more negative baseline.

Second, across projects, cooperative capacity and prosocial motivation appear broadly preserved. Measures of social preferences and fairness sensitivity, as well as both active and reactive forms of cooperation, did not indicate a shift toward selfishness or divergent social motives. Even in a repeated and interactive paradigm, participants with BPD demonstrated a reliable ability to act prosocially, when their partner needed help. This convergence is important as it demonstrates that the general willingness to cooperate is intact in individuals with BPD.

The combination of these two findings – preserved willingness to cooperate alongside pessimistic social expectations – helps reconcile two observations that often appear contradictory: individuals with BPD can show intact cooperation, fairness sensitivity, and social motives in controlled settings, yet experience pronounced instability in everyday relationships. The present results suggest that what becomes problematic is not prosocial motivation itself, but that negatively biased partner representations and pessimistic expectations may constrain subsequent cooperation, particularly in everyday settings where

intentions must be inferred and social feedback is noisy. Negative self-views and pessimistic social expectations may also promote confirmatory sampling such that attention is preferentially directed toward information that confirms negative beliefs (Liebke et al., 2018; Van Schie et al., 2020, 2024). This interpretation is consistent with accounts proposing biased processing of social feedback in BPD (Jeung et al., 2018; Korn et al., 2016; Waite et al., 2024) and provides a conceptual bridge from the present findings on negative priors and pessimistic expectations to the self-reinforcing dynamics that characterize unstable relationships.

7.3 Convergence and divergence across paradigms

Building on this integrative account, one important question is *when* biases in social decision-making are likely to become visible across different experimental paradigms. From this perspective, the present findings offer a parsimonious explanation for heterogeneity across studies of social decision-making in BPD.

Methodological choices strongly shape what is measured when we claim to assess "cooperation". Under anonymous, one-shot conditions, social behavior is relatively decontextualized and primarily reflects baseline preferences and capacities. In such paradigms, small or absent group differences should not be interpreted as evidence that interpersonal dysfunction in BPD is unrelated to cooperative behavior per se. Rather, they indicate that these tasks might not engage the processes that are most diagnostically relevant. At the same time, preserved preferences can represent genuine opportunities for positive interactions.

The present findings also refine distinctions between active and reactive cooperation in BPD. In the present thesis, paradigms that were formally reactive – such as the one-shot Ultimatum Game – did not reveal group differences when interactions were anonymous and non-dynamic. This indicates that basic reactive responses to unfairness can remain intact in BPD under conditions that do not involve ongoing interaction or learning about a partner. At the same time, this pattern suggests that reactive cooperation deficits proposed in the literature (De Panfilis et al., 2019; King-Casas et al., 2008; Thielmann et al., 2014) may depend on dynamic, reciprocal interactions, in which partner behavior must be interpreted over time and incorporated into future decisions, rather than on reactive responding per se.

In Project 2, the present thesis also included a novel repeated, interactive cooperation paradigm, allowing a more direct comparison between one-shot and interactive settings. Importantly, interaction in this task was structured and pre-programmed, rather than based on reciprocal behavior from a human partner. Within this framework, cooperative behavior in individuals with BPD was largely intact when helping was clearly required. That is, when the partner's outcome depended on the participant's choice, both groups reliably increased cooperative behavior. This demonstrates that interactive task structure

alone is not sufficient to elicit group differences, and that cooperative engagement can remain preserved in BPD under clear helping demands.

Crucially, group differences did not emerge in the interaction per se, but selectively under conditions of weak external constraints – that is, when personal risk was low and the decision was less normatively prescribed. Here, "normatively prescribed" refers to dependent-partner trials, where the cooperative option is to help because the partner depends on help. By contrast, in independent-partner trials, help was not necessarily needed, making the cooperative option more mixed-motive (self- and other-regarding) and therefore more strongly dependent on internally generated valuation rather than adherence to an obvious norm (i.e., helping when help is needed). Behavioral group differences emerged under low risk across social contexts, whereas neural differences appeared more selectively under low-risk, independent-partner trials. Together, this pattern suggests that group differences in this task are most visible when both costs and normative guidance are minimal, indicating reduced engagement with potentially beneficial actions in situations with weak external constraints. It was precisely under these weakly constrained conditions that subtle group differences emerged, with controls increasing engagement more strongly under low risk than participants with BPD, despite preserved sensitivity to risk for partner expectations.

Taken together, these observations indicate that group differences are not uniformly tied to cooperation itself but instead seem to depend on whether a paradigm elicits expectation-dependent engagement and regulation under weak normative guidance. An important next question concerns how these context-sensitive behavioral patterns map onto neural systems that support valuation, control, and social motive components.

7.4 Neural implications

At the neural level, the findings of Project 2 support a context-sensitive interpretation of prefrontal recruitment rather than a general disruption of medial prefrontal cortex (PFC) function in BPD during cooperative decision-making. Neural group differences emerged most clearly in contexts with weak constraints, characterized by low personal risk and less normatively prescribed decisions (i.e., low risk, independent-partner conditions). Independent-partner conditions are described as less normatively prescribed because the partner did not necessarily need help, such that decisions relied more strongly on internally generated valuation of mixed self- and other-regarding motives than on adherence to obvious helping norms. Within these weak-constraint contexts, controls showed stronger recruitment of lateral prefrontal regions, most notably bilateral inferior frontal gyrus (IFG), alongside a smaller cluster in the orbitofrontal cortex (OFC).

This pattern is notable given converging evidence that inferior frontal regions constitute a less discussed but potentially important locus of neural alteration in BPD. Recent meta-analyses report reduced IFG

engagement in social-cognitive and affective paradigms (Schurz et al., 2024), with broader task-general meta-analytic evidence across fMRI studies likewise indicating reduced inferior frontal convergence in BPD relative to healthy controls (Degasperi et al., 2021). Functionally, the IFG has been implicated in inhibitory and regulatory control processes (Aron et al., 2014; Christopoulos et al., 2009; Juan & Muggleton, 2012; Maia & Frank, 2011; Menon et al., 2001), as well as in the reappraisal of others' intentions (Grecucci et al., 2013; Wager et al., 2008). Taken together, these findings tentatively suggest that the IFG may be a relevant node for understanding social decision-making in BPD, particularly in situations that are weakly constrained and involve the integration of self- and other-regarding considerations. The reduced IFG recruitment observed here may reflect diminished engagement of regulatory or evaluative processes under such conditions, potentially including the flexible interpretation or reappraisal of others' needs or intentions. This interpretation, however, is necessarily indirect and should be treated with caution, as these functional links are inferred from prior literature rather than directly tested in the present thesis.

The medial prefrontal findings further refine this interpretation. Across groups, independent-partner (relative to dependent-partner) conditions were associated with stronger engagement of medial orbitofrontal cortex (OFC) and a broader medial prefrontal network, alongside recruitment of regions implicated in salience and memory processing (including insula and hippocampus). Given that independent-partner conditions required the integration of self- and other-regarding incentives under reduced normative guidance, this pattern aligns with prior work linking orbitofrontal and ventromedial prefrontal regions to the representation of diverse and also abstract rewards (Bray et al., 2010; Cutler & Campbell-Meiklejohn, 2019; Diekhof et al., 2012; Peters & Büchel, 2010), as well as to integrative valuation and mentalizing processes (Gangopadhyay et al., 2021; Morelli et al., 2015; Schurz et al., 2014; Tusche & Bas, 2021; Zoh et al., 2022). Notably, OFC activity during prosocial behavior has also been linked to rewarding aspects of cooperation itself (Cutler & Campbell-Meiklejohn, 2019; Decety et al., 2004). Within the task logic, dependent-partner conditions reduced ambiguity because helping constituted a clearly normatively prescribed response, whereas independent-partner conditions increased demands on valuation and context integration. Accordingly, medial prefrontal engagement in the present task appears to reflect integrative valuation under weaker constraints rather than the salience of helping per se.

Importantly, behavioral and neural effects showed partial dissociation. Behavioral group differences in cooperative choice emerged under low personal risk across social contexts, whereas neural group differences were more selective, with reduced lateral prefrontal and OFC engagement in individuals with BPD during low-risk, independent-partner trials. This dissociation suggests that behavioral sensitivity to low costs may generalize across contexts, while alterations in prefrontal engagement become most apparent when both personal costs and normative constraints are minimal.

Project 4 adds a complementary and mechanistically distinct neural perspective by suggesting that the amygdala may shape how social value is computed through differential weighting of prosocial motives (inequality aversion versus joint-gain maximization), rather than determining overall prosocial orientation. In the Urbach-Wiethe case, selective bilateral amygdala damage was associated with a marked shift in motive composition despite preserved prosocial orientation, a pattern not observed in the BPD sample in Project 3. Given the substantial differences in sample, method, and task demands, these findings should not be directly compared with the fMRI results of Project 2. Instead, they provide converging but independent evidence that distinct neural systems may differentially contribute to regulatory engagement during cooperative decision-making (lateral prefrontal systems) and to valuation and weighting of prosocial motives during allocation (amygdala-related mechanisms).

Taken together, these neural findings across projects motivate a neurocomputational picture in which lateral prefrontal systems, including the IFG, support context-sensitive engagement and regulatory control when external constraints are weak, with medial OFC contributions reflecting valuation under mixed-motive conditions, while the amygdala contributes to how different prosocial motives are weighted during valuation. Crucially, these results suggest that neural signatures most informative for understanding interpersonal dysfunction in BPD may not be those elicited by overt helping demands, but rather those that emerge when decisions require internally generated regulation and value integration under reduced normative and motivational structure.

7.5 Clinical implications: targets and opportunities

The multi-staged framework developed in this thesis highlights several processes that may be particularly informative for clinical intervention.

A first target concerns negatively biased self-evaluations, which may shape how individuals with BPD enter interpersonal situations and evaluate their own role and worth within them. These biased starting points may foster an early negative focus and may initiate a self-reinforcing cycle in subsequent interaction.

A second, closely related target concerns negatively biased evaluations of others and pessimistic expectations about their behavior, especially in situations of interdependence. Across projects, cooperative behavior was often preserved, yet expectations about others' willingness to help were reduced when participants depended on their partner. Such pessimistic expectations may bias the interpretation of ambiguous social signals and increase the likelihood that neutral or mildly negative behavior is perceived as rejection, neglect, or exploitation.

Third, dependence might be a particularly critical context for these biases. When outcomes hinge on another person's action, expectations about reliability and support become personally consequential. If expectations are pessimistic precisely under dependence, individuals may disengage or escalate when

cooperation would otherwise be adaptive, thereby contributing to instability even in the absence of reduced prosocial intent.

Finally, the thesis identifies important and clear opportunities for intervention. Preserved learning mechanisms in structured tasks suggest that individuals with BPD are capable of updating beliefs when contingencies are clear and feedback is reliable, pointing to the potential value of interventions that explicitly target the recalibration of self- and other-related priors. In addition, preserved need-sensitivity and baseline willingness to cooperate indicate that prosocial motivation remains accessible. When expectations are made explicit and responsibility is clearly defined, cooperative engagement can potentially be supported, stabilized, and generalized beyond highly structured contexts.

7.6 Limitations and strengths

Several limitations constrain the generalization of the present findings. First, although Projects 1-3 relied on a comparatively large and carefully characterized clinical sample, sample size still limits power for detecting subtle group-by-context interactions. The focus on women constrains inference to the broader BPD population. At the same time, this sampling strategy reflects the composition of many treatment-seeking cohorts and supports comparability with much of the existing literature. Relatedly, it remains possible that individuals with higher symptom load would show stronger deviations in the component processes examined here. Importantly, however, the present sample still represents a clinically valid and well-characterized group, and the observed patterns therefore remain informative for mechanistic assumptions.

Second, ecological validity is inherently limited by the abstraction of controlled tasks. This concern is particularly salient for baseline, anonymous one-shot paradigms, which deliberately remove relationship history, reputation concerns, and reciprocity dynamics. In this thesis, the design choice was strategic: baseline variants provide a necessary reference for isolating component processes before adding interpersonal meaning, uncertainty, and dynamic feedback. Nonetheless, it implies that key features of real-world relationships, such as ambiguity of intent, emotional salience, and trust repair, were not modeled here.

Third, neural inference is currently restricted. Only one project employed fMRI, which limits conclusions about neural generality across stages of the framework. Moreover, the lesion-based mechanistic comparison in Project 4 relies on a single Urbach-Wiehte patient, which necessarily constrains population-level inference and raises the possibility of compensatory development or lesion-specific idiosyncrasies. At the same time, Project 2 is among the few fMRI studies with an interactive social decision-making paradigm in BPD, and the Urbach-Wiethe case provides a unique opportunity to test mechanistic predictions about amygdala contributions to prosocial valuation.

Finally, while the projects align conceptually with a multi-stage framework, they do not yet provide a single empirical test in which the same individuals form partner representations about the same interaction partners, carry those beliefs into repeated interdependent cooperation, and subsequently make allocation decisions that translate interaction history into payoff allocations. As a result, cross-stage links are currently supported by convergent patterns across paradigms rather than by a unified design within a single task. Still, approaching the framework from multiple angles strengthens the overall account.

These limitations are balanced by several strengths. This thesis draws on converging evidence from multiple tasks and methods rather than relying on a single paradigm. Baseline behavior was established using standard economic games and explicit evaluation tasks, followed by a newly developed, repeated and interdependent cooperation task that isolates partner dependence, risk, and expectations, with additional computational and neural levels of analysis alongside a mechanistic lesion contrast. Together, this multi-method approach helps to explain mixed findings in cooperation tasks in BPD by assessing how different tasks recruit different processes, and it provides a structured basis for future within-participant cross-stage designs.

7.7 Future directions

An important next step is to operationalize the full staged framework into a single continuous paradigm. Participants should first learn about different partners, then interact repeatedly under varying dependence and risk, and finally allocate realized gains across multiple time points and different roles (allocator, receiver, joint decision-maker). Such designs would allow direct tests of the proposed cascade: whether negatively biased priors predict pessimistic expectations of partner behavior, whether these expectations prospectively predict reduced action engagement under varying constraints, and whether these processes translate into allocation behavior.

A second direction is increasing ecological validity through real partners or paradigms that allow reciprocity, learning, and repair. After establishing baseline behavior, richer interpersonal cues (e.g., direct communication, facial affect, explicit norm violation, and repair attempts) can be added in a controlled, stepwise manner to map when and why interactions destabilize.

A third direction is addressing heterogeneity using dimensional and symptom-based approaches (Herpertz et al., 2017; Insel et al., 2010). If negative priors, dependence-related pessimism, and context-insensitive engagement differ across symptom dimensions or maladaptive trait profiles, this would allow mechanistic subgrouping and prediction. However, this requires substantially larger samples and, ideally, multi-site designs to obtain stable estimates and to test moderation effects with sufficient power.

A fourth direction concerns cross-clinical inference. Project 4 highlights that pathology may shift depending on why people act prosocially rather than whether they act prosocially per se. Applying a

multi-stage framework across clinical groups with interpersonal dysfunctions could reveal both disorder-general and disorder-specific signatures. For example, whether the central alteration lies in priors, expectation formation, control engagement, motive weighting during allocation, or in the coupling between these components across stages.

Finally, given the context sensitivity of effects and the often modest effect sizes in low-arousal economic tasks, cumulative inference will benefit from open science practices, meta-analytic synthesis, and collaborative datasets that enable robust estimation, principled comparisons across paradigms, and tests of heterogeneity.

8 Conclusion

This thesis supports a multi-staged account of cooperation in which interpersonal dysfunction in BPD is *not* best explained by a generalized inability to cooperate or by globally atypical fairness preferences. Instead, the convergent pattern points to negatively biased self-evaluations and partner representations, paired with pessimistic expectations that become particularly salient under dependence on the partner. In addition, differences were most likely to emerge when situations provided fewer external constraints – when helping was not clearly required and decisions were more mixed-motive – probably because these contexts place greater demands on internally generated valuation and context-sensitive engagement.

A key implication is also encouraging: several core capacities relevant to cooperation appear preserved under clear contingencies, including broadly typical baseline social preferences and sensitivity to partner need. This suggests clinically relevant opportunities. Interventions may profit from targeting the initial stages and inputs of cooperation (like negative priors, pessimistic expectations, and the interpretation of ambiguous partner behavior), while leveraging intact learning potential and willingness to act prosocially when social structure is clear. Moving forward, I am confident that my work provides the basis for an informative agenda to test the full framework within dynamic, contingent interactions and to identify which symptom dimensions and individual differences determine where and why cooperative interactions become unstable.

References

Abramov, G., Miellet, S., Kautz, J., Grenyer, B. F. S., & Deane, F. P. (2020). The paradoxical decline and growth of trust as a function of borderline personality disorder trait count: Using discontinuous growth modelling to examine trust dynamics in response to violation and repair. *PLOS ONE*, *15*(7), e0236170. https://doi.org/10.1371/journal.pone.0236170

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *55*(3), 387–395. https://doi.org/10.1037/0022-3514.55.3.387

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268–277. https://doi.org/10.1038/nrn1884

Apps, M. A. J., Lockwood, P. L., & Balsters, J. H. (2013). The role of the midcingulate cortex in monitoring others' decisions. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00251

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*(4), 177–185. https://doi.org/10.1016/j.tics.2013.12.003

Asherson, P., Young, A. H., Eich-Höchli, D., Moran, P., Porsdal, V., & Deberdt, W. (2014). Differential diagnosis, comorbidity, and treatment of attention-deficit/hyperactivity disorder in relation to bipolar disorder or borderline personality disorder in adults. *Current Medical Research and Opinion*, *30*(8), 1657–1672. https://doi.org/10.1185/03007995.2014.915800

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & Van Der Maas, H. L. J. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Ball, J. S., & Links, P. S. (2009). Borderline personality disorder and childhood trauma: Evidence for a causal relationship. *Current Psychiatry Reports*, *11*(1), 63–68. https://doi.org/10.1007/s11920-009-0010-4

Barnow, S., Stopsack, M., Grabe, H. J., Meinke, C., Spitzer, C., Kronmüller, K., & Sieswerda, S. (2009). Interpersonal evaluation bias in borderline personality disorder. *Behaviour Research and Therapy*, *47*(5), 359–365. https://doi.org/10.1016/j.brat.2009.02.003

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bateman, A., & Fonagy, P. (2009). Randomized Controlled Trial of Outpatient Mentalization-Based Treatment Versus Structured Clinical Management for Borderline Personality Disorder. *American Journal of Psychiatry*, *166*(12), 1355–1364. https://doi.org/10.1176/appi.ajp.2009.09040539

Bertsch, K., Krauch, M., Stopfer, K., Haeussler, K., Herpertz, S. C., & Gamer, M. (2017). Interpersonal Threat Sensitivity in Borderline Personality Disorder: An Eye-Tracking Study. *Journal of Personality Disorders*, *31*(5), 647–670. https://doi.org/10.1521/pedi_2017_31_273

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: Social information processing in the human striatum. *WIREs Cognitive Science*, *5*(1), 61–73. https://doi.org/10.1002/wcs.1266

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*(4), 645–657. https://doi.org/10.1037/0022-3514.62.4.645

Borkenau, P., & Ostendorf, F. (2007). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (NEO-FFI; 2. Neu normierte und vollständig überarb. Aufl.)*. Hogrefe.

Bott, A., Brockmann, L., Denneberg, I., Henken, E., Kuper, N., Kruse, F., & Degner, J. (2024). Spontaneous Trait Inferences From Behavior: A Systematic Meta-Analysis. *Personality & Social Psychology Bulletin*, *50*(1), 78–102. https://doi.org/10.1177/01461672221100336

Bozzatello, P., Rocca, P., Baldassarri, L., Bosia, M., & Bellino, S. (2021). The Role of Trauma in Early Onset Borderline Personality Disorder: A Biopsychosocial Perspective. *Frontiers in Psychiatry*, *12*, 721361. https://doi.org/10.3389/fpsyt.2021.721361

Bray, S., Shimojo, S., & O'Doherty, J. P. (2010). Human Medial Orbitofrontal Cortex Is Recruited During Experience of Imagined and Real Rewards. *Journal of Neurophysiology*, *103*(5), 2506–2512. https://doi.org/10.1152/jn.01030.2009

Calvo, N., Valero, S., Sáez-Francàs, N., Gutiérrez, F., Casas, M., & Ferrer, M. (2016). Borderline Personality Disorder and Personality Inventory for DSM-5 (PID-5): Dimensional personality assessment with DSM-5. *Comprehensive Psychiatry*, *70*, 105–111. https://doi.org/10.1016/j.comppsych.2016.07.002

Camerer, C. F., & Fehr, E. (2004). 3 Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientist. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis (Eds.), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (p. 0). Oxford University Press. https://doi.org/10.1093/0199262055.003.0003

Cavicchioli, M., & Maffei, C. (2020). Rejection sensitivity in borderline personality disorder and the cognitive–affective personality system: A meta-analytic review. *Personality Disorders: Theory, Research, and Treatment*, *11*(1), 1–12. https://doi.org/10.1037/per0000359

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. *Neuron*, *70*(3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience*, *13*(6), 637–647. https://doi.org/10.1080/17470919.2018.1518834

Cho, J. C., & Knowles, E. D. (2013). I'm like you and you're like me: Social projection and self-stereotyping both help explain self–other correspondence. *Journal of Personality and Social Psychology*, *104*(3), 444–456. https://doi.org/10.1037/a0031017

Christopoulos, G. I., Tobler, P. N., Bossaerts, P., Dolan, R. J., & Schultz, W. (2009). Neural Correlates of Value, Risk, and Risk Aversion Contributing to Decision Making under Risk. *The Journal of Neuroscience*, *29*(40), 12574–12583. https://doi.org/10.1523/JNEUROSCI.2614-09.2009

Clarkin, J. F., Levy, K. N., Lenzenweger, M. F., & Kernberg, O. F. (2007). Evaluating Three Treatments for Borderline Personality Disorder: A Multiwave Study. *American Journal of Psychiatry*, *164*(6), 922–928. https://doi.org/10.1176/ajp.2007.164.6.922

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory Professional Manual*. Psychological Assessment Resources.

Cutler, J., & Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *NeuroImage*, *184*, 227–241. https://doi.org/10.1016/j.neuroimage.2018.09.009

Czekalla, N., Schröder, A., Mayer, A. V., Stierand, J., Stolz, D. S., Kube, T., Korn, C. W., Wilhelm-Groch, I., Klein, J. P., Paulus, F. M., Krach, S., & Müller-Pinzler, L. (2024). *Neurocomputational Mechanisms Underlying Maladaptive Self-Belief Formation in Depression*. Neuroscience. https://doi.org/10.1101/2024.05.09.593087

D'Adda, F., Sighinolfi, G., Mitolo, M., Scala, M., Guidi, L., Motta, L., Cirignotta, L., Manners, D. N., Tonon, C., Lodi, R., & Menchetti, M. (2025). Neurobiological correlates of personality dimensions in borderline personality disorder using graph analysis of functional connectivity. *Scientific Reports*, *15*(1), 12623. https://doi.org/10.1038/s41598-025-85989-x

Daros, A. R., Ruocco, A. C., & Rule, N. O. (2016). Identifying Mental Disorder from the Faces of Women with Borderline Personality Disorder. *Journal of Nonverbal Behavior*, *40*(4), 255–281. https://doi.org/10.1007/s10919-016-0237-9

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

De Panfilis, C., Riva, P., Preti, E., Cabrino, C., & Marchesi, C. (2015). When social inclusion is not enough: Implicit expectations of extreme inclusion in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *6*(4), 301–309. https://doi.org/10.1037/per0000132

De Panfilis, C., Schito, G., Generali, I., Gozzi, L. A., Ossola, P., Marchesi, C., & Grecucci, A. (2019). Emotions at the border: Increased punishment behavior during fair interpersonal exchanges in borderline personality disorder. *Journal of Abnormal Psychology*, *128*(2), 162–172. https://doi.org/10.1037/abn0000404

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: An fMRI investigation. *NeuroImage*, *23*(2), 744–751. https://doi.org/10.1016/j.neuroimage.2004.05.025

Degasperi, G., Cristea, I. A., Di Rosa, E., Costa, C., & Gentili, C. (2021). Parsing variability in borderline personality disorder: A meta-analysis of neuroimaging studies. *Translational Psychiatry*, *11*(1), 314. https://doi.org/10.1038/s41398-021-01446-z

Diekhof, E. K., Kaps, L., Falkai, P., & Gruber, O. (2012). The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude – An activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and outcome processing. *Neuropsychologia*, *50*(7), 1252–1266. https://doi.org/10.1016/j.neuropsychologia.2012.02.007

Domes, G., Schulze, L., & Herpertz, S. C. (2009). Emotion Recognition in Borderline Personality Disorder—A Review of the Literature. *Journal of Personality Disorders*, *23*(1), 6–19. https://doi.org/10.1521/pedi.2009.23.1.6

Domsalla, M., Koppe, G., Niedtfeld, I., Vollstädt-Klein, S., Schmahl, C., Bohus, M., & Lis, S. (2014). Cerebral processing of social rejection in patients with borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *9*(11), 1789–1797. https://doi.org/10.1093/scan/nst176

Donegan, N. H., Sanislow, C. A., Blumberg, H. P., Fulbright, R. K., Lacadie, C., Skudlarski, P., Gore, J. C., Olson, I. R., McGlashan, T. H., & Wexler, B. E. (2003). Amygdala hyperreactivity in borderline personality disorder: Implications for emotional dysregulation. *Biological Psychiatry*, *54*(11), 1284–1293. https://doi.org/10.1016/S0006-3223(03)00636-X

Elder, J. J., Davis, T. H., & Hughes, B. L. (2023). A Fluid Self-Concept: How the Brain Maintains Coherence and Positivity across an Interconnected Self-Concept While Incorporating Social Feedback. *The Journal of Neuroscience*, *43*(22), 4110–4128. https://doi.org/10.1523/JNEUROSCI.1951-22.2023

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Franzen, N., Hagenhoff, M., Baer, N., Schmidt, A., Mier, D., Sammer, G., Gallhofer, B., Kirsch, P., & Lis, S. (2011). Superior 'theory of mind' in borderline personality disorder: An analysis of interaction behavior in a virtual trust game. *Psychiatry Research*, *187*(1–2), 224–233. https://doi.org/10.1016/j.psychres.2010.11.012

Frith, C. D., & Frith, U. (2012). *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Frith, C. D., & Singer, T. (2008). The role of social cognition in decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3875–3886. https://doi.org/10.1098/rstb.2008.0156

Frolichs, K. M. M., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. *Nature Communications*, *13*(1), 6205. https://doi.org/10.1038/s41467-022-33418-2

Gangopadhyay, P., Chawla, M., Dal Monte, O., & Chang, S. W. C. (2021). Prefrontal–amygdala circuits in social decision-making. *Nature Neuroscience*, *24*(1), 5–18. https://doi.org/10.1038/s41593-020-00738-9

Grant, B. F., Chou, S. P., Goldstein, R. B., Huang, B., Stinson, F. S., Saha, T. D., Smith, S. M., Dawson, D. S., Pulay, A. J., Pickering, R. P., & Ruan, W. J. (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. *The Journal of Clinical Psychiatry*, *69*(4), 533–545. https://doi.org/10.4088/JCP.v69n0404

Graumann, L., Kulakova, E., Cho, A. B., Deuter, C. E., Wolf, O. T., Schell, J., Hellmann-Regen, J., Roepke, S., Otte, C., & Wingenfeld, K. (2025). Elevated testosterone and prosocial behavior in female patients with borderline personality disorder independent of social exclusion. *Psychoneuroendocrinology*, *171*, 107232. https://doi.org/10.1016/j.psyneuen.2024.107232

Grecucci, A., Giorgetta, C., Bonini, N., & Sanfey, A. G. (2013). Reappraising social emotions: The role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00523

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Gunderson, J. G. (2007). Disturbed Relationships as a Phenotype for Borderline Personality Disorder. *American Journal of Psychiatry*, *164*(11), 1637–1640. https://doi.org/10.1176/appi.ajp.2007.07071125

Gunderson, J. G. (2011). Ten-Year Course of Borderline Personality Disorder: Psychopathology and Function From the Collaborative Longitudinal Personality Disorders Study. *Archives of General Psychiatry*, *68*(8), 827. https://doi.org/10.1001/archgenpsychiatry.2011.37

Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nature Reviews Disease Primers*, *4*(1), 18029. https://doi.org/10.1038/nrdp.2018.29

Haber, S. N., & Knutson, B. (2010). The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology*, *35*(1), 4–26. https://doi.org/10.1038/npp.2009.129

Haruno, M., & Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature Neuroscience*, *13*(2), 160–161. https://doi.org/10.1038/nn.2468

Haruno, M., Kimura, M., & Frith, C. D. (2014). Activity in the Nucleus Accumbens and Amygdala Underlies Individual Differences in Prosocial and Individualistic Economic Choices. *Journal of Cognitive Neuroscience*, *26*(8), 1861–1870. https://doi.org/10.1162/jocn_a_00589

Hengartner, M. P., Ajdacic-Gross, V., Rodgers, S., Müller, M., & Rössler, W. (2013). Childhood adversity in association with personality disorder dimensions: New findings in an old debate. *European Psychiatry*, *28*(8), 476–482. https://doi.org/10.1016/j.eurpsy.2013.04.004

Hepp, J., Hilbig, B. E., Moshagen, M., Zettler, I., Schmahl, C., & Niedtfeld, I. (2014). Active versus reactive cooperativeness in Borderline Psychopathology: A dissection based on the HEXACO model of personality. *Personality and Individual Differences*, *56*, 19–23. https://doi.org/10.1016/j.paid.2013.08.013

Hepp, J., Kieslich, P. J., Schmitz, M., Schmahl, C., & Niedtfeld, I. (2021). Negativity on two sides: Individuals with borderline personality disorder form negative first impressions of others and are perceived negatively by them. *Personality Disorders: Theory, Research, and Treatment*, *12*(6), 514–525. https://doi.org/10.1037/per0000412

Hepp, J., Kieslich, P. J., Wycoff, A. M., Bertsch, K., Schmahl, C., & Niedtfeld, I. (2021). Mouse-tracking reveals cognitive conflict during negative impression formation in women with Borderline Personality Disorder or Social Anxiety Disorder. *PLOS ONE*, *16*(3), e0247955. https://doi.org/10.1371/journal.pone.0247955

Hepp, J., Störkel, L. M., Kieslich, P. J., Schmahl, C., & Niedtfeld, I. (2018). Negative evaluation of individuals with borderline personality disorder at zero acquaintance. *Behaviour Research and Therapy*, *111*, 84–91. https://doi.org/10.1016/j.brat.2018.09.009

Herpertz, S. C., Huprich, S. K., Bohus, M., Chanen, A., Goodman, M., Mehlum, L., Moran, P., Newton-Howes, G., Scott, L., & Sharp, C. (2017). The Challenge of Transforming the Diagnostic System of Personality Disorders. *Journal of Personality Disorders*, *31*(5), 577–589. https://doi.org/10.1521/pedi_2017_31_338

Herpertz, S. C., Jeung, H., Mancke, F., & Bertsch, K. (2014). Social Dysfunctioning and Brain in Borderline Personality Disorder. *Psychopathology*, *47*(6), 417–424. https://doi.org/10.1159/000365106

Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, *54*(5), 598–603. https://doi.org/10.1016/j.paid.2012.11.008

Hiser, J., & Koenigs, M. (2018). The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. *Biological Psychiatry*, *83*(8), 638–647. https://doi.org/10.1016/j.biopsych.2017.10.030

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLOS Computational Biology*, *14*(2), e1005935. https://doi.org/10.1371/journal.pcbi.1005935

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jaksic, C., & Schlegel, K. (2020). Accuracy in Judging Others' Personalities: The Role of Emotion Recognition, Emotion Understanding, and Trait Emotional Intelligence. *Journal of Intelligence*, *8*(3), 34. https://doi.org/10.3390/jintelligence8030034

Jeung, H., Schwieren, C., & Herpertz, S. C. (2016). Rationality and self-interest as economic-exchange strategy in borderline personality disorder: Game theory, social preferences, and interpersonal behavior. *Neuroscience & Biobehavioral Reviews*, *71*, 849–864. https://doi.org/10.1016/j.neubiorev.2016.10.030

Jeung, H., Vollmann, M., Herpertz, S. C., & Schwieren, C. (2020). Consider others better than yourself: Social decision-making and partner preference in Borderline Personality Disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *67*, 101436. https://doi.org/10.1016/j.jbtep.2018.11.004

Jeung, H., Walther, S., Korn, C. W., Bertsch, K., & Herpertz, S. C. (2018). Emotional responses to receiving peer feedback on opinions in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *9*(6), 595–600. https://doi.org/10.1037/per0000292

Johnson, J. G., Cohen, P., Kasen, S., Skodol, A. E., & Oldham, J. M. (2008). Cumulative prevalence of personality disorders between adolescence and adulthood. *Acta Psychiatrica Scandinavica*, *118*(5), 410–413. https://doi.org/10.1111/j.1600-0447.2008.01231.x

Juan, C.-H., & Muggleton, N. G. (2012). Brain stimulation and inhibitory control. *Brain Stimulation*, *5*(2), 63–69. https://doi.org/10.1016/j.brs.2012.03.012

Kandler, C., & Bleidorn, W. (2015). Personality Differences and Development: Genetic and Environmental Contributions. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 884–890). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.25011-3

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The Rupture and Repair of Cooperation in Borderline Personality Disorder. *Science*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

Koenigsberg, H. W., Siever, L. J., Lee, H., Pizzarello, S., New, A. S., Goodman, M., Cheng, H., Flory, J., & Prohovnik, I. (2009). Neural correlates of emotion processing in borderline personality disorder. *Psychiatry Research: Neuroimaging*, *172*(3), 192–199. https://doi.org/10.1016/j.pscychresns.2008.07.010

Kopala-Sibley, D. C., Zuroff, D. C., Russell, J. J., Moskowitz, D. S., & Paris, J. (2012). Understanding heterogeneity in borderline personality disorder: Differences in affective reactivity explained by the traits of dependency and self-criticism. *Journal of Abnormal Psychology*, *121*(3), 680–691. https://doi.org/10.1037/a0028513

Korn, C. W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine*, *46*(3), 575–587. https://doi.org/10.1017/S003329171500207X

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Korzekwa, M. I., Dell, P. F., Links, P. S., Thabane, L., & Webb, S. P. (2008). Estimating the prevalence of borderline personality disorder in psychiatric outpatients using a two-phase procedure. *Comprehensive Psychiatry*, *49*(4), 380–386. https://doi.org/10.1016/j.comppsych.2008.01.007

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine*, *42*(9), 1879–1890. https://doi.org/10.1017/S0033291711002674

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lazarus, S. A., Cheavens, J. S., Festa, F., & Zachary Rosenthal, M. (2014). Interpersonal functioning in borderline personality disorder: A systematic review of behavioral and laboratory-based assessments. *Clinical Psychology Review*, *34*(3), 193–205. https://doi.org/10.1016/j.cpr.2014.01.007

Leary, M. R. (2007). Motivational and Emotional Aspects of the Self. *Annual Review of Psychology*, *58*(1), 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Lee, V. K., & Harris, L. T. (2013). How social cognition can inform social decision making. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00259

Lenzenweger, M. F., Lane, M. C., Loranger, A. W., & Kessler, R. C. (2007). DSM-IV Personality Disorders in the National Comorbidity Survey Replication. *Biological Psychiatry*, *62*(6), 553–564. https://doi.org/10.1016/j.biopsych.2006.09.019

Lévay, E. E., Bajzát, B., & Unoka, Z. S. (2021). Expectation of Selfishness From Others in Borderline Personality Disorder. *Frontiers in Psychology*, *12*, 702227. https://doi.org/10.3389/fpsyg.2021.702227

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Liebenow, B., Jones, R., DiMarco, E., Trattner, J. D., Humphries, J., Sands, L. P., Spry, K. P., Johnson, C. K., Farkas, E. B., Jiang, A., & Kishida, K. T. (2022). Computational reinforcement learning, reward (and punishment), and dopamine in psychiatric disorders. *Frontiers in Psychiatry*, *13*, 886297. https://doi.org/10.3389/fpsyt.2022.886297

Liebke, L., Bungert, M., Thome, J., Hauschild, S., Gescher, D. M., Schmahl, C., Bohus, M., & Lis, S. (2017). Loneliness, social networks, and social functioning in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *8*(4), 349–356. https://doi.org/10.1037/per0000208

Liebke, L., Koppe, G., Bungert, M., Thome, J., Hauschild, S., Defiebre, N., Izurieta Hidalgo, N. A., Schmahl, C., Bohus, M., & Lis, S. (2018). Difficulties with being socially accepted: An experimental study in borderline personality disorder. *Journal of Abnormal Psychology*, *127*(7), 670–682. https://doi.org/10.1037/abn0000373

Linehan, M. M., Korslund, K. E., Harned, M. S., Gallop, R. J., Lungu, A., Neacsiu, A. D., McDavid, J., Comtois, K. A., & Murray-Gregory, A. M. (2015). Dialectical Behavior Therapy for High Suicide Risk in Individuals With Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry*, *72*(5), 475. https://doi.org/10.1001/jamapsychiatry.2014.3039

Lis, S., Schaedler, A., Liebke, L., Hauschild, S., Thome, J., Schmahl, C., Stahlberg, D., Kleindienst, N., & Bohus, M. (2018). Borderline Personality Disorder Features and Sensitivity to Injustice. *Journal of Personality Disorders*, *32*(2), 192–206. https://doi.org/10.1521/pedi_2017_31_292

Lischke, A., Herpertz, S. C., Berger, C., Domes, G., & Gamer, M. (2017). Divergent effects of oxytocin on (para-)limbic reactivity to emotional and neutral scenes in females with and without borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *12*(11), 1783–1792. https://doi.org/10.1093/scan/nsx107

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, *98*(1), 111–126. https://doi.org/10.1348/000712606X109648

Liu, Y., Li, S., Lin, W., Li, W., Yan, X., Wang, X., Pan, X., Rutledge, R. B., & Ma, Y. (2019). Oxytocin modulates social value representations in the amygdala. *Nature Neuroscience*, *22*(4), 633–641. https://doi.org/10.1038/s41593-019-0351-1

Lockwood, P. L., & Klein-Flügge, M. C. (2020). Computational modelling of social cognition and behaviour—A reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, nsaa040. https://doi.org/10.1093/scan/nsaa040

MacCrimmon, K. R., & Messick, D. M. (1976). A framework for social motives. *Behavioral Science*, *21*(2), 86–100. https://doi.org/10.1002/bs.3830210203

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. https://doi.org/10.1038/nn.2723

Masland, S. R., & Hooley, J. M. (2020). When Trust Does Not Come Easily: Negative Emotional Information Unduly Influences Trustworthiness Appraisals for Individuals With Borderline

Personality Features. *Journal of Personality Disorders*, *34*(3), 394–409. https://doi.org/10.1521/pedi_2019_33_404

Masland, S. R., Schnell, S. E., & Shah, T. V. (2020). Trust Beliefs, Biases, and Behaviors in Borderline Personality Disorder: Empirical Findings and Relevance to Epistemic Trust. *Current Behavioral Neuroscience Reports*, *7*(4), 239–249. https://doi.org/10.1007/s40473-020-00220-7

Mauchnik, J., & Schmahl, C. (2010). The Latest Neuroimaging Findings in Borderline Personality Disorder. *Current Psychiatry Reports*, *12*(1), 46–55. https://doi.org/10.1007/s11920-009-0089-7

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835. https://doi.org/10.1073/pnas.211415698

Menon, V., Adleman, N. E., White, C. D., Glover, G. H., & Reiss, A. L. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, *12*(3), 131–143. https://doi.org/10.1002/1097-0193(200103)12:3%253C131::AID-HBM1010%253E3.0.CO;2-C

Mier, D., Lis, S., Esslinger, C., Sauer, C., Hagenhoff, M., Ulferts, J., Gallhofer, B., & Kirsch, P. (2013). Neuronal correlates of social cognition in borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *8*(5), 531–537. https://doi.org/10.1093/scan/nss028

Minzenberg, M. J., Fan, J., New, A. S., Tang, C. Y., & Siever, L. J. (2007). Fronto-limbic dysfunction in response to facial emotion in borderline personality disorder: An event-related fMRI study. *Psychiatry Research: Neuroimaging*, *155*(3), 231–243. https://doi.org/10.1016/j.pscychresns.2007.03.006

Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage*, *112*, 244–253. https://doi.org/10.1016/j.neuroimage.2014.12.056

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring Social Value Orientation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1804189

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality Judgments Based on Physical Appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. https://doi.org/10.1177/0146167209346309

Niedtfeld, I., & Kroneisen, M. (2020). Impaired memory for cooperative interaction partners in borderline personality disorder. *Borderline Personality Disorder and Emotion Dysregulation*, *7*(1), 22. https://doi.org/10.1186/s40479-020-00137-3

Niedtfeld, I., Renkewitz, F., Mädebach, A., Hillmann, K., Kleindienst, N., Schmahl, C., & Schulze, L. (2020). Enhanced memory for negative social information in borderline personality disorder. *Journal of Abnormal Psychology*, *129*(5), 480–491. https://doi.org/10.1037/abn0000540

Niv, Y., & Langdon, A. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67–73. https://doi.org/10.1016/j.cobeha.2016.04.005

Oltmanns, T. F., Friedman, J. N. W., Fiedler, E. R., & Turkheimer, E. (2004). Perceptions of people with personality disorders based on thin slices of behavior. *Journal of Research in Personality*, *38*(3), 216–229. https://doi.org/10.1016/S0092-6566(03)00066-7

Pagura, J., Stein, M. B., Bolton, J. M., Cox, B. J., Grant, B., & Sareen, J. (2010). Comorbidity of borderline personality disorder and posttraumatic stress disorder in the U.S. population. *Journal of Psychiatric Research*, *44*(16), 1190–1198. https://doi.org/10.1016/j.jpsychires.2010.04.016

Peters, J., & Büchel, C. (2010). Neural representations of subjective reward value. *Behavioural Brain Research*, *213*(2), 135–141. https://doi.org/10.1016/j.bbr.2010.04.031

Pike, A. C., & Robinson, O. J. (2022). Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *79*(4), 313. https://doi.org/10.1001/jamapsychiatry.2022.0051

Pisor, A. C., Gervais, M. M., Purzycki, B. G., & Ross, C. T. (2020). Preferences and constraints: The value of economic games for studying human behaviour. *Royal Society Open Science*, *7*(6), 192090. https://doi.org/10.1098/rsos.192090

Poggi, A., Richetin, J., & Preti, E. (2019). Trust and Rejection Sensitivity in Personality Disorders. *Current Psychiatry Reports*, *21*(8), 69. https://doi.org/10.1007/s11920-019-1059-3

Preuss, N., Brändle, L. S., Hager, O. M., Haynes, M., Fischbacher, U., & Hasler, G. (2016). Inconsistency and social decision making in patients with Borderline Personality Disorder. *Psychiatry Research*, *243*, 115–122. https://doi.org/10.1016/j.psychres.2016.06.017

Reiter, A. M. F., Heinze, H.-J., Schlagenhauf, F., & Deserno, L. (2017). Impaired Flexible Reward-Based Decision-Making in Binge Eating Disorder: Evidence from Computational Modeling and Functional Neuroimaging. *Neuropsychopharmacology*, *42*(3), 628–637. https://doi.org/10.1038/npp.2016.95

Renneberg, B., Herm, K., Hahn, A., Staebler, K., Lammers, C., & Roepke, S. (2012). Perception of Social Participation in Borderline Personality Disorder. *Clinical Psychology & Psychotherapy*, *19*(6), 473–480. https://doi.org/10.1002/cpp.772

Rosenblau, G., Frolichs, K., & Korn, C. W. (2023). A neuro-computational social learning framework to facilitate transdiagnostic classification and treatment across psychiatric disorders. *Neuroscience & Biobehavioral Reviews*, *149*, 105181. https://doi.org/10.1016/j.neubiorev.2023.105181

Rosenblau, G., Korn, C. W., Dutton, A., Lee, D., & Pelphrey, K. A. (2021). Neurocognitive Mechanisms of Social Inferences in Typical and Autistic Adolescents. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(8), 782–791. https://doi.org/10.1016/j.bpsc.2020.07.002

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562. https://doi.org/10.1038/nrn3776

Samuel, D., & Widiger, T. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis☆. *Clinical Psychology Review*, *28*(8), 1326–1342. https://doi.org/10.1016/j.cpr.2008.07.002

Saulsman, L. M., & Page, A. C. (2004). The five-factor model and personality disorder empirical literature: A meta-analytic review. *Clinical Psychology Review*, *23*(8), 1055–1085. https://doi.org/10.1016/j.cpr.2002.09.001

Saunders, K. E. A., Goodwin, G. M., & Rogers, R. D. (2015). Borderline personality disorder, but not euthymic bipolar disorder, is associated with a failure to sustain reciprocal cooperative behaviour: Implications for spectrum models of mood disorders. *Psychological Medicine*, *45*(8), 1591–1600. https://doi.org/10.1017/S0033291714002475

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Scheunemann, J., Jelinek, L., Biedermann, S. V., Lipp, M., Yassari, A. H., Kühn, S., Gallinat, J., & Moritz, S. (2023). Can you trust this source? Advice taking in borderline personality disorder. *European Archives of Psychiatry and Clinical Neuroscience*, *273*(4), 875–885. https://doi.org/10.1007/s00406-022-01539-w

Schuermann, B., Kathmann, N., Stiglmayr, C., Renneberg, B., & Endrass, T. (2011). Impaired decision making and feedback evaluation in borderline personality disorder. *Psychological Medicine*, *41*(9), 1917–1927. https://doi.org/10.1017/S003329171000262X

Schulze, L., Schulze, A., Renneberg, B., Schmahl, C., & Niedtfeld, I. (2019). Neural Correlates of Affective Disturbances: A Comparative Meta-analysis of Negative Affect Processing in Borderline Personality Disorder, Major Depressive Disorder, and Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(3), 220–232. https://doi.org/10.1016/j.bpsc.2018.11.004

Schurz, M., Berenz, J.-P., Maerz, J., Perla, R., Buchheim, A., & Labek, K. (2024). Brain Activation for Social Cognition and Emotion Processing Tasks in Borderline Personality Disorder: A Meta-Analysis of Neuroimaging Studies. *Brain Sciences*, *14*(4), 395. https://doi.org/10.3390/brainsci14040395

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303

Seres, I., Unoka, Z., & K\' eri, S. (2009). The broken trust and cooperation in borderline personality disorder. *NeuroReport*, *20*(4), 388–392. https://doi.org/10.1097/wnr.0b013e328324eb4d

Staebler, K., Renneberg, B., Stopsack, M., Fiedler, P., Weiler, M., & Roepke, S. (2011). Facial emotional expression in reaction to social exclusion in borderline personality disorder. *Psychological Medicine*, *41*(9), 1929–1938. https://doi.org/10.1017/S0033291711000080

Stepp, S. D., Pilkonis, P. A., Yaggi, K. E., Morse, J. Q., & Feske, U. (2009). Interpersonal and Emotional Experiences of Social Interactions in Borderline Personality Disorder. *Journal of Nervous & Mental Disease*, *197*(7), 484–491. https://doi.org/10.1097/NMD.0b013e3181aad2e7

Suppes, P. (1966). Models of Data. In *Studies in Logic and the Foundations of Mathematics* (Vol. 44, pp. 252–261). Elsevier. https://doi.org/10.1016/S0049-237X(09)70592-0

Sutton, R. S., & Barto, A. (2014). *Reinforcement learning: An introduction* (Nachdruck). The MIT Press.

Tebartz Van Elst, L., Hesslinger, B., Thiel, T., Geiger, E., Haegele, K., Lemieux, L., Lieb, K., Bohus, M., Hennig, J., & Ebert, D. (2003). Frontolimbic brain abnormalities in patients with borderline personality disorder. *Biological Psychiatry*, *54*(2), 163–171. https://doi.org/10.1016/S0006-3223(02)01743-2

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic Games: An Introduction and Guide for Research. *Collabra: Psychology*, *7*(1), 19004. https://doi.org/10.1525/collabra.19004

Thielmann, I., Hilbig, B. E., & Niedtfeld, I. (2014). Willing to Give but Not to Forgive: Borderline Personality Features and Cooperative Behavior. *Journal of Personality Disorders*, *28*(6), 778–795. https://doi.org/10.1521/pedi_2014_28_135

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. https://doi.org/10.1037/bul0000217

Thome, J., Liebke, L., Bungert, M., Schmahl, C., Domes, G., Bohus, M., & Lis, S. (2016). Confidence in facial emotion recognition in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *7*(2), 159–168. https://doi.org/10.1037/per0000142

Tomko, R. L., Trull, T. J., Wood, P. K., & Sher, K. J. (2014). Characteristics of Borderline Personality Disorder in a Community Sample: Comorbidity, Treatment Utilization, and General Functioning. *Journal of Personality Disorders*, *28*(5), 734–750. https://doi.org/10.1521/pedi_2012_26_093

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The Prevalence of Personality Disorders in a Community Sample. *Archives of General Psychiatry*, *58*(6), 590. https://doi.org/10.1001/archpsyc.58.6.590

Torgersen, S., Lygren, S., Øien, P. A., Skre, I., Onstad, S., Edvardsen, J., Tambs, K., & Kringlen, E. (2000). A twin study of personality disorders. *Comprehensive Psychiatry*, *41*(6), 416–425. https://doi.org/10.1053/comp.2000.16560

Tusche, A., & Bas, L. M. (2021). Neurocomputational models of altruistic decision-making and social motives: Advances, pitfalls, and future directions. *WIREs Cognitive Science*, *12*(6), e1571. https://doi.org/10.1002/wcs.1571

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*(4), 399–409. https://doi.org/10.1521/pedi.2009.23.4.399

Van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

Van Schie, C. C., Chiu, C.-D., Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2020). Stuck in a negative me: fMRI study on the role of disturbed self-views in social feedback processing in borderline personality disorder. *Psychological Medicine*, *50*(4), 625–635. https://doi.org/10.1017/S0033291719000448

Van Schie, C. C., Whiting, L., & Grenyer, B. F. S. (2024). How negative self-views may interfere with building positive relationships: An experimental analogue of identity dysfunction in borderline personality disorder. *PLOS ONE*, *19*(3), e0301196. https://doi.org/10.1371/journal.pone.0301196

Vater, A., Schröder-Abé, M., Weißgerber, S., Roepke, S., & Schütz, A. (2015). Self-concept structure and borderline personality disorder: Evidence for negative compartmentalization. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 50–58. https://doi.org/10.1016/j.jbtep.2014.08.003

Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A., & Ochsner, K. N. (2008). Prefrontal-Subcortical Pathways Mediating Successful Emotion Regulation. *Neuron*, *59*(6), 1037–1050. https://doi.org/10.1016/j.neuron.2008.09.006

Waite, E. E., Savalia, T., Cohen, A. L., Haliczer, L. A., Huffman, S., & Dixon-Gordon, K. L. (2024). Borderline personality disorder and learning: The influences of emotional state and social versus nonsocial feedback. *Journal of Affective Disorders*, *363*, 474–482. https://doi.org/10.1016/j.jad.2024.07.072

Weinbrecht, A., Niedeggen, M., Roepke, S., & Renneberg, B. (2018). Feeling excluded no matter what? Bias in the processing of social participation in borderline personality disorder. *NeuroImage: Clinical*, *19*, 343–350. https://doi.org/10.1016/j.nicl.2018.04.031

Werner, A. M., Tibubos, A. N., Rohrmann, S., & Reiss, N. (2019). The clinical trait self-criticism and its relation to psychopathology: A systematic review – Update. *Journal of Affective Disorders*, *246*, 530–547. https://doi.org/10.1016/j.jad.2018.12.069

Widiger, T. A., & Weissman, M. M. (1991). Epidemiology of Borderline Personality Disorder. *Psychiatric Services*, *42*(10), 1015–1021. https://doi.org/10.1176/ps.42.10.1015

Winter, D., Bohus, M., & Lis, S. (2017). Understanding Negative Self-Evaluations in Borderline Personality Disorder—A Review of Self-Related Cognitions, Emotions, and Motives. *Current Psychiatry Reports*, *19*(3), 17. https://doi.org/10.1007/s11920-017-0771-0

Wischniewski, J., & Brüne, M. (2013). How Do People With Borderline Personality Disorder Respond to Norm Violations? Impact of Personality Factors on Economic Decision-Making. *Journal of Personality Disorders*, *27*(4), 531–546. https://doi.org/10.1521/pedi_2012_26_036

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology*, *99*(1), 174–190. https://doi.org/10.1037/a0019390

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089

Zimmerman, M., & Mattia, J. I. (1999). Axis I diagnostic comorbidity and borderline personality disorder. *Comprehensive Psychiatry*, *40*(4), 245–252. https://doi.org/10.1016/S0010-440X(99)90123-2

Zimmermann, J., Altenstein, D., Krieger, T., Holtforth, M. G., Pretsch, J., Alexopoulos, J., Spitzer, C., Benecke, C., Krueger, R. F., Markon, K. E., & Leising, D. (2014). The Structure and Correlates of Self-Reported DSM-5 Maladaptive Personality Traits: Findings From Two German-Speaking Samples. *Journal of Personality Disorders*, *28*(4), 518–540. https://doi.org/10.1521/pedi_2014_28_130

Zoh, Y., Chang, S. W. C., & Crockett, M. J. (2022). The prefrontal cortex and (uniquely) human cooperation: A comparative perspective. *Neuropsychopharmacology*, *47*(1), 119–133. https://doi.org/10.1038/s41386-021-01092-5

Erklärung des Eigenanteils an den Publikationen

Studie I:

LMD sammelte die Daten, führte die statistischen Analysen durch und verfasste den Originalentwurf des Manuskripts. RP führte die Modellierungsanalysen durch. KMMF konzipierte die Studie und führte Datenanalysen durch. GR verfasste den Originalentwurf und betreute das Projekt. SCH leistete wesentliche fachliche Unterstützung im Umgang mit der klinischen Stichprobe und überarbeitete das Manuskript. CWK konzipierte die Studie, verfasste den Originalentwurf, betreute das Projekt und verantwortete die Drittmittelakquise. Alle Autorinnen und Autoren prüften und überarbeiteten das Manuskript und genehmigten die finale Version zur Publikation.

Studie III:

LMD und CWK konzipierten die Studie und verfassten den Originalentwurf des Manuskripts. LMD sammelte die Daten und führte die Datenanalysen durch. CWK betreute das Projekt, verantwortete die Drittmittelakquise und unterstützte bei den Datenanalysen. JL sammelte Daten und verfasste Teile des Originalentwurfs. CN und SCH unterstützten die Entwicklung des Studiendesigns, leisteten wesentliche Unterstützung im Umgang mit der klinischen Stichprobe und überarbeiteten das Manuskript. Alle Autorinnen und Autoren prüften und überarbeiteten das Manuskript und genehmigten die finale Version zur Publikation.

Studie IV:

LMD analysierte die Daten und verfasste den Originalentwurf des Manuskripts. CWK analysierte die Daten. CWK, DRB und RH planten die Experimente und erhoben die Daten. Alle Autorinnen und Autoren prüften und überarbeiteten das Manuskript und genehmigten die finale Version zur Publikation.

Verwendung KI-basierter Tools:

Generative künstlichen Intelligenz (KI) Werkzeuge wurden in dieser Arbeit ausschließlich zur Unterstützung bei sprachlicher Überarbeitung und beim Korrekturlesen eingesetzt. Konkret wurden sie verwendet, um grammatikalische Fehler zu identifizieren, Formulierungen im Sinne einer besseren Verständlichkeit zu optimieren sowie Kohärenz, Konsistenz und Einheitlichkeit im gesamten Manuskript sicherzustellen. KI-Tools wurden nicht zur Generierung von Ideen, zur Konzeption der Studien, zur Datenanalyse, zur Interpretation von Ergebnissen oder zur Erstellung wissenschaftlicher Aussagen eingesetzt.

Acknowledgments

Abschließend möchte ich mich bei all den Menschen bedanken, die mich in den letzten Jahren begleitet und unterstützt haben.

Mein besonderer Dank gilt zuerst Christoph Korn, meinem Supervisor. Lieber Christoph, danke, dass du mir diese Möglichkeit eröffnet hast. Danke für deine Unterstützung, dein Verständnis und dein Vertrauen und besonders für deinen Enthusiasmus und deinen grenzenlosen Optimismus.

Ich danke allen Patientinnen und Probandinnen, die an diesem Projekt teilgenommen haben für ihre Zeit und ihr Vertrauen.

Thank you to all current and former members of my lab for the support, feedback, and great discussions, as well as the fun breaks: Ben, Koen, Sihui, Sergej, Ismail, Xiong, Raphael, Jiatong, Moritz, Mariana, Franny, Josué, Heike and Gabriela.

A special thank you to the inspiring people I met along this academic journey and am incredibly lucky to call my friends: Lena, Koen, Mana, Sihui, Elena, Hanna, and Alina.

Ich möchte mich außerdem bei allen bedanken, die mein Leben in Heidelberg so viel leichter und schöner machen: Steffi, Corinna, Chrissy, Isabelle, Maya, und Char.

Danke an meine Liebsten in Wien, die dafür sorgen, dass sich Wien immer wie zuhause anfühlt: Dani, Lena, Verena, Raphi und Sabine.

Danke an meine Familie für den Rückhalt, besonders an meine Omas und Ingrid, Gabi und Hef.

Und zuletzt: Alina, Mama und Papa, danke, dass ihr immer für mich da seid und mir zeigt, was wirklich wichtig ist.

Lebenslauf aus datenschutzrechtlichen Gründen nicht enthalten

Lebenslauf aus datenschutzrechtlichen Gründen nicht enthalten

Lebenslauf aus datenschutzrechtlichen Gründen nicht enthalten

Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe, insbesondere ohne entgeltliche Hilfe von Vermittlungs- und Beratungsdiensten, verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe. Das gilt insbesondere auch für alle Informationen aus Internetquellen.

Soweit beim Verfassen der Dissertation KI-basierte Tools („Chatbots“) verwendet wurden, versichere ich ausdrücklich, den daraus generierten Anteil deutlich kenntlich gemacht zu haben. Die „Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG“ aus September 2023 wurde dabei beachtet.

Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.

Ich erkläre mich damit einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Datum Unterschrift

Supplementary material for all studies

Study I - Supplementary material

Supplementary material

Supplementary methods

Power analysis

We conducted a power analysis using G*Power version 3.1.9.2 (1) to estimate the required sample size, based on a previous study (2) on group differences in updating undesirable feedback in individuals with BPD (N = 22) and controls (N = 42). The effect size reported in Korn et al. (2) was d = 0.65. Using an alpha level of .05 and a power of 1-β = .80, the analysis indicated a minimum sample size of N = 30 per group for a one-tailed two-sample t-test.

Participants

Distribution of BPD criteria according to the IPDE

In the BPD group, endorsement was highest for recurrent suicidal behavior/self-harm (90%), followed by impulsivity (80%), unstable and intense relationships (73.33%), affective instability (76.67%), chronic feelings of emptiness (73.33%). Moderate endorsement appeared for stress-related paranoid ideation or dissociation (60%), frantic efforts to avoid abandonment (56.67%), and identity disturbance (56.57%), while inappropriate/intense anger was least frequent (26.67%). In this sample the controls showed no endorsement (0%) for any of the nine criteria.

Table S1 Comorbid diagnoses of the BPD group according to the mini-DIPS diagnostic interview

Comorbid Diagnosis	*N*	**%**
Panic disorder	4	13.33
Agoraphobia with panic	1	3.33
Specific phobia	5	16.67
Social anxiety disorder	13	43.33
Major depressive episode	2	6.67
Major depressive disorder, recurrent	15	50
Persistent depressive disorder	3	10
Obsessive-compulsive disorder	1	3.33
Body dysmorphic disorder	1	3.33
Posttraumatic stress disorder	10	33.33
Bulimia nervosa	1	3.33
Binge eating disorder	2	6.67
Illness anxiety disorder	1	3.33
Substance abuse, alcohol	1	3.33
Substance abuse, cannabis	2	6.67
No axis I diagnosis	4	13.33

Table S2 Psychotropic medication intake of the BPD group

Medication	*N*	**%**
Antidepressant, SNRI	4	13.33
Antidepressant, SSRI	10	33.33
Atypical antipsychotics	3	10
Mood stabilizer (Lamotrigine)	2	6.67
No medication	14	46.67

Task instructions and additional information

Task instructions for the personality learning task (other-ratings)

The following task instructions are translated from the German on-screen instructions:

In this task, you are supposed to evaluate six different people. Imagine that you could meet these people in the future. First, you will see a short description of a person. After that, you should evaluate the person on various personality adjectives. After a fixation cross, you will see how the person has evaluated themselves. You can then compare your evaluation with the person's self-evaluation.

Please give your rating on a scale from 1 to 8.

1 means: The trait does NOT apply at all.

8 means: The trait applies VERY STRONGLY.

At the beginning, of course, you know almost nothing about the person. Over time, however, you may get a feeling for how the person generally evaluates themselves.

You have 4 seconds for each answer; after that, the program will continue.

Additional information on the profiles

Names used: Emma, Hannah, Sophia, Marie, Lena, Julia

Ages used: 22, 24, 26, 27, 31, 33

Task instructions for the self-ratings

The following task instructions are translated from the German on-screen instructions:

Now, please evaluate yourself. You have 4 seconds for each answer; after that, the program will continue.

Table S3 The 40 trait items used in the personality trait learning task

Item number	German trait words used	Approximate English translation	Valence
Factor 1: Neuroticism			
1	gelassen	composed	positive
2	locker	easy-going	positive
3	selbstständig	self-reliant	positive
4	souverän	confident	positive
5	ängstlich	anxious	negative
6	launisch	moody	negative
7	unentschlossen	Indecisive	negative
8	wehleidig	whiny	negative
Factor 2: Extraversion			
9	enthusiastisch	enthusiastic	positive
10	lebenslustig	fun-loving	positive
11	schlagfertig	articulate	positive
12	tatkräftig	dynamic	positive
13	humorlos	humorless	negative
14	kalt	cold-hearted	negative
15	scheu	unassertive	negative
16	unnahbar	inapproachable	negative
Factor 3: Openness			
17	kreativ	creative	positive
18	spontan	spontaneous	positive
19	tolerant	tolerant	positive
20	vielseitig	versatile	positive
21	engstirnig	narrow-minded	negative

22	träge	lazy	negative
23	voreingenommen	biased	negative
24	freundlich	friendly	negative
Factor 4: Agreeableness			
25	großzügig	generous	positive
26	hilfsbereit	helpful	positive
27	höflich	polite	positive
28	arrogant	arrogant	negative
29	egoistisch	selfish	negative
30	hinterhältig	conniving	negative
31	stur	stubborn	negative
32	unsympathisch	unpleasant	negative
Factor 5: Conscientiousness			
33	aufrichtig	honest	positive
34	bescheiden	modest	positive
35	fleißig	hard-working	positive
36	ordentlich	tidy	positive
37	leichtsinnig	foolhardy	negative
38	pedantisch	pedantic	negative
39	unpünktlich	tardy	negative
40	zwanghaft	obsessive	negative

Inclusion and exclusion procedure

We recruited female participants aged 18-40 who were fluent in German. Our preregistered criteria initially required individuals in the BPD group to meet at least five BPD criteria based on the IPDE interview. However, during recruitment, we found that about 20% of the participants with a formal BPD diagnosis did not meet this threshold. To ensure feasibility of data collection and enhance the clinical generalizability of our results, we expanded our inclusion criteria to include sub-threshold BPD participants meeting four IPDE criteria.

Exclusion criteria included severe neurological or physical illness, BMI < 15, severe visual impairments, cognitive disorders, a lifetime diagnosis of schizophrenia, schizoaffective or bipolar disorder, current narcissistic or avoidant personality disorder, acute suicidality, and being a psychology student (except first semester). Additional exclusions for BPD participants were moderate to severe alcohol or drug dependence and psychotropic medication use (except antidepressants or anxiolytics). However, due to challenges in recruiting enough eligible participants, we broadened our inclusion to individuals who received Quetiapine prescriptions to improve sleeping (up to 100 mg extended-release and 50 mg immediate-release) and on-demand medication for the BPD group. For controls, exclusions included current alcohol or drug dependence, any lifetime personality disorder, current or past psychiatric diagnoses or interventions (within the last 5 years), use of psychotropic medication, more than one BPD symptom (according to the IPDE interview), or current clinically relevant symptoms (according to the Mini-Dips).

We aimed to collect 30 valid datasets from women with BPD and 30 from an age- and gender-matched control group. A power analysis confirmed that this sample size was sufficient (see section Power analysis for details). As participants in this study also completed behavioral and fMRI experiments for a larger project mentioned above, we ultimately collected data from 100 participants (44 individuals with BPD and 56 controls).

Prior to analysis, 25 datasets (4 BPD, 21 controls) had to be excluded due to a programming error in the task code that we only noticed after collecting these datasets (i.e., they did not receive the correct feedback due to a mistake in the randomization setup). We excluded an additional 10 participants in the BPD group for the following reasons: four did not meet the minimum requirement of four IPDE diagnostic criteria, two showed task comprehension difficulties, two had taken medications listed in our exclusion criteria, and two reported substance use before testing. Finally, we excluded four control participants who displayed clinically relevant symptoms according to the Mini-Dips during the diagnostic interview.

Computational models

To model participants' learning behavior in the behavioral task, we employed eight models introduced by Frolichs et al. (3) and extended them into a hierarchical paradigm. These models capture the learning and generalization processes at three levels of granularity, forming three distinct model classes: **no learning**, **coarse granularity**, and **fine granularity**. Each of these classes includes at least one instance that utilizes **reference points**. These reference points represent average trait values, either reflecting average population values (population reference points) or an individual's own trait values (self-reference points).

No learning models

The first model class implements a learning-free approach, where predictions are derived from linear transformations of reference points. For each participant j and trait t with reference point x_t:

$$y_{t,j} = b_{0j} + b_{1j}x_t + \epsilon_{t,j}, \quad \epsilon_{t,j} \sim \mathcal{N}(0, \sigma_j^2),$$

with participant-level parameters:

$$b_{0j} \sim \mathcal{N}(\mu_{b0}, \sigma_{b0}^2),$$

$$b_{1j} \sim \mathcal{N}(\mu_{b1}, \sigma_{b1}^2).$$

Since there are two types of reference points (population and self), this class consists of two models.

Coarse granularity models

The second class of models represents a coarse learning approach, where predictions are generalized across the Big Five personality dimensions within a Rescorla-Wagner learning framework. We denote the initial expectation for trait as $V_{0,t}$, treating it as an item-specific parameter drawn from a common distribution:

$$V_{0,t} \sim \mathcal{N}(\mu_{V_0}, \sigma_{V_0}^2).$$

For each subsequent trial i, the dimension d-specific expectation $V_{i,j,d}$ is then updated according to:

$$V_{i+1,j,d} = V_{i,j,d} + \alpha_j(r_{i,j,d} - V_{i,j,d}).$$

Here, $\alpha_j \in [0, 1]$ is a participant-specific learning rate that modulates the influence of the discrepancy between the feedback $r_{i,j,d}$ and the previous expectation $V_{i,j,d}$ on the updated expectation. As with the no learning model class, parameters are modeled hierarchically. To ensure the learning rate respects its natural bounds, we employed the probit (normal CDF) transform:

$$\alpha_j = \Phi(\eta_j), \quad \eta_j \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2).$$

Given a participant's expectation $V_{i,j,d}$, we estimate their prediction as:

$$y_{i,j,d} = V_{i,j,d} + \epsilon_{i,j,d}, \quad \epsilon_{i,j,d} \sim \mathcal{N}(0, \sigma_j^2),$$

where $\sigma_j^2 = \ln(1 + e^{z_j}), \quad z_j \sim \mathcal{N}(\mu_z, \sigma_z^2).$

The addition of the error term ϵ and its associated participant-specific variance σ_j^2 models stochastic deviations from the otherwise deterministic learning rule.

Fine granularity models

Finally, the third model class addresses the generalization process at the finest level by modeling separate expectations $\mathbf{V}_{i,j} \in \mathbb{R}^t$ for each trait t. These expectations are updated using a similarity matrix $\mathbf{R} \in \mathbb{R}^{t \times t}$, which was obtained by computing pairwise correlations from data in previous studies (2,4):

$$\mathbf{V}_{i+1,j} = \mathbf{V}_{i,j} + \alpha_j (r_{i,t,j} - V_{i,t,j}) \mathbf{R}_{:,t}.$$

Both the coarse and the fine granularity model class include models that make use of the population or self specific reference points $\mathbf{X} \in \mathbb{R}^t$. For the fine granularity models, the predictions are made as:

$$y_{i,j,t} = \gamma_j V_{i,j,t} + (1 - \gamma_j) x_t + \epsilon_{i,j,t}, \quad \epsilon_{i,j,t} \sim \mathcal{N}(0, \sigma_j^2),$$

where $\gamma_j \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$ is a participant-specific weighting parameter that modulates the influence of expectations versus reference points on the predictions. Predictions for the coarse granularity models are constructed similarly, using personality dimension d instead of trait t. Thus, both the coarse and the final granularity classes contain three models each (none, self, or population reference points).

Model fitting

All models were fitted using Stan (5) and the rstan interface (6). Stan is a probabilistic programming language designed for Bayesian statistical inference, utilizing techniques such as Hamiltonian Markov Chain Monte Carlo sampling. The framework lends itself well to hierarchical estimation which often surpasses non-hierarchical methods in accuracy (7,8). To obtain parameter estimates, we ran the NUTS algorithm with eight MCMC chains and 2000 iterations each, ensuring that all diagnostic criteria remained within ranges desired for accurate inference (9). Given that the computational load of full Hamiltonian MCMC runs is prohibitively expensive in combination with cross-validation, we used Stan's variational algorithm for approximate posterior sampling during model comparison.

Model comparison

All models include a participant-specific residual variance parameter σ_j that captures the variability in predictions for each participant, accounting for individual differences not explained by the model. As a byproduct, this facilitates the calculation of the model likelihood $P(D|\theta)$ via the normal distribution's log probability density function:

$$\ln P(D|\theta) = -\frac{N}{2}\ln(2\pi\sigma_j^2) - \sum_j \frac{(\mathbf{Y}_j - \mathbf{V}_j)^2}{2\sigma_j^2}.$$

We employed a Bayesian group-K-fold cross-validation scheme to estimate the out of sample likelihood on unseen participants $P(D_{new}|D,\theta)$. In the case of Bayesian models, this likelihood takes the form of the expected predictive density over the posterior draws based on the training folds. Specifically, after dividing the participants into K disjoint subsets (folds) of approximately equal size, each model is trained on $K-1$ folds, leaving out one fold for validation. This process is repeated K times so that each participant subset serves as a hold-out set exactly once. For Bayesian models, let y_i be an outcome for participant i in the held-out fold and $p(\theta|y_{(-k)})$ denote the posterior distribution of the parameters θ based on the training folds $y_{(-k)}$. The predictive likelihood for each held-out data point is then given by the posterior predictive integral:

$$\log p(y_i \mid y_{(-k)}) = \log \int p(y_i \mid \theta) p(\theta \mid y_{(-k)})\, d\theta.$$

Because $p(\theta|y_{(-k)})$ is available only through Monte Carlo draws from the posterior, we approximate the integral using S samples $\{\theta^{k,s}\}_{s=1}^{S}$ drawn from the posterior of the k-th training fold:

$$\widehat{\mathrm{elpd}}_i = \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta^{k,s})\right).$$

Summing across all held-out points from all folds, we obtain an estimate of the expected log predictive density:

$$\widehat{\mathrm{elpd}}_{\mathrm{xval}} = \sum_{i=1}^{n} \widehat{\mathrm{elpd}}_i,$$

where n is the total number of data points. The model with the highest elpd score is considered the best fit to the data. To get an idea of the uncertainty associated with the ranking of models based on their elpd estimates, we compute standard errors of their differences. Although no formal decision boundary akin to a frequentist significance level exists, performance differences between models A and B are sometimes considered "significant" when $|\widehat{\mathrm{elpd}}_{\mathrm{A}} - \widehat{\mathrm{elpd}}_{\mathrm{B}}| \geq 2\mathrm{se}(\widehat{\mathrm{elpd}}_{\mathrm{A}} - \widehat{\mathrm{elpd}}_{\mathrm{B}})$, which is approximately equivalent to performing a conventional significance test (10).

For our study, we set the number of holdout folds, k, to the number of participants in the data. This effectively performs leave-one-participant-out cross-validation, which is simultaneously the most accurate and computationally expensive approach available for our data.

Supplementary results

Self-ratings on trait adjectives

Table S4 LME model on the effect of Group on Self-Ratings

Predictors	*b*	*SE*	*95% CI*	*t*	*p*	*df*
(Intercept)	5.44	0.14	5.16 – 5.73	37.57	**<0.001**	61.63
Group [CON]	0.77	0.11	0.56 – 0.99	7.16	**<0.001**	58.76
Random Effects						
σ^2	2.19					
$\tau_{00\ sub}$	0.12					
$\tau_{00\ item}$	0.60					
ICC	0.25					
N_{sub}	61					
N_{item}	40					
Observations	2406					
Marginal R^2 / Conditional R^2	0.049 / 0.285					

Note. Significant results in bold font

Table S5 Type III ANOVA on the effect of Group on Self-Ratings

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Group	112.45	112.45	1	58.76	51.30	**< 0.001**

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S6 LME model on the effect of Group, Valence, and their interaction on Self-Ratings

Predictors	*b*	*SE*	*95% CI*	*t*	*p*	*df*
(Intercept)	3.89	0.19	3.52 – 4.27	20.77	**<0.001**	48.72
Group [CON]	-1.05	0.11	-1.27 – -0.84	-9.57	**<0.001**	120.51
Valence [positive]	1.89	0.26	1.37 – 2.41	7.39	**<0.001**	42.87
Group [CON] × Valence [positive]	1.54	0.12	1.31 – 1.78	12.71	**<0.001**	2305.41
Random Effects						
σ^2	2.22					
$\tau_{00\ sub}$	0.07					
$\tau_{00\ item}$	0.58					
ICC	0.23					
$N_{\ sub}$	61					
$N_{\ item}$	40					
Observations	2406					
Marginal R^2 / Conditional R^2	0.406 / 0.541					

Note. Significant results in bold font

Table S7 Type III ANOVA on the effect of Group, Valence, and their interaction on Self-Ratings

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Group	20.88	20.88	1	58.92	9.40	**0.003**
Valence	255.35	255.35	1	38.00	115.02	**< 0.001**
Group × Valence	358.40	358.40	1	2,305.41	161.43	**< 0.001**

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S8 EMMs Contrasts of Group Differences Within Each Valence

Valence	*Contrast*	*Estimate*	*SE*	*df*	*CI(LL)*	*CI(UL)*	*t*	*p*
Negative	Group BPD - Group CON	1.054	0.11	120.644	0.836	1.272	9.574	**< 0.001**
Positive	Group BPD - Group CON	-0.490	0.11	121.671	-0.709	-0.272	-4.441	**< 0.001**

Note. Contrasts reflect differences in predicted values of Self-Ratings between Groups within each Valence. Degrees-of-freedom method: Kenward-Roger. *CI*: 95% Confidence Intervals with lower limits (*LL*) and upper limits (*UL*). Significant results in bold font

Self-ratings on personality inventories

NEO-FFI

Table S9 Multivariate Analysis of Variance (MANOVA) on the effect of Group on the NEO-FFI personality factors

Effect	*DF*	*Pillai's Trace*	*F (approx..)*	*df1*	*df2*	*p*
(Intercept)	1.00	0.99	1,431.04	5.00	55.00	**<0.001**
Group	1.00	0.77	36.10	5.00	55.00	**<0.001**
Residuals	59.00					

DF = degrees of freedom for the terms; df1 = degrees of freedom in the model; df2 = degrees of freedom associated with the model errors; Significant results in bold font

Table S10 Post-hoc Welch's t-tests comparing BPD and control groups on NEO-FFI personality factors

	BPD		CON								
Factor	*M*	*SD*	*M*	*SD*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*d*	*p*	*p (fdr)*
N	3.09	0.44	1.55	0.52	12.54	57.90	1.30	1.79	3.20	**<0.001**	**<0.001**
E	1.93	0.63	2.33	0.66	-2.43	59.00	-0.73	-0.07	-0.62	**0.018**	**0.022**
O	2.49	0.56	2.31	0.46	1.33	55.80	-0.09	0.44	0.34	0.190	0.190
A	2.39	0.49	2.98	0.44	-4.88	58.00	-0.82	-0.34	-1.25	**<0.001**	**<0.001**
C	2.21	0.55	2.94	0.61	-4.92	58.60	-1.03	-0.43	-1.26	**<0.001**	**<0.001**

Factors: N = Neuroticism, E = Extraversion, O = Openness, A= Agreeableness, C = Conscientiousness; CI = 95% confidence interval; LL = lower limit; UL = upper limit; d = Cohen's d; p (fdr) = p-value adjusted for false discovery rate; Significant results in bold font

PID-5-BF

Table S11 Multivariate Analysis of Variance (MANOVA) on the effect of Group on the PID-5-BF domains

Effect	*DF*	*Pillai's Trace*	*F (approx..)*	*df1*	*df2*	*p*
(Intercept)	1.00	0.92	125.79	5.00	55.00	**<0.001**
Group	1.00	0.68	23.86	5.00	55.00	**<0.001**
Residuals	59.00					

DF = degrees of freedom for the terms; df1 = degrees of freedom in the model; df2 = degrees of freedom associated with the model errors; Significant results in bold font

Table S12 Post-hoc Welch's t-tests comparing BPD and control groups on the PID-5-BF domains

	BPD		CON								
Factor	*M*	*SD*	*M*	*SD*	*t*	*DF*	*CI (LL)*	*CI (UL)*	*d*	*p*	*p (fdr)*
Negative Affect	10.50	2.01	4.61	2.79	9.48	54.60	4.64	7.13	2.41	**<0.001**	**<0.001**
Detachment	6.63	2.58	2.87	2.47	5.81	58.70	2.47	5.06	1.49	**<0.001**	**<0.001**
Antagonism	4.23	3.50	1.39	1.71	4.02	41.70	1.42	4.28	1.04	**<0.001**	**<0.001**
Disinhibition	8.13	3.08	2.65	2.47	7.66	55.50	4.05	6.92	1.97	**<0.001**	**<0.001**
Psychoticism	7.70	3.63	1.84	1.73	8.00	41.30	4.38	7.34	2.07	**<0.001**	**<0.001**

CI = 95% confidence interval; LL = lower limit; UL = upper limit; d = Cohen's d; p (fdr) = p-value adjusted for false discovery rate; Significant results in bold font

Other-ratings

Table S13 LME model on the effect of Group, Valence, and their interaction on Other-Ratings

Predictors	*b*	*SE*	*95% CI*	*t*	*p*	*df*
(Intercept)	3.78	0.10	3.58 – 3.97	38.26	**<0.001**	65.03
Group [CON]	-0.14	0.07	-0.29 – 0.01	-1.83	0.071	70.82
Valence [positive]	1.88	0.12	1.64 – 2.13	15.48	**<0.001**	40.69
Group [CON] × Valence [positive]	0.23	0.04	0.14 – 0.31	5.20	**<0.001**	13958.68
Random Effects						
σ^2	1.69					
$\tau_{00\ sub}$	0.07					
$\tau_{00\ item}$	0.14					
ICC	0.11					
$N_{\ sub}$	61					
$N_{\ item}$	40					
Observations	14059					
Marginal R^2 / Conditional R^2	0.346 / 0.418					

Note. Significant results in bold font

Table S14 Type III ANOVA on the effect of Group, Valence, and their interaction on Other-Ratings

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Group	0.17	0.17	1	58.96	0.10	0.750
Valence	471.03	471.03	1	38.02	278.99	**< 0.001**
Group × Valence	45.72	45.72	1	13,958.68	27.08	**< 0.001**

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S15 EMMs Contrasts of Group Differences Within Each Valence

Valence	*Contrast*	*Estimate*	*SE*	*z*	*p*
Negative	Group BPD - Group CON	0.137	0.075	1.835	0.067
Positive	Group BPD - Group CON	-0.091	0.075	-1.224	0.221

Note. Contrasts reflect differences in predicted values of Other-Ratings between Groups within each Valence. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Significant results in bold font

Table S16 LME model on the effect of Group, Profile Type, and their interaction on Other-Ratings

Predictors	*b*	*SE*	*95% CI*	*t*	*p*	*df*
(Intercept)	5.41	0.09	5.24 – 5.58	62.64	**<0.001**	82.50
Group [CON]	-0.01	0.08	-0.16 – 0.14	-0.17	0.863	69.90
Profile Type [CON]	0.07	0.03	0.01 – 0.13	2.30	**0.022**	13963.05
Group [CON] × Profile Type [CON]	0.26	0.04	0.17 – 0.34	5.86	**<0.001**	13964.15
Random Effects						
σ^2	1.67					
$\tau_{00\ sub}$	0.07					
$\tau_{00\ item}$	0.18					
ICC	0.13					
N_{sub}	61					
N_{item}	40					
Observations	14059					
Marginal R^2 / Conditional R^2	0.009 / 0.140					

Note. Significant results in bold font

Table S17 Type III ANOVA on the effect of Group, Profile Type, and their interaction on Other-Ratings

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Group	4.13	4.13	1	59.08	2.47	0.121
Profile Type	139.23	139.23	1	13,964.49	83.36	**< 0.001**
Group × Profile Type	57.39	57.39	1	13,964.15	34.36	**< 0.001**

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S18 EMMs Contrasts of Group Differences Within Each Profile Type

Profile Type	*Contrast*	*Estimate*	*SE*	*z*	*p*
BPD	Group BPD - Group CON	0.013	0.076	0.173	0.863
CON	Group BPD - Group CON	-0.243	0.076	-3.185	**0.001**

Note. Contrasts reflect differences in predicted values of Other-Ratings between Groups within each Profile Type. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Significant results in bold font

Prediction errors

Table S19 LME model on the effect of Group, Profile Type, and their interaction on PEs

Predictors	*b*	*SE*	*95% CI*	*t*	*p*	*df*
(Intercept)	1.52	0.06	1.40 – 1.63	25.55	**<0.001**	87.10
Group [CON]	0.00	0.05	-0.11 – 0.11	0.06	0.956	77.25
Profile Type [CON]	-0.09	0.03	-0.15 – -0.04	-3.32	**0.001**	13966.57
Group [CON] × Profile Type [CON]	-0.17	0.04	-0.24 – -0.09	-4.28	**<0.001**	13968.22
Random Effects						
σ^2	1.33					
$\tau_{00\ sub}$	0.03					
$\tau_{00\ item}$	0.08					
ICC	0.08					
N_{sub}	61					
N_{item}	40					
Observations	14059					
Marginal R^2 / Conditional R^2	0.008 / 0.086					

Note. Significant results in bold font

Table S20 Type III ANOVA on the effect of Group, Profile Type, and their interaction on PEs

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Group	3.31	3.31	1	59.15	2.48	0.120
Profile Type	108.53	108.53	1	13,968.82	81.30	**< 0.001**
Group × Profile Type	24.50	24.50	1	13,968.22	18.36	**< 0.001**

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S21 EMMs Contrasts of Group Differences Within Each Profile Type

Profile Type	*Contrast*	*Estimate*	*SE*	*z*	*p*
BPD	Group BPD - Group CON	-0.003	0.055	-0.056	0.956
CON	Group BPD - Group CON	0.164	0.055	2.996	**0.003**

Note. Contrasts reflect differences in predicted values of PEs between Groups within each Profile Type. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Significant results in bold font

Table S22 EMMs Contrasts of Profile Type Differences Within Each Group

Group	*Contrast*	*Estimate*	*SE*	*z*	*p*
BPD	Profile Type BPD - Profile Type CON	0.092	0.028	3.324	**0.001**
CON	Profile Type BPD - Profile Type CON	0.260	0.027	9.473	**<0.001**

Note. Contrasts reflect differences in predicted values of PEs between Profile Type within each Group. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Significant results in bold font

Table S23 LME model on the effect of Trial on PEs

Predictors	*b*	*SE*	*95% CI*	*t*	*p*	*df*
(Intercept)	1.49	0.05	1.38 – 1.60	27.13	**<0.001**	76.34
Trial	-0.00	0.00	-0.00 – -0.00	-3.10	**0.002**	13875.47
Random Effects						
σ^2	1.34					
$\tau_{00\ sub}$	0.03					
$\tau_{00\ item}$	0.08					
ICC	0.08					
N_{sub}	61					
N_{item}	40					
Observations	14059					
Marginal R^2 / Conditional R^2	0.001 / 0.079					

Note. Significant results in bold font

Table S24 Type III ANOVA on the effect of Trial on PEs

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Trial	12.94	12.94	1	13,875.47	9.63	**0.002**

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S 25 LME model on the effect of Trial, Group, Profile Type, and their interaction on PEs

Predictors	b	SE	95% CI	t	p	df
(Intercept)	1.54	0.07	1.40 – 1.67	22.16	**<0.001**	164.82
Trial	-0.00	0.00	-0.00 – 0.00	-0.61	0.540	13969.69
Group [CON]	0.02	0.07	-0.12 – 0.17	0.32	0.748	246.21
Profile Type [CON]	-0.04	0.06	-0.15 – 0.08	-0.62	0.534	13988.44
Trial × Group [CON]	-0.00	0.00	-0.01 – 0.00	-0.42	0.674	13953.68
Trial × profile Type [CON]	-0.00	0.00	-0.01 – 0.00	-1.05	0.292	13967.61
Group [CON] × Profile Type [CON]	-0.18	0.08	-0.34 – -0.02	-2.23	**0.026**	13956.60
Trial × Group [CON] × Profile Type [CON]	0.00	0.00	-0.01 – 0.01	0.17	0.862	13954.14
Random Effects						
σ^2	1.33					
$\tau_{00\ sub}$	0.03					
$\tau_{00\ item}$	0.08					
ICC	0.08					
N_{sub}	61					
N_{item}	40					
Observations	14059					
Marginal R^2 / Conditional R^2	0.008 / 0.086					

Note. Significant results in bold font

Table S26 Type III ANOVA on the effect of Trial, Group, Profile Type, and their interaction on PEs

Predictors	*Sum of Squares*	*Mean Square*	df_1	df_2	*F*	*p*
Trial	12.26	12.26	1	13,872.31	9.18	**0.002**
Group	1.51	1.51	1	127.64	1.13	0.289
Profile Type	11.55	11.55	1	13,893.68	8.65	**0.003**
Trial × Group	0.23	0.23	1	13,954.29	0.17	0.680
Trial × Profile Type	2.19	2.19	1	13,788.83	1.64	0.200
Group × Profile Type	6.64	6.64	1	13,956.60	4.98	**0.026**
Trial × Group × Profile Type	0.04	0.04	1	13,954.14	0.03	0.862

Note. df_1 = numerator degrees of freedom; df_2 = denominator degrees of freedom; Significant results in bold font

Table S27 EMMs Contrasts of Group Differences Within Each Profile Type

Profile Type	*Contrast*	*Estimate*	*SE*	*z*	*p*
BPD	Group BPD - Group CON	-0.003	0.055	-0.054	0.957
CON	Group BPD - Group CON	0.164	0.055	2.991	**0.003**

Note. Contrasts reflect differences in predicted values of PEs between Groups within each Profile Type. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Significant results in bold font

Table S28 Interaction Contrasts of Group × Profile Type on the Linear Effect of Trial (Slopes) on PEs

Contrast	*Estimate*	*SE*	*z*	*p*
Group BPD Profile Type BPD - Group CON Profile Type BPD	0.001	0.002	0.42	0.975
Group BPD Profile Type BPD - Group BPD Profile Type CON	0.003	0.003	1.055	0.717
Group BPD Profile Type BPD - Group CON Profile Type CON	0.003	0.003	1.222	0.613
Group CON Profile Type BPD - Group BPD Profile Type CON	0.002	0.003	0.668	0.909
Group CON Profile Type BPD - Group CON Profile Type CON	0.002	0.002	0.835	0.838
Group BPD Profile Type CON - Group CON Profile Type CON	0.000	0.002	0.167	0.998

Note**.** Contrasts test pairwise differences in the slopes of the Trial effect on PEs across combinations of Group and Profile Type. p-values adjusted using Tukey's method for multiple comparisons. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Significant results in bold font

Table S29 Estimated Trends (Slopes) for Trial Effects on PE by Group and Profile Type

Group	Profile Type	Trend (Trial)	SE	LCL	UCL
BPD	Profile Type BPD	-0.001	0.002	-0.005	0.002
CON	Profile Type BPD	-0.002	0.002	-0.005	0.001
BPD	Profile Type CON	-0.004	0.002	**-0.007**	**0.000**
CON	Profile Type CON	-0.004	0.002	**-0.008**	**-0.001**

Note. Trends represent the estimated linear effect of trial (centered) on PEs for each group × profile type combination. Degrees-of-freedom method: asymptotic. Df values were not included due to high memory and computation load. Confidence level: 95%. LCL = lower confidence limit; UCL = upper confidence limit. Significant results in bold font

Computational models

Table S30. Differences in population means of learning rates and initial expectations.

Contrast	Parameter	Posterior Mean Δ	Lower 95% SPI	Upper 95% SPI
CON on CON vs. CON on BPD	alpha_mu	0.04	-0.10	0.18
	V0_mu	0.22	0.00	0.43
CON on CON vs. CON on BPD	alpha_mu	-0.02	-0.14	0.11
	V0_mu	-0.18	-0.40	0.06
CON on CON vs. BPD on CON	alpha_mu	0.02	-0.12	0.16
	V0_mu	**0.42***	0.19	0.65
CON on BPD vs. CON on BPD	alpha_mu	-0.03	-0.15	0.09
	V0_mu	0.02	-0.19	0.23

References

1. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007 May;39(2):175–91. https://doi.org/10.3758/BF03193146

2. Korn CW, La Rosée L, Heekeren HR, Roepke S. Social feedback processing in borderline personality disorder. Psychol Med. 2016 Feb;46(3):575–87. https://doi.org/10.1017/S003329171500207X

3. Frolichs KMM, Rosenblau G, Korn CW. Incorporating social knowledge structures into computational models. Nat Commun. 2022 Oct 20;13(1):6205. https://doi.org/10.1038/s41467-022-33418-2

4. Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR. Positively Biased Processing of Self-Relevant Social Feedback. J Neurosci. 2012 Nov 21;32(47):16832–44. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

5. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. *Stan*: A Probabilistic Programming Language. J Stat Softw [Internet]. 2017 [cited 2025 Jul 7];76(1). Available from: http://www.jstatsoft.org/v76/i01/10.18637/jss.v076.i01

6. Gabry J, Češnovar R, Johnson A, Bronder S. cmdstanr: R Interface to "CmdStan" [Internet]. 2025. Available from: https://mc-stan.org/cmdstanr/

7. Lee MD, Wagenmakers EJ. Bayesian cognitive modeling: a practical course. Cambridge, United Kingdom New York: Cambridge University Press; 2014. https://doi.org/10.1017/CBO9781139087759

8. Zhang L, Lengersdorff L, Mikus N, Gläscher J, Lamm C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. Soc Cogn Affect Neurosci. 2020 Jul 30;15(6):695–707. https://doi.org/10.1093/scan/nsaa089

9. Makowski D, Ben-Shachar M, Lüdecke D. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. J Open Source Softw. 2019 Aug;4(40):1541. https://doi.org/10.21105/joss.01541

10. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017 Sep;27(5):1413–32. https://doi.org/10.48550/ARXIV.1507.04544

Study II - Supplementary material

Supplementary material

Sample description

Table S1. Demographic characteristics of both groups

	BPD (*N* = 31)		Control (*N* = 34)		
	N	*%*	*N*	*%*	*p*
Education					
Secondary general school (9 years)	1	3.23	0	0.00	**0.027**
Intermediate school (10 years)	6	19.35	1	2.94	
High school diploma (13 years)	24	77.42	33	97.06	
Professional qualification					
Non, in professional training	7	22.58	21	61.76	**< 0.001**
Non, not in professional training	5	16.13	0	0.00	
Apprenticeship	6	19.35	1	2.94	
Vocational school	8	25.81	3	8.82	
Technical school	0	0.00	1	2.94	
University	5	16.13	8	23.53	
Employment status					
Student	5	16.13	28	82.35	**< 0.001**
In training	6	19.35	0	0.00	
Employed	8	25.81	3	8.82	
Unemployed	5	16.13	1	2.94	
Retired or homemaker	2	6.45	0	0.00	
Other	5	16.13	2	5.88	

Note. *p* = p-values from Fisher's exact test for count data. Significant results in bold font

Table S2. Psychotropic medication intake of the BPD group

Medication	*N*	**%**
Antidepressant, SNRI	5	16.13
Antidepressant, SSRI	8	25.81
Antidepressant, atypical	1	3.23
Atypical antipsychotics	3	9.68
Mood stabilizer (Lamotrigine)	2	6.45
No medication	15	48.39

List of interviews and questionnaires

Table S3. Interviews and Questionnaires

Name	Type
Mini-DIPS Open Access (Diagnostic Short-Interview for Mental Disorders; Margraf & Cwik, 2017)	Interview
IPDE (International Personality Disorder Examination; Loranger, 1994)	Interview
BSL (Borderline Symptom List-23; Bohus et al., 2009)	Self-report
BSCL (Brief-Symptom-Checklist; Franke, 2017)	Self-report
BDI-II (Beck Depression Inventory; Beck et al., 2006)	Self-report
ASRS-v1.1 (Adult ADHD Self-Report Scale; Kessler et al., 2005)	Self-report
ADHS-SB (ADHS-Self-Report for adults; Rösler et al., 2004)	Self-report
WURS-K (Wender Utah Rating Scale; Retz-Junginger et al., 2002)	Self-report
MDBF (Multidimensional Mood Questionnaire – Short Version A; Steyer et al., 1997)	Self-report
NEO-FFI (Five-Factor Inventory; Borkenau & Ostendorf, 2007)	Self-report
PID-5-BF (Personality Inventory for DSM-5-Brief Form; Bach et al., 2020; Zimmermann et al., 2014)	Self-report
STAXI-2 (State-Trait Anger Expression Inventory-2; Rohrmann et al., 2013)	Self-report
BIS-11 (Barratt Impulsiveness Scale; Preuss et al., 2008)	Self-report

IERQ (Interpersonal Emotion Regulation Questionnaire; Hofmann & Kashdan, 2010; Pruessner et al., 2020)	Self-report
mini-q (A Three-Minute Intelligence Screening; Baudson & Preckel, 2016)	Self-report
PiCD (The Personality Inventory for ICD-11; Damovsky et al., 2021; Oltmanns & Widiger, 2018)	Self-report
LPFS-BF (Level of Personality Functioning Scale-Brief Form; Spitzer et al., 2021; Weekers et al., 2019)	Self-report
IES-27 (Impulsivity and Emotion Dysregulation Scale for Borderline Personality Disorder; Kröger & Kosfelder, 2011)	Self-report

Behavioral results: model summaries and model comparisons

Decision task

Table S4. Decision task Model 1 - GLMM Summary (OR)

	Choice				
Predictors	*Odds Ratios*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.62	0.09	0.46 – 0.83	-3.21	**0.001**
Group	1.30	0.19	0.97 – 1.74	1.77	0.078
Partner	3.40	0.11	3.19 – 3.62	38.63	**<0.001**
Risk	0.93	0.03	0.88 – 0.99	-2.42	**0.015**
Group × Partner	1.05	0.03	0.99 – 1.12	1.54	0.124
Group × Risk	0.88	0.03	0.83 – 0.94	-4.12	**<0.001**
Partner × Risk	0.79	0.02	0.75 – 0.84	-7.66	**<0.001**
Group × Partner × Risk	1.04	0.03	0.98 – 1.10	1.20	0.230
Random Effects					
σ^2	3.29				
$\tau_{00\ participant}$	1.36				
ICC	0.29				
$N_{\ participant}$	65				
Observations	9074				
Marginal R^2 / Conditional R^2	0.238 / 0.461				

Table S5. Decision task Model 2 – GLMM Summary (OR)

	Choice				
Predictors	*Odds Ratios*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.62	0.09	0.47 – 0.83	-3.18	**0.001**
Group	1.29	0.19	0.97 – 1.72	1.72	0.085
Partner	3.39	0.11	3.19 – 3.61	38.73	**<0.001**
Risk	0.93	0.03	0.87 – 0.98	-2.57	**0.010**
Group × Partner	1.06	0.03	1.00 – 1.12	1.88	0.061
Group × Risk	0.90	0.02	0.85 – 0.95	-4.03	**<0.001**
Partner × Risk	0.80	0.02	0.75 – 0.85	-7.59	**<0.001**
Random Effects					
σ^2	3.29				
$\tau_{00\ participant}$	1.36				
ICC	0.29				
$N_{\ participant}$	65				
Observations	9074				
Marginal R^2 / Conditional R^2	0.237 / 0.460				

Table S6. Decision task Model 3 – GLMM Summary (OR)

	Choice				
Predictors	*Odds Ratios*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.64	0.09	0.48 – 0.86	-3.02	**0.003**
Group	1.27	0.19	0.95 – 1.70	1.64	0.100
Partner	3.26	0.10	3.07 – 3.46	38.62	**<0.001**
Risk	0.84	0.02	0.80 – 0.88	-6.68	**<0.001**
Group × Partner	1.07	0.03	1.00 – 1.13	2.08	**0.037**
Group × Risk	0.91	0.02	0.86 – 0.95	-3.74	**<0.001**
Random Effects					
σ^2	3.29				
$\tau_{00\ participant}$	1.34				
ICC	0.29				
$N_{\ participant}$	65				
Observations	9074				
Marginal R^2 / Conditional R^2	0.231 / 0.453				

Table S7. Likelihood ratio test comparing the decision task Model 1 to Model 2

Model	*npar*	*AIC*	*BIC*	*logLik*	*Chisq*	*df*	*p*
Model 2 (reduced)	8	9,198.863	9,255.768	-4,591.431			
Model 1 (full)	9	9,199.421	9,263.439	-4,590.710	1.442	1	0.230

Table S8. Likelihood ratio test comparing the decision task Model 2 to Model 3

Model	*npar*	*AIC*	*BIC*	*logLik*	*Chisq*	*df*	*p*
Model 3 (reduced)	7	9,255.754	9,305.547	-4,620.877			
Model 2	8	9,198.863	9,255.768	-4,591.431	58.892	1	**< .001**

Expectation task

Table S9. Expectation task Model 1 – GLMM Summary (OR)

	Expectation				
Predictors	*Odds Ratios*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.45	0.06	0.35 – 0.57	-6.47	**<0.001**
Group	1.12	0.14	0.88 – 1.43	0.91	0.362
Partner	4.01	0.20	3.64 – 4.43	27.52	**<0.001**
Risk	0.59	0.03	0.54 – 0.65	-10.84	**<0.001**
Group × Partner	1.30	0.07	1.18 – 1.44	5.28	**<0.001**
Group × Risk	0.96	0.05	0.88 – 1.06	-0.75	0.455
Partner × Risk	0.79	0.04	0.72 – 0.87	-4.82	**<0.001**
Group × Partner × Risk	0.94	0.05	0.86 – 1.03	-1.26	0.209
Random Effects					
σ^2	3.29				
$\tau_{00\ participant}$	0.84				
ICC	0.20				
$N_{\ participant}$	65				
Observations	4499				
Marginal R^2 / Conditional R^2	0.358 / 0.489				

Table S10. Expectation task Model 2 – GLMM Summary (OR)

	Expectation				
Predictors	*Odds Ratios*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.45	0.06	0.35 – 0.57	-6.49	**<0.001**
Group	1.12	0.14	0.88 – 1.43	0.93	0.354
Partner	4.01	0.20	3.63 – 4.43	27.50	**<0.001**
Risk	0.59	0.03	0.54 – 0.65	-10.87	**<0.001**
Group × Partner	1.29	0.06	1.17 – 1.43	5.17	**<0.001**
Group × Risk	0.93	0.04	0.86 – 1.00	-1.86	0.062
Partner × Risk	0.80	0.04	0.72 – 0.87	-4.75	**<0.001**
Random Effects					
σ^2	3.29				
$\tau_{00\ participant}$	0.84				
ICC	0.20				
$N_{\ participant}$	65				
Observations	4499				
Marginal R^2 / Conditional R^2	0.358 / 0.488				

Table S11. Expectation task Model 3 – GLMM Summary (OR)

	Expectation				
Predictors	*Odds Ratios*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.43	0.05	0.34 – 0.54	-7.44	**<0.001**
Group	1.10	0.12	0.88 – 1.37	0.80	0.423
Partner	3.88	0.19	3.53 – 4.27	27.68	**<0.001**
Group × Partner	1.26	0.06	1.14 – 1.39	4.74	**<0.001**
Random Effects					
σ^2	3.29				
$\tau_{00\ participant}$	0.69				
ICC	0.17				
$N_{\ participant}$	65				
Observations	4499				
Marginal R^2 / Conditional R^2	0.297 / 0.418				

Table S12. Likelihood ratio test comparing the expectation task Model 1, Model 2, and Model 3

Model	*npar*	*AIC*	*BIC*	*logLik*	*Chisq*	*df*	*p*
Model 3 (group, partner and interaction)	5	4,820.490	4,852.548	-2,405.245			
Model 2 (group, partner, risk, and 2-way interactions)	8	4,480.765	4,532.058	-2,232.383	345.724	3	**< .001**
Model 1 (group, partner, risk, and all interactions)	9	4,481.187	4,538.891	-2,231.593	1.578	1	0.209

Response times decision task

Table S13. Response time decision task Model 1 – GLMM Summary

	log(RT)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.41	0.02	0.37 – 0.45	18.70	**<0.001**
Group	0.02	0.02	-0.02 – 0.06	0.86	0.387
Partner	0.06	0.00	0.06 – 0.07	18.20	**<0.001**
Risk	0.01	0.00	0.00 – 0.02	3.18	**0.001**
Group × Partner	0.00	0.00	-0.00 – 0.01	0.99	0.323
Group × Risk	0.00	0.00	-0.00 – 0.01	0.85	0.394
Partner × Risk	0.03	0.00	0.02 – 0.04	8.00	**<0.001**
Group × Partner × Risk	0.00	0.00	-0.00 – 0.01	1.23	0.220
Random Effects					
σ^2	0.10				
$\tau_{00\ participant}$	0.03				
ICC	0.24				
$N_{\ participant}$	65				
Observations	9074				
Marginal R^2 / Conditional R^2	0.041 / 0.274				

Table S14. Response time decision task Model 2 – GLMM Summary

	log(RT)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.41	0.02	0.37 – 0.45	18.70	**<0.001**
Group	0.02	0.02	-0.02 – 0.06	0.84	0.403
Partner	0.06	0.00	0.06 – 0.07	18.18	**<0.001**
Risk	0.01	0.00	0.00 – 0.02	3.15	**0.002**
Group × Partner	0.00	0.00	-0.00 – 0.01	1.18	0.238
Group × Risk	0.00	0.00	-0.00 – 0.01	1.40	0.161
Partner × Risk	0.03	0.00	0.02 – 0.04	8.08	**<0.001**
Random Effects					
σ^2	0.10				
$\tau_{00\ participant}$	0.03				
ICC	0.24				
$N_{\ participant}$	65				
Observations	9074				
Marginal R^2 / Conditional R^2	0.041 / 0.274				

Table S15. Response time decision task Model 3 – GLMM Summary

	log(RT)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	0.41	0.02	0.36 – 0.45	18.51	**<0.001**
Group	0.02	0.02	-0.02 – 0.06	0.84	0.399
Partner	0.07	0.00	0.06 – 0.07	19.50	**<0.001**
Risk	0.02	0.00	0.02 – 0.03	6.55	**<0.001**
Group × Partner	0.00	0.00	-0.00 – 0.01	1.13	0.257
Group × Risk	0.00	0.00	-0.00 – 0.01	1.41	0.158
Random Effects					
σ^2	0.10				
$\tau_{00\ participant}$	0.03				
ICC	0.24				
$N_{\ participant}$	65				
Observations	9074				
Marginal R^2 / Conditional R^2	0.036 / 0.269				

Table S16. Likelihood ratio test comparing the response time decision task Model 1, Model 2, and Model 3

Model	*npar*	*AIC*	*BIC*	*logLik*	*Chisq*	*df*	*p*
Model 3 (without partner x risk interaction)	8	4,725.767	4,782.672	-2,354.883			
Model 2 (group, partner, risk, and 2-way interactions)	9	4,662.654	4,726.672	-2,322.327	65.113	1	**< .001**
Model 1 (group, partner, risk, and all interactions)	10	4,663.149	4,734.281	-2,321.575	1.505	1	0.220

Table S17. Response time decision task – Type III F-tests (Satterthwaite's approximation for Model 2)

Term	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F value*	*Pr(>F)*
(Intercept)	0.066	0.066	1	63.366	0.699	0.406
Group	31.440	31.440	1	9,004.296	330.524	**0.000**
Partner	0.947	0.947	1	9,004.276	9.951	**0.002**
Risk	0.133	0.133	1	9,004.353	1.395	0.238
Group x Partner	0.187	0.187	1	9,004.130	1.961	0.161
Group x Risk	6.213	6.213	1	9,004.230	65.312	**0.000**

Response times expectation task

Table S18. Response time expectation task Model 1 – GLMM Summary

	log(RT)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	-0.71	0.03	-0.78 – -0.65	-21.66	**<0.001**
Group	0.04	0.03	-0.03 – 0.10	1.14	0.253
Partner	0.06	0.01	0.03 – 0.08	3.74	**<0.001**
Risk	0.01	0.01	-0.02 – 0.04	0.65	0.515
Group × Partner	-0.02	0.01	-0.05 – 0.01	-1.29	0.197
Group × Risk	-0.01	0.01	-0.04 – 0.02	-0.52	0.602
Partner × Risk	0.02	0.01	-0.01 – 0.05	1.60	0.111
Group × Partner × Risk	0.02	0.01	-0.01 – 0.05	1.37	0.171
Random Effects					
σ^2	0.81				
$\tau_{00\ participant}$	0.06				
ICC	0.06				
$N_{\ participant}$	65				
Observations	4499				
Marginal R^2 / Conditional R^2	0.006 / 0.070				

Table S19. Response time expectation task Model 2 – GLMM Summary

	log(RT)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	-0.71	0.03	-0.78 – -0.65	-21.65	**<0.001**
Group	0.03	0.03	-0.03 – 0.10	1.03	0.301
Partner	0.06	0.01	0.03 – 0.08	3.73	**<0.001**
Risk	0.01	0.01	-0.02 – 0.04	0.63	0.528
Group × Partner	-0.02	0.01	-0.04 – 0.01	-1.05	0.292
Group × Risk	0.00	0.01	-0.03 – 0.03	0.04	0.968
Partner × Risk	0.02	0.01	-0.00 – 0.05	1.66	0.098
Random Effects					
σ^2	0.81				
$\tau_{00\ participant}$	0.06				
ICC	0.06				
$N_{\ participant}$	65				
Observations	4499				
Marginal R^2 / Conditional R^2	0.005 / 0.070				

Table S20. Response time expectation task Model 3 – GLMM Summary

	log(RT)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	-0.72	0.03	-0.78 – -0.65	-21.86	**<0.001**
Group	0.03	0.03	-0.03 – 0.10	1.04	0.300
Partner	0.06	0.01	0.03 – 0.09	4.11	**<0.001**
Risk	0.02	0.01	-0.01 – 0.05	1.43	0.153
Group × Partner	-0.02	0.01	-0.04 – 0.01	-1.05	0.292
Group × Risk	0.00	0.01	-0.03 – 0.03	0.05	0.963
Random Effects					
σ^2	0.81				
$\tau_{00\ participant}$	0.06				
ICC	0.06				
$N_{\ participant}$	65				
Observations	4499				
Marginal R^2 / Conditional R^2	0.005 / 0.069				

Table S21. Likelihood ratio test comparing the response time expectation task Model 1, Model 2, and Model 3

Model	*npar*	*AIC*	*BIC*	*logLik*	*Chisq*	*df*	*p*
Model 3 (without partner x risk interaction)	8	11,947.12	11,998.42	-5,965.562			
Model 2 (group, partner, risk, and 2-way interactions)	9	11,946.38	12,004.08	-5,964.190	2.744	1	0.098
Model 1 (group, partner, risk, and all interactions)	10	11,946.51	12,010.62	-5,963.253	1.874	1	0.171

Table S22. Response time expectation task – Type III F-tests (Satterthwaite's approximation for Model 3)

Term	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F value*	*Pr(>F)*
(Intercept)	0.870	0.870	1	67.056	1.074	0.304
Group	13.693	13.693	1	4,431.458	16.893	**0.000**
Partner	1.660	1.660	1	4,430.315	2.047	0.153
Risk	0.901	0.901	1	4,431.458	1.111	0.292
Group x Partner	0.002	0.002	1	4,430.315	0.002	0.963

Behavioral results: post-hoc tests

Decision task

Table S23. EMMs for partner across all group and risk combinations

Partner	Group	Risk	*logit*	*SE (logit)*	*95% CI (logit)*	*probability*	*95% CI (probability)*
dependent	CON	high-risk	0.653	0.208	[0.246, 1.060]	0.658	[0.561, 0.743]
independent	CON	high-risk	-1.454	0.215	[-1.876, -1.032]	0.189	[0.133, 0.263]
dependent	BPD	high-risk	0.242	0.219	[-0.186, 0.671]	0.560	[0.454, 0.662]
independent	BPD	high-risk	-1.635	0.227	[-2.080, -1.189]	0.163	[0.111, 0.233]
dependent	CON	low-risk	1.470	0.210	[1.058, 1.882]	0.813	[0.742, 0.868]
independent	CON	low-risk	-1.537	0.221	[-1.971, -1.104]	0.177	[0.122, 0.249]
dependent	BPD	low-risk	0.631	0.219	[0.202, 1.061]	0.653	[0.550, 0.743]
independent	BPD	low-risk	-2.147	0.237	[-2.611, -1.682]	0.105	[0.068, 0.157]

Table S24. EMMs pairwise contrasts of partner across all group and risk combinations

Contrast	Group	Risk	*OR*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
dependent / independent	CON	high-risk	8.222	0.766	22.627	[6.851, 9.868]	[6.516, 10.375]	**< .001**	**< .001**
dependent / independent	BPD	high-risk	6.536	0.666	18.413	[5.352, 7.981]	[5.066, 8.431]	**< .001**	**< .001**
dependent / independent	CON	low-risk	20.236	2.190	27.784	[16.367, 25.018]	[15.442, 26.518]	**< .001**	**< .001**
dependent / independent	BPD	low-risk	16.085	1.916	23.324	[12.736, 20.314]	[11.946, 21.658]	**< .001**	**< .001**

Notes: Adjusted using Bonferroni. Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 4 estimates Intervals are back-transformed from the log odds ratio scale P value adjustment: bonferroni method for 4 tests Tests are performed on the log odds ratio scale

Table S25. EMMs for risk across all risk and partner combinations

Risk	Group	Partner	*Logit*	*SE (logit)*	*95% CI (logit)*	*Probability*	*95% CI (probability)*
high-risk	CON	dependent	0.653	0.208	[0.246, 1.060]	0.658	[0.561, 0.743]
low-risk	CON	dependent	1.470	0.210	[1.058, 1.882]	0.813	[0.742, 0.868]
high-risk	BPD	dependent	0.242	0.219	[-0.186, 0.671]	0.560	[0.454, 0.662]
low-risk	BPD	dependent	0.631	0.219	[0.202, 1.061]	0.653	[0.550, 0.743]
high-risk	CON	independent	-1.454	0.215	[-1.876, -1.032]	0.189	[0.133, 0.263]
low-risk	CON	independent	-1.537	0.221	[-1.971, -1.104]	0.177	[0.122, 0.249]
high-risk	BPD	independent	-1.635	0.227	[-2.080, -1.189]	0.163	[0.111, 0.233]
low-risk	BPD	independent	-2.147	0.237	[-2.611, -1.682]	0.105	[0.068, 0.157]

Table S26. EMMs pairwise contrasts of risk across all group and partner combinations

Contrast	Group	Partner	*OR*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
high-risk / low-risk	CON	dependent	0.442	0.036	-10.108	[0.377, 0.517]	[0.361, 0.540]	**< .001**	**< .001**
high-risk / low-risk	BPD	dependent	0.678	0.056	-4.672	[0.576, 0.798]	[0.551, 0.834]	**< .001**	**< .001**
high-risk / low-risk	CON	independent	1.087	0.120	0.756	[0.876, 1.349]	[0.825, 1.431]	0.450	1.000
high-risk / low-risk	BPD	independent	1.668	0.197	4.343	[1.324, 2.102]	[1.243, 2.239]	**< .001**	**< .001**

Notes: Adjusted using Bonferroni. Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 4 estimates Intervals are back-transformed from the log odds ratio scale P value adjustment: bonferroni method for 4 tests Tests are performed on the log odds ratio scale

Table S27. EMMs pairwise contrasts of risk across partner (averaged over group)

Contrast	Partner	*OR*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
high-risk / low-risk	dependent	0.547	0.034	-9.654	[0.484, 0.618]	[0.476, 0.629]	< .001	**< .001**
high-risk / low-risk	independent	1.347	0.136	2.948	[1.105, 1.641]	[1.074, 1.688]	0.003	**0.006**

Notes: Adjusted using Bonferroni. Results are averaged over the levels of: group Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 2 estimates Intervals are back-transformed from the log odds ratio scale P value adjustment: bonferroni method for 2 tests Tests are performed on the log odds ratio scale

Table 28. EMMs pairwise contrasts of group across risk (averaged over partner)

Contrast	Risk	*OR*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
CON / BPD	high-risk	1.344	0.402	0.987	[0.747, 2.416]	[0.687, 2.629]	0.324	0.647
CON / BPD	low-risk	2.063	0.624	2.393	[1.140, 3.732]	[1.047, 4.063]	0.017	**0.033**

Notes: Adjusted using Bonferroni. Results are averaged over the levels of: partner_factor Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 2 estimates Intervals are back-transformed from the log odds ratio scale P value adjustment: bonferroni method for 2 tests Tests are performed on the log odds ratio scale

Expectation task

Table S29. EMMs pairwise contrasts of group across partner (averaged over risk)

Contrast	Partner	*OR*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
CON / BPD	dependent	2.108	0.514	3.060	[1.307, 3.399]	[1.221, 3.640]	**0.002**	**0.004**
CON / BPD	independent	0.751	0.217	-0.993	[0.427, 1.322]	[0.393, 1.433]	0.320	0.641

Notes: Adjusted using Bonferroni. Results are averaged over the levels of: risk_factor Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 2 estimates Intervals are back-transformed from the log odds ratio scale P value adjustment: bonferroni method for 2 tests Tests are performed on the log odds ratio scale

Table 30. EMMs pairwise contrasts of risk across group (averaged over partner)

Contrast	Group	*OR*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
high-risk / low-risk	CON	0.301	0.038	-9.406	[0.235, 0.387]	[0.227, 0.401]	**< .001**	**< .001**
high-risk / low-risk	BPD	0.403	0.049	-7.524	[0.318, 0.510]	[0.307, 0.528]	**< .001**	**< .001**

Notes: Adjusted using Bonferroni. Results are averaged over the levels of: risk_factor Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 2 estimates Intervals are back-transformed from the log odds ratio scale P value adjustment: bonferroni method for 2 tests Tests are performed on the log odds ratio scale

Response times decision task

Table S31. EMMs pairwise contrasts of risk across partner (averaged over risk)

Contrast	Partner	*Ratio*	*SE*	*z-ratio*	*95% CI (raw)*	*95% CI (Bonferroni)*	*p (raw)*	*p-adj (Bonferroni)*
high-risk / low-risk	dependent	1.082	0.009	9.982	[1.065, 1.099]	[1.063, 1.101]	< .001	**< .001**
high-risk / low-risk	independent	0.966	0.011	-2.975	[0.944, 0.988]	[0.941, 0.992]	0.003	**0.006**

Notes: Adjusted using Bonferroni. Results are averaged over the levels of: group Degrees-of-freedom method: asymptotic Confidence level used: 0.95 Conf-level adjustment: bonferroni method for 2 estimates Intervals are back-transformed from the log scale P value adjustment: bonferroni method for 2 tests Tests are performed on the log scale

fMRI results

Table S32. Main effect of partner (dependent > independent; t > 3.1339; p < 0.0010; df = 189; minimum extent = 120)

Region Label	Extent	t-value	MNI Coordinates x	y	z
R middle occipital gyrus	7837	9.422	36	-88	14
R inferior frontal gyrus, opercular part	825	6.136	36	6	28
L precentral gyrus extending to inferior frontal gyrus, opercular part	508	5.819	-36	20	20
L superior frontal gyrus	204	5.448	-24	2	48
L supplementary motor area	236	5.351	-8	10	54
R angular gyrus	120	5.039	48	-58	30
R superior frontal gyrus	155	4.878	24	52	28
R postcentral gyrus	216	4.439	46	-32	50

Table S33. Main effect of partner (independent > dependent; t > 3.1339; p < 0.0010; df = 189; minimum extent = 121)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R superior frontal gyrus, medial orbital	405	6.326	6	52	-12
R parahippocampal gyrus extending into hippocampus and insula	248	5.973	18	-8	-20
L supramarginal gyrus	522	5.904	-64	-36	28
R inferior frontal gyrus, triangular part	180	5.798	52	38	2
R lingual gyurs	240	5.155	10	-74	-6
R middle temporal gyrus	1252	5.145	64	-48	2
L hippocampus	255	5.129	-24	-16	-20
L middle temporal gyrus	148	5.103	-62	-54	-6
R posterior orbital gyrus	191	5.078	24	30	-12
L superior frontal gyrus	134	4.923	-24	26	38
L inferior frontal gyrus, pars orbitalis	121	4.836	-38	36	-12
L inferior frontal gyrus, triangular part	229	4.698	-52	34	0
L middle cingulate & paracingulate gyri	210	4.691	-8	-36	44
L superior frontal gyrus, medial	149	4.576	-8	46	26
L insula	130	4.203	-34	-18	-6

Table S34. Main effect of partner in mPFC ROI (mask-based cluster inference; independent > dependent; $t > 3.13$; $p < 0.001$; $df = 189$; minimum extent = 121)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
L superior frontal gyrus, medial	142	4.576	-8	46	26

Table S35. Group x partner x risk interaction (CON > BPD, independent > dependent, low > high; t > 3.1339; p < 0.0010; df = 189; minimum extent = 120)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R middle temporal gyrus	152	5.246	52	-70	6
R middle temporal gyrus	224	5.018	58	-48	6
L superior frontal gyrus, medial orbital	120	4.795	-4	32	-12
L lingual gyurs	131	4.507	-12	-58	2
L inferior frontal gyrus	189	4.435	-52	28	12
R inferior frontal gyrus, triangular part	126	4.320	52	24	12
L calcarine fissure	233	4.247	-2	-96	8

Table S36. Main effect of risk (high > low; t > 3.1339; p < 0.0010; df = 189; minimum extent = 224)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
L insula extending into inferior frontal gyrus, triangular part	531	7.959	-34	26	0
L lingual gyrus extending into	3481	7.674	-12	-76	-4
R insula	617	7.613	34	22	10
L middle cingulate & paracingulate gyri	1432	6.781	-6	20	36
L postcentral gyrus	955	5.913	-44	-20	52
R superior parietal gyrus	1400	5.794	18	-60	62
R precentral gyrus	224	4.981	42	-6	48
R inferior occipital gyrus	368	4.879	38	-82	-2

Table S37. Main effect of risk (low > high; t > 3.1339; p < 0.0010; df = 189; minimum extent = 278)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
L middle frontal gyrus	3562	7.320	-44	38	16
R middle temporal gyrus	1435	7.003	48	-60	22
L supramarginal gyurs	1824	6.922	-62	-40	36
R inferior frontal gyrus, triangular part	519	6.711	44	40	0
R medial orbital gyrus	278	6.233	22	34	-16
R supramarginal gyrus	330	5.983	60	-18	26
R insula	290	5.584	34	-4	12
R superior frontal gyrus	412	5.583	26	20	50

Table S38. Main effect of risk in mPFC ROI (mask-based cluster inference; high > low; t > 3.13, p < 0.001; df = 189; minimum extent = 224)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
L middle cingulate & paracingulate gyri	468	6.781	-6	20	36

Table S39. Main effect of risk in mPFC ROI (mask-based cluster inference; low > high; t > 3.13; p < 0.001; df = 189; minimum extent = 278)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R superior frontal gyrus, medial	404	5.341	12	64	18

Table S40. Group x risk interaction (BPD > CON, high > low; t > 3.1339; p < 0.0010; df = 189; minimum extent = 134)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R calcarine fissure	134	4.358	16	-84	14

Table S41. Partner x risk interaction (independent > dependent, low > high; t > 3.1339; p < 0.0010; df = 189; minimum extent = 149)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R precuneus	1102	5.901	6	-56	20
Brainstem	149	5.512	8	-28	-4
L caudate nucleus	206	5.390	-10	24	16
L anterior cingulate cortex, subgenual	212	5.015	0	36	-4
L postcentral gyrus	507	5.011	-44	-18	56

Table S42. Partner x risk interaction in mPFC ROI (independent > dependent, low > high; t > 3.13; p < 0.001; df = 189; minimum extent = 149)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
L anterior cingulate cortex, subgenual	199	5.015	0	36	-4

Table S43. Main effect of outcome (positive >negative; t > 3.2247; p < 0.0010; df = 63; minimum extent = 121)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
L caudate nucleus	21908	9.420	-18	0	26
L superior occipital gyrus	757	7.039	-12	-98	6
L inferior frontal gyrus, pars orbitalis	147	5.058	-38	42	-12
L superior frontal gyrus	121	4.498	-22	32	54

Table S44. Main effect of outcome (negative > positive; t > 3.2247; p < 0.0010; df = 63; minimum extent = 148)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R insula extending into inferior frontal gyrus, triangular part	1223	6.155	50	20	6
L insula extending into inferior frontal gyrus, triangular part	580	5.707	-40	20	2
R middle temporal gyrus	435	5.640	54	2	-20
L fusiform gyrus	217	5.603	-26	-68	-10
L superior frontal gyrus, medial	302	5.383	-8	50	16
L middle temporal gyrus	148	5.344	-56	-18	-12
L middle temporal gyrus	191	5.340	-56	-36	-2
R anterior cingulate cortex, supracallosal	407	5.325	6	32	28
R superior frontal gyrus	177	5.040	26	52	12

Table S45. Group x outcome interaction (BPD > CON; positive > negative; t > 3.2247; p < 0.0010; df = 63; minimum extent = 216)

			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
R superior parietal gyrus	216	4.148	28	-66	60

Table S46. mPFC ROI – Type III F-tests (Satterthwaite's approximation)

Term	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F value*	*Pr(>F)*
Group	0.003	0.003	1	63	0.011	0.918
Partner	1.575	1.575	1	189	5.019	**0.026**
Risk	0.521	0.521	1	189	1.659	0.199
Group x Partner	0.036	0.036	1	189	0.115	0.734
Group x Risk	0.048	0.048	1	189	0.154	0.695
Partner x Risk	0.001	0.001	1	189	0.003	0.959
Group x Partner x Risk	1.496	1.496	1	189	4.769	**0.030**

Table S47. dmPFC ROI – Type III F-tests (Satterthwaite's approximation)

Term	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F value*	*Pr(>F)*
Group	0.350	0.350	1	63	0.506	0.480
Partner	2.403	2.403	1	189	3.470	0.064
Risk	0.070	0.070	1	189	0.101	0.751
Group x Partner	0.254	0.254	1	189	0.367	0.545
Group x Risk	0.077	0.077	1	189	0.111	0.739
Partner x Risk	1.543	1.543	1	189	2.228	0.137
Group x Partner x Risk	2.122	2.122	1	189	3.064	0.082

References

Bach, B., Kerber, A., Aluja, A., Bastiaens, T., Keeley, J. W., Claes, L., Fossati, A., Gutierrez, F., Oliveira, S. E. S., Pires, R., Riegel, K. D., Rolland, J.-P., Roskam, I., Sellbom, M., Somma, A., Spanemberg, L., Strus, W., Thimm, J. C., Wright, A. G. C., & Zimmermann, J. (2020). International Assessment of DSM-5 and ICD-11 Personality Disorder Traits: Toward a Common Nosology in DSM-5.1. *Psychopathology*, *53*(3–4), 179–188. https://doi.org/10.1159/000507589

Baudson, T. G., & Preckel, F. (2016). mini-q: Intelligenzscreening in drei Minuten. *Diagnostica*, *62*(3), 182–197. https://doi.org/10.1026/0012-1924/a000150

Beck, A. T., Steer, R. A., & Brown, G. K. (2006). *BDI-II. Beck-Depressionsinventar [Beck Depression Inventory II]*. Pearson Verlag.

Bohus, M., Kleindienst, N., Limberger, M. F., Stieglitz, R.-D., Domsalla, M., Chapman, A. L., Steil, R., Philipsen, A., & Wolf, M. (2009). The Short Version of the Borderline Symptom List (BSL-23): Development and Initial Data on Psychometric Properties. *Psychopathology*, *42*(1), 32–39. https://doi.org/10.1159/000173701

Borkenau, P., & Ostendorf, F. (2007). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (NEO-FFI; 2. Neu normierte und vollständig überarb. Aufl.)*. Hogrefe.

Damovsky, F., Zettl, M., Zimmermann, J., Herbold, W., Curtius, T., Buecker, S., Taubner, S., & Volkert, J. (2021). *Das Persönlichkeitsinventar für ICD-11: Reliabilität und Validität der deutschen Version in einer klinischen und nicht-klinischen Stichprobe* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/cm3qv

Franke, G. H. (2017). *BSCL-53®-S. Brief symptom-checklist–standard–German manual*. Hogrefe Publishing Group.

Hofmann, S. G., & Kashdan, T. B. (2010). The Affective Style Questionnaire: Development and Psychometric Properties. *Journal of Psychopathology and Behavioral Assessment*, *32*(2), 255–263. https://doi.org/10.1007/s10862-009-9142-4

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., Howes, M. J., Jin, R., Secnik, K., Spencer, T., Ustun, T. B., & Walters, E. E. (2005). The World Health Organization adult ADHD self-report scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, *35*(2), 245–256. https://doi.org/10.1017/S0033291704002892

Kröger, C., & Kosfelder, J. (2011). *IES-27: Skala zur Erfassung der Impulsivität und emotionalen Dysregulation der Borderline-Persönlichkeitsstörung*. Hogrefe.

Loranger, A. W., Sartorius, N., & Andreoli, A. (1994). The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health

Administration International Pilot Study of Personality Disorders. *Archives of General Psychiatry*, *51*(3), 215. https://doi.org/10.1001/archpsyc.1994.03950030051005

Margraf, J., & Cwik, J. C. (2017). *Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen* [PDF]. https://doi.org/10.13154/RUB.102.91

Oltmanns, J. R., & Widiger, T. A. (2018). A self-report measure for the ICD-11 dimensional trait model proposal: The Personality Inventory for ICD-11. *Psychological Assessment*, *30*(2), 154–169. https://doi.org/10.1037/pas0000459

Preuss, U. W., Rujescu, D., Giegling, I., Watzke, S., Koller, G., Zetzsche, T., Meisenzahl, E. M., Soyka, M., & Möller, H. J. (2008). Psychometrische Evaluation der deutschsprachigen Version der Barratt-Impulsiveness-Skala. *Der Nervenarzt*, *79*(3), 305–319. https://doi.org/10.1007/s00115-007-2360-7

Pruessner, L., Holt, D. V., Gölz, R., Sevcenko, N., Hofmann, S. G., & Backenstrass, M. (2020). Psychometrische Eigenschaften der deutschsprachigen Version des Interpersonal Emotion Regulation Questionnaire. *Diagnostica*, *66*(1), 62–73. https://doi.org/10.1026/0012-1924/a000239

Retz-Junginger, P., Retz, W., Blocher, D., Weijers, H.-G., Trott, G.-E., Wender, P. H., & Rössler, M. (2002). Wender Utah Rating Scale (WURS-k) Die deutsche Kurzform zur retrospektiven Erfassung des hyperkinetischen Syndroms bei Erwachsenen: Die deutsche Kurzform zur retrospektiven Erfassung des hyperkinetischen Syndroms bei Erwachsenen. *Der Nervenarzt*, *73*(9), 830–838. https://doi.org/10.1007/s00115-001-1215-x

Rohrmann, S., Hodapp, V., Schnell, K., Tibubos, A. N., Schwenkmezger, P., & Spielberger, C. D. (2013). *Das State-Trait-Ärgerausdrucks-Inventar-2 (STAXI-2). Deutschsprachige Adaptation des State-Trait Anger Expression Inventory-2 (STAXI-2) von Charles D. Spielberger [State-Trait Anger-Expression-Inventory-2 (STAXI-2). German Adaptation of the State-Trait-Anger-Expression-Inventory-2 (STAXI-2) of Charles D. Spielberger]*. Hans Huber.

Rösler, M., Retz, W., Retz-Junginger, P., Thome, J., Supprian, T., Nissen, T., Stieglitz, R.-D., Blocher, D., Hengesch, G., & Trott, G. E. (2004). Instrumente zur Diagnostik der Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS) im Erwachsenenalter: Selbstbeurteilungsskala (ADHS-SB) und Diagnosecheckliste (ADHS-DC). *Der Nervenarzt*, *75*(9). https://doi.org/10.1007/s00115-003-1622-2

Spitzer, C., Müller, S., Kerber, A., Hutsebaut, J., Brähler, E., & Zimmermann, J. (2021). Die deutsche Version der Level of Personality Functioning Scale-Brief Form 2.0 (LPFS-BF): Faktorenstruktur, konvergente Validität und Normwerte in der Allgemeinbevölkerung. *PPmP -*

Psychotherapie · Psychosomatik · Medizinische Psychologie, *71*(07), 284–293. https://doi.org/10.1055/a-1343-2396

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *The multidimensional mental state questionnaire: Manual*. Hogrefe.

Weekers, L. C., Hutsebaut, J., & Kamphuis, J. H. (2019). The Level of Personality Functioning Scale-Brief Form 2.0: Update of a brief instrument for assessing level of personality functioning: The Level of Personality Functioning Scale - Brief Form 2.0. *Personality and Mental Health*, *13*(1), 3–14. https://doi.org/10.1002/pmh.1434

Zimmermann, J., Altenstein, D., Krieger, T., Holtforth, M. G., Pretsch, J., Alexopoulos, J., Spitzer, C., Benecke, C., Krueger, R. F., Markon, K. E., & Leising, D. (2014). The Structure and Correlates of Self-Reported DSM-5 Maladaptive Personality Traits: Findings From Two German-Speaking Samples. *Journal of Personality Disorders*, *28*(4), 518–540. https://doi.org/10.1521/pedi_2014_28_130

Study III - Supplementary material

Supplementary material

List of interviews and questionnaires

Supplementary Table S1. Interviews and Questionnaires

Name	Type
Mini-DIPS Open Access (German version of the Anxiety Disorders Interview Schedule; 1)	Interview
IPDE (International Personality Disorder Examination; 2)	Interview
BSL-23 (Borderline Symptom List-23; 3)	Self-report
BSCL (Brief-Symptom-Checklist; 4)	Self-report
BDI-II (Beck Depression Inventory; 5)	Self-report
ASRS-v1.1 (Adult ADHD Self-Report Scale; 6)	Self-report
ADHS-SB (ADHS-Self-Report for adults; 7)	Self-report
WURS-K (German version of the Wender Utah Rating Scale; 8)	Self-report
MDBF (Multidimensional Mood Questionnaire – Short Version A; 9)	Self-report
NEO-FFI (German version; 10)	Self-report
PID-5-BF (German version of the Personality Inventory for DSM-5-Brief Form; 11,12)	Self-report
STAXI-2 (German Version of the State-Trait Anger Expression Inventory-2; 13)	Self-report
BIS-11 (German Version of the Barratt Impulsiveness Scale; 14)	Self-report
IERQ (Interpersonal Emotion Regulation Questionnaire; 15,16)	Self-report
mini-q (A Three-Minute Intelligence Screening; 17)	Self-report
PiCD (The Personality Inventory for ICD-11; 18,19)	Self-report
LPFS-BF (Level of Personality Functioning Scale-Brief Form; 20,21)	Self-report
IES-27 (Impulsivity and Emotion Dysregulation Scale for Borderline Personality Disorder; 22)	Self-report

Supplementary Methods

Participants

Power analysis

We performed a power analysis using G*Power version 3.1.9.2 (23) for sample size estimation, based on data from a previous study (24) on group differences in updating for undesirable feedback in BPD patients (N = 22) and healthy controls (N = 42). The effect size in Korn et al. (24) was d = 0.65. With an alpha criterion of .05 and a power of 1-β = .80, the minimum sample size needed is N = 30 for the patients' group and N = 30 for the control group for a one-tailed two-sample t-test.

Inclusion and exclusion procedure

We excluded individuals with severe neurological disorders or head/brain injuries, severe physical illness, BMI < 15, severe visual impairments, cognitive disorders, lifetime diagnosis of schizophrenia, schizoaffective or bipolar disorder, or acute suicidality from participation. Additionally, psychology students, except those in their first semester, were not included.

Participants with BPD had to meet four or more BPD criteria according to the IPDE interview to be included (i.e., we also included participants with sub-threshold BPD who met less than the threshold of 5 criteria but had a prior diagnosis). For the BPD group further exclusion criteria included current diagnosis of narcissistic or avoidant personality disorder, current moderate or severe alcohol or drug abuse, and the use of psychotropic medication, except for antidepressants, anxiolytic medication, on-demand medication, and limited Quetiapin prescriptions for sleep efficiency (up to 100 mg retarded and 50 mg non-retarded). Control group participants were also subject to additional exclusion criteria, including current alcohol or drug abuse, lifetime diagnosis of personality disorders, other psychiatric diagnoses (within the last 5 years), recent psychological or psychiatric interventions (within the last 5 years), presence of more than one current BPD symptoms (according to the IPDE), current clinically relevant symptoms (assessed via the Mini-DIPS, 1), and the use of psychotropic medication.

We planned to collect 30 valid datasets per group (i.e., datasets that fulfill our preregistered inclusion criteria). We started collecting data in October 2021 and completed data collection in May 2023 after collecting 60 valid datasets for all experiments of the larger project. Since all participants who completed the behavioral or fMRI experiments mentioned above also completed the behavioral tasks of this study, we ended up with more than 60 datasets (we collected 100 datasets in total, 44 patients and 56 controls). We excluded nine patients (one because of task comprehension problems, two because of intake of medication listed in our exclusion criteria, four because they fulfilled less than four IPDE criteria, two because of drug use before testing) and six controls (due to clinically relevant symptoms assessed via the diagnostic interview) before data analysis.

Task specific exclusion procedure

We applied additional preregistered exclusion criteria for the different tasks resulting in slightly different sample sizes. For the analysis of the SVO scores and categories, we excluded participants with intransitive choices (see SVO task in Analyses section) in the primary items (25) resulting in a sample of 29 patients and 50 controls. For the analysis of the prosocial motivation scores of the SVO task, we could only include participants with prosocial SVO scores according to both primary and secondary items (see Analyses and 26), resulting in a sample of 24 patients and 44 controls. Although we did not initially specify any exclusion criteria for the UG, we decided to exclude one participant from the control group due to an implausible response pattern which indicates that this participant misunderstood the task (specifically, the participant accepted all unfair offers from 1 to 3 and rejected all fair offers from 4 to 6). Consequently, the sample sizes for the UG analysis were 35 participants in the BPD group and 49 participants in the control group.

We did not have to exclude any of the participants due to implausible response patterns in the Fairness Ratings or the JPE task. In this case, we defined “implausible” as selecting the same choice option more than 90% of the trials in the Fairness Ratings.

Ordering and randomization of the tasks

The order of the tasks was not randomized on purpose. We chose a specific order to minimize the influence between the tasks. That is, tasks on social preferences (SVO) and evaluation of allocations (JPE) were presented before participants were asked to actively allocate money between themselves and other participants (DG). Participants performed the active allocation (DG) before responding to allocations from others (UG). We asked them about the minimum offer they would be willing to accept (Minimum Acceptance Rating) after they accepted or rejected several offers (UG). Finally, we asked the participants to rate the fairness of hypothetical offers because we wanted to be able to differentiate between the decisions that participants actually make, and the ones they consider fair.

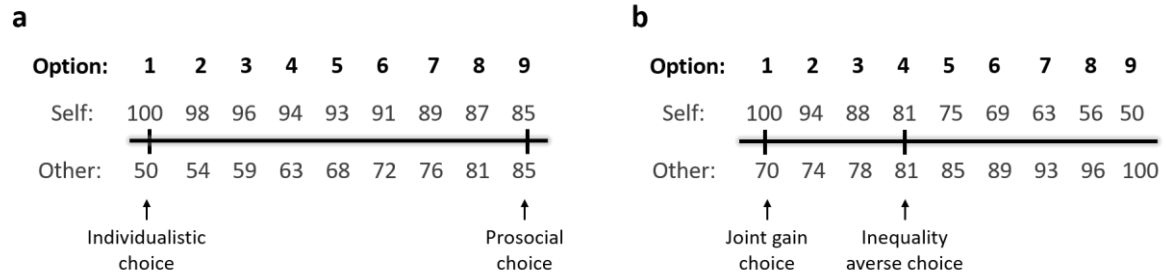
For the items in the SVO task, we used the two different versions of the task (same items in different order) provided by Murphy (http://ryanomurphy.com/styled-2/downloads/index.html). The version of the SVO task given to the participants was based on their individual subject number. The order of the items within the UG task, the JPE task, and the Fairness Ratings were fully randomized. All participants saw the exact same items in a different order. For randomization, we use the MATLAB function *randperm*. All tasks were presented using the MATLAB toolbox Psychtoolbox (psychtoolbox.org). Response time was not restricted in any of the tasks.

Experimental procedures

Participants were recruited via (online) advertisements, flyers, and existing databases. For the recruitment of participants with BPD, we contacted psychotherapists in Heidelberg and surrounding towns and recruited inpatients at the Department of General Psychiatry at the University of Heidelberg. Healthy participants were paid a minimum of 50€ and participants with BPD were paid a minimum of 60€ (due to differences in the amount of time required for the diagnostic interview) for completing all experiments that are part of the project. In addition, all participants received a bonus depending on their decisions during the experiments. In case participants participated only in some of the experiments, they were paid a standard hourly fee.

Tasks

SVO task



Supplementary Figure S1. Example items of the SVO Slider task. (a) Primary item 6 measures general Social Value Orientation and (b) secondary item 9 measures prosocial motivation (two different motives for being prosocial: joint gain maximization and inequality aversion).

Supplementary analyses

Calculation of SVO scores

The items of the SVO task are sampled from an allocation plane defined by two orthogonal dimensions: outcomes for self (participant) and outcomes for other (anonymous participant). The endpoints (options 1 and 9 of each item) lie on a circle centered at 50 with a diameter of 50. The SVO score is an angle on this allocation plane (see Supplementary Material for details of calculation).

Participants with intransitive responses for the primary items were excluded from the analyses. Transitivity implies that if a person prefers choice A over choice B, and choice B over choice C, then they should also prefer choice A over choice C. However, if responses are random, it is highly likely that the choices will be intransitive, meaning they won't follow this logical order of preferences.

To calculate the SVO score, we determined the individual mean allocation for self and the other from the primary items, then subtracted 50 from both means to shift the base of the resulting SVO angle to the center of the circle. The SVO angle for each participant was calculated by taking the inverse tangent of the ratio of the adjusted means. SVO scores can be classified into four categories: competitive (less than -12.04°), individualistic (between -12.04° and 22.45°), prosocial (between 22.45° and 57.15°), and altruistic (more than 57.15°).

For computing the prosocial motivation score from the nine secondary items, we first calculated the "inequality distance score" which quantifies the distance between the participant's chosen options and the options that would minimize the inequality between the payoffs for self and the other. Second, we calculated the "joint gain distance score," which quantifies the distance between the participant's chosen options and those that would maximize joint gain. The prosocial motivation score was derived by dividing the inequality distance score by the sum of the inequality distance score and the joint-gain

distance score. This results in a single score ranging from 0 (perfect inequality aversion) to 1 (perfect joint gain maximization). Based on this score, participants can be categorized as either inequality averse (prosocial motivation score < 0.5) or joint gain maximizing (> 0.5). Participants who scored exactly 0.5 are not categorized (26).

Economic models for the JPE task

We compared the following economic models (see Supplementary Material in 27 for a detailed model description):
In all models, Utility (U) depends on various (combinations of) weighted predictors. In the first five models, utility is determined by the weighted payoff for Self ($\$Self$) and/or other ($\$Other$), or a combination of these payoffs (e.g., the sum, the joint payoff, and/or the absolute difference between payoffs). The Fehr-Schmidt model (28) separately considers the payoff for Self, Disadvantageous Inequality, and Advantageous Inequality. If the difference between $Other and $Self is negative, the term max(*$Other - $Self*,0) sets the result to zero (preserving positive values). The same logic applies to the difference between $Self and $Other (*$Self - $Other*). Parameters were not restricted, allowing for the possibility of altruistic and competitive motives. The ERC model (29) accounts for quadratic and normalized inequality, without explicitly distinguishing between advantageous and disadvantageous inequality.

1. Individualism model:
   $U = \alpha * \$Self$
2. Prosocial model:
   $U = \alpha * \$Self + \beta * \$Other$
3. Altruism model:
   $U = \beta * \$Other$
4. Joint-gain model:
   $U = \alpha * (\$Self + \$Other)$
5. Inequality model:
   $U = \alpha * abs(\$Self - \$Other)$
6. Inequality & Joint-gain model:
   $U = \alpha * abs(\$Self - \$Other) + \beta * (\$Self + \$Other)$
7. Fehr-Schmidt model:
   $U = \alpha * \$Self + \beta * \max(\$Other - \$Self, 0) + \gamma * \max(\$Self - \$Other, 0)$

8. Fehr-Schmidt model with separate parameters for positive and negative payoffs for self and disadvantageous inequality so that:

   $\alpha = \alpha_{pos}$ for \$Self ≥ 0,

   $\alpha = \alpha_{neg}$ for \$Self < 0,

   $\beta = \beta_{pos}$ for \$Self ≥ 0,

   $\beta = \beta_{neg}$ for \$Self ≥ 0,

9. ERC (Equity, Reciprocity, Competition) model:

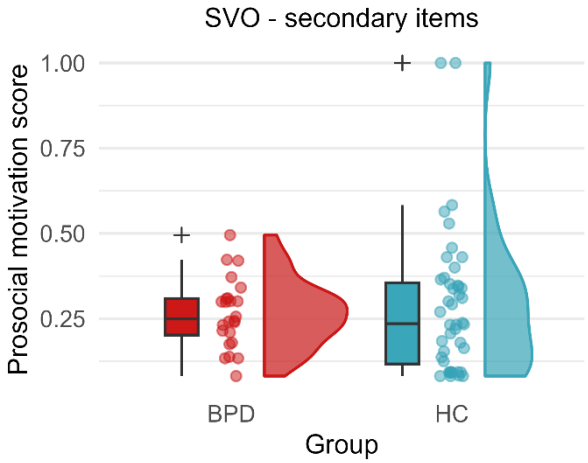
   $$U = \alpha * \$Self + \frac{\beta}{2} * \left(\frac{\$Self}{\$Self + \$Other} - \frac{1}{2}\right)^2$$

To assess model evidence, we conducted both fixed-effects analysis (assumption that all participants use the same model) and random-effects analysis (assumption that different participants may use different models). For the fixed-effects analysis, we calculated log-group Bayes factors. We first, calculated the Bayesian information criterion (BIC; 30) for each participant and each model, using the standard formula: BIC = n * ln(RSS/n) + k * ln(n), where RSS is the residual sum of squared errors, n is the number of trials, k is the number of free parameters in the model, and "ln" is the natural logarithm. The BIC penalizes model complexity (i.e., the number of parameters in the model). To compute log-group Bayes factors, we summed the BIC scores across the group and subtracted the BIC of a reference model from the BIC of each tested model (31). The model with the smallest log-group Bayes factor is considered the best fit.

For the random-effects analysis, we calculated protected exceedance probabilities using the Bayesian Model Selection (BMS), a procedure implemented in the MATLAB toolbox SPM12 (http://www.fil.ion.ucl.ac.uk/spm/;spm_BMS). The protected exceedance probability indicates the probability that a given model is more likely across subjects than any other model, accounting for differences in model evidence due to chance (32).

Supplementary results

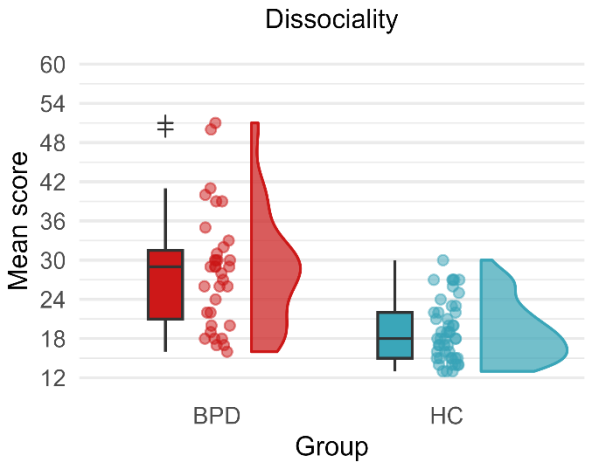
SVO prosocial motivation



Supplementary Figure S2. Active cooperation task results. Raincloud plots show the distribution of prosocial motivation scores (secondary items of the SVO task) in the BPD group and the HC group.

Dissociality scores

Since we used dissociality scores to examine their potential association with reduced reactive cooperation in the UG, we also tested whether dissociality scores varied between the groups. We found that participants from the BPD group scored significantly higher (M_{BPD} = 28.314, SD_{BPD} = 8.878, N_{BPD} = 35) than participants form the control group (M_{HC} = 19.082, SD_{HC} = 4.559, N_{HC} = 49; Welch's two-sample t-test $t_{46.839}$ = 5.644, $p < 0.001$, 95% CI [5.942, 12.524], d = 1.31; Wilcoxon rank sum test z = 5.157, $p < 0.001$, r = 0.563; **Supplementary Figure S3**). Note that the sample size used here is the same as in the UG analysis.



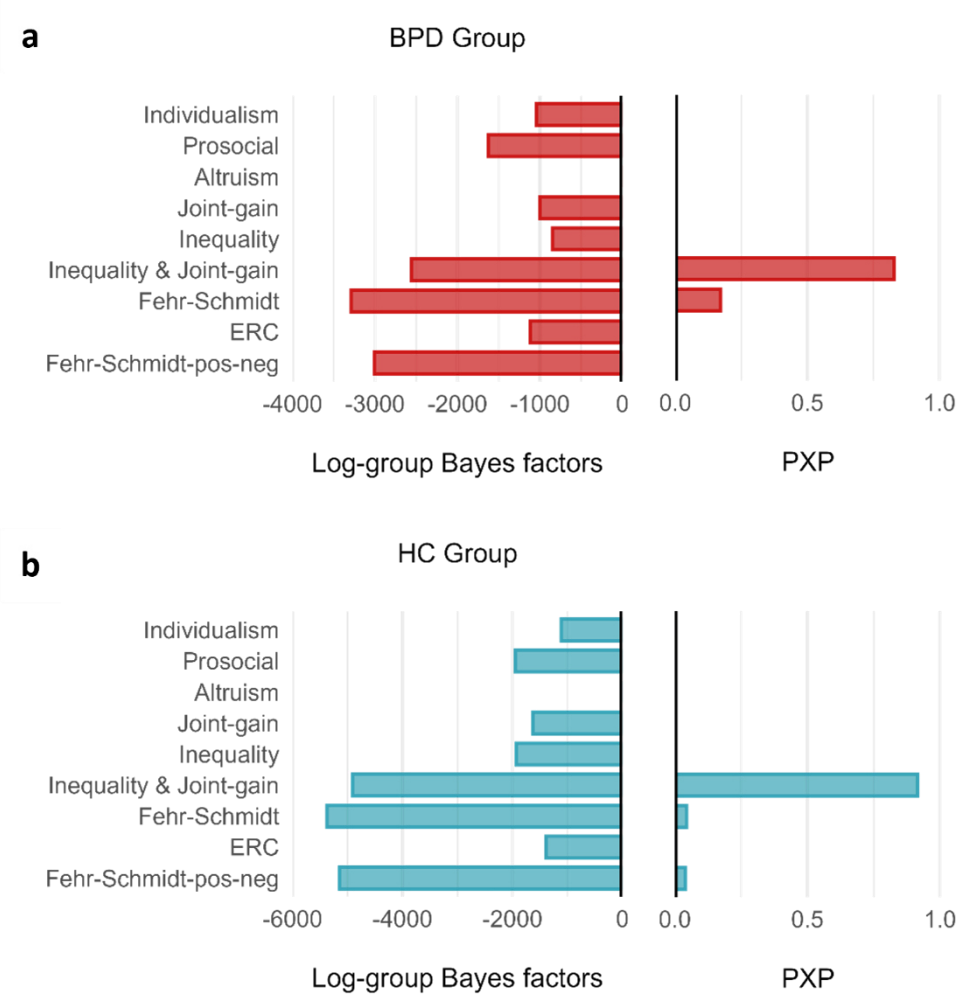
Supplementary Figure S3. Raincloud plots show the distribution of dissociality scores in the BPD group and the HC group. The plot includes individual dissociality scores for each participant, a box plot, and a probability density plot. The boxplot visualizes the median, the 25th and 75th percentiles, two whiskers (extending from the upper and lower hinges to the largest and lowest value within 1.5 times the inter-quartile range, and all outliers individually. The density plot was generated using Gaussian kernel density estimation.

Fairness ratings

Supplementary Table S2. Means and standard deviations of Fairness Ratings

	BPD		HC	
Offer range	*M*	*SD*	*M*	*SD*
Disadvantageous offers	1.921	0.65	2.355	1.001
Fair offers	6.629	1.081	6.473	1.518
Advantageous offers	3.029	1.84	2.780	1.317

JPE model comparison



Supplementary Figure S4. Model comparisons of the JPE task data according to fixed-effects analyses (Log-group Bayes factors) and random-effects analyses (protected exceedance probabilities, PXPs). Log-group Bayes factors are presented relative to a reference model (Altruism model). Lower numbers indicate better model fit in the Log-group Bayes factors and higher numbers indicate better model fit in the PXPs. The Fehr-Schmidt model performed best in both groups according to the fixed-effects analyses, and the Inequality & Joint-gain model performed best in both groups according to the random-effects analyses.

Mini meta-analysis on ultimatum games

Eligibility criteria

We included studies that met the following criteria: (a) experimental designs, (b) assessment based on a clinical diagnosis of Borderline Personality Disorder (BPD) or BP(D) features, (c) use of a variant of the Ultimatum Game, and (d) design with participants in the responder role. For studies involving diagnosed BPD samples, (e) a comparison with a healthy control group was required. Additionally, (f) the studies had to measure rejections in the Ultimatum Game as the primary outcome. Only studies published in English or German were considered eligible for inclusion. Both published and unpublished studies were included, and we actively searched for gray literature.

Search strategy and selection

We conducted a comprehensive search for eligible studies using the databases PubMed and PsycINFO (last search on October 3, 2023). The search strings used were "Borderline AND Ultimatum Game". Due to the limited number of studies found, we extended our search by issuing a call for data through the German Psychological Society (DGPs) and by reaching out to researchers in the same field at conferences. For study selection, we first screened abstracts and subsequently assessed full texts for eligibility based on our predefined criteria. The study selection process is depicted in a flowchart in **Supplementary Figure S5**.

Data extraction

From each article, we extracted the following data: (a) the first author, (b) the year of publication, (c) the sample size (separated by groups if applicable), (d) BPD feature scores (if applicable), (e) sample characteristics, including whether the study assessed a BPD vs. healthy control (HC) group or a sample with BPD features, along with demographic information such as gender and age, (f) the specific Ultimatum Game outcome measured, (g) data required for effect size calculation, and (h) any additional conditions assessed. If certain information was not directly available in the article, we made up to two attempts to contact the corresponding authors and excluded articles if we didn't receive a response.

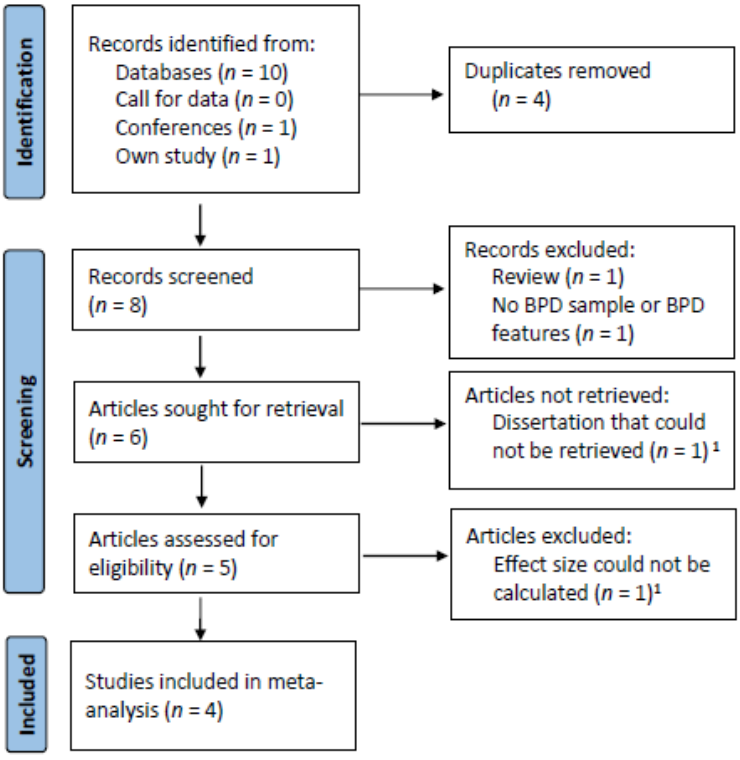
Meta-Analysis procedure

The meta-analysis was conducted in R (33) using the meta package (34), following the guidelines outlined by Harrer et al. (35). For each study, between-group effect sizes (standardized mean differences) were calculated as Cohen's d. For the study reporting BPD feature scores, participants were divided into groups based on established cut-off criteria (36). If studies reported outcomes separately

for additional conditions, the data were aggregated across these conditions before calculating the effect sizes.

We used a random-effects model because we expected heterogeneity between studies. To calculate the 95% confidence interval for the pooled effect size, we used the Knapp-Hartung adjustments (37), which are recommended for meta-analyses with a small number of studies. Additionally, alongside the heterogeneity statistic I^2, which can be biased in smaller meta-analyses, we employed the Paule-Mandel estimator (38) to assess heterogeneity variance (τ^2). For the targeted outcome—rejections in the Ultimatum Game—all available data were pooled, regardless of whether rejections were the primary or secondary outcome in the studies. To manage potential outliers, we examined if any of the studies´ 95% confidence intervals fell outside the pooled effect size. None of the studies met this criterion, indicating no outliers. Publication bias was evaluated using a funnel plot.

The systematic search initially identified 12 studies. After removing duplicates and excluding records based on abstract screening, five studies were selected for full-text analysis. One study was subsequently excluded due to insufficient data for effect size calculation, as no response was received from the authors. Ultimately, four studies met the inclusion criteria and were included in the final meta-analysis. Detailed characteristics of these studies are presented in **Supplementary Table S3.**



Supplementary Figure S5. PRISMA Flowchart of the Study Selection Process. [1]Authors were contacted and asked for the data needed. For the excluded articles no response was received.

Supplementary Table S3. Characteristics of the studies included in the Meta-Analysis

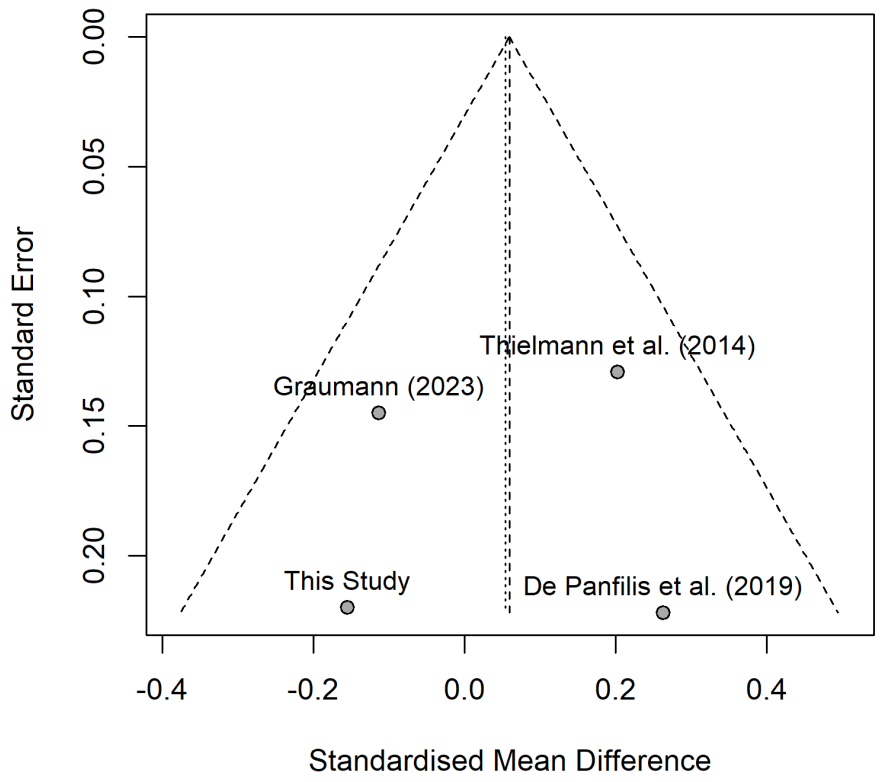
Study	Sample Type	BPD Measure	*N HC*	*N BPD*	Female (%)	Age (M)	Outcome	Additional Conditions
De Panfilis et al. (39)	HC, BPD	SCID-II	41	41	76.8	37	Percentage of offers rejected[1]	Emotion regulation strategies
Graumann et al. (40)	HC, BPD	SCID	97	94	100	28	Percentage of offers rejected[1]	Overinclusion and exclusion
Thielmann et al. (41)	Community sample (BPD features)[2]	VEI-BOR	94	168	82[3]	31.2[3]	Offers rejected after breakdown	Cooperation breakdown severity
This Study	HC, BPD	IPDE	49	36	100	25.8[3]	Percentage of offers rejected	none

Note. HC = Healthy Control group; BPD = Borderline Personality Disorder group. BPD Group Measure = measure of BPD symptoms or features. SCID/SCID-II = *Structured Clinical Interview* (BPD criterion: 5 out of 9 symptoms), VEI-BOR = *Verhaltens-Erlebens-Inventar* (Cut-off for BPD group inclusion: total score >38), IPDE = *International Personality Disorder Examination* (BPD criterion: 4 out of 9 symptoms). Outcome = outcome that was used for the meta-analysis. Additional conditions = conditions in the original study that varied within or between participants.

[1]Studies reported the outcome separately for additional conditions. For the meta-analysis, data was aggregated across conditions.

[2]No grouped sample, authors recruited participants via online communities, BPD-related web pages, and mailing lists and continuously assessed their BPD features.

[3]Studies reported descriptives for the whole sample, but only a sub-sample was included in the task for the target outcome. This was because of separate exclusion criteria or because only a sub-sample got assigned the condition with the target outcome in designs with different tasks.



Supplementary Figure S6. Funnel Plot for Meta-Analysis. The effect size of each study is depicted on the x-axis and the respective standard error on the y-axis. The vertical line represents the average effect size. All studies are within the funnel. Studies with a lower standard error lie closer to each other, resulting in an overall symmetrical pattern not indicative of publication bias.

References

1. Margraf J, Cwik JC. Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen. 2017 [cited 2022 May 17]; Available from: https://omp.ub.rub.de/index.php/RUB/catalog/book/102

2. Loranger AW. The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration International Pilot Study of Personality Disorders. Arch Gen Psychiatry. 1994 Mar 1;51(3):215.

3. Bohus M, Kleindienst N, Limberger MF, Stieglitz RD, Domsalla M, Chapman AL, et al. The Short Version of the Borderline Symptom List (BSL-23): Development and Initial Data on Psychometric Properties. Psychopathology. 2009;42(1):32–9.

4. Franke GH. BSCL-53®-S. Brief symptom-checklist–standard–German manual. Göttingen: Hogrefe Publishing Group; 2017.

5. Beck AT, Steer RA, Brown GK. BDI-II. Beck-Depressionsinventar [Beck Depression Inventory II]. Frankfurt (Germany): Pearson Verlag; 2006.

6. Kessler RC, Adler L, Ames M, Demler O, Faraone S, Hiripi E, et al. The World Health Organization adult ADHD self-report scale (ASRS): a short screening scale for use in the general population. Psychol Med. 2005 Feb;35(2):245–56.

7. Rösler M, Retz W, Retz-Junginger P, Thome J, Supprian T, Nissen T, et al. Instrumente zur Diagnostik der Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS) im Erwachsenenalter: Selbstbeurteilungsskala (ADHS-SB) und Diagnosecheckliste (ADHS-DC). Nervenarzt [Internet]. 2004 Oct [cited 2022 May 17];75(9). Available from: http://link.springer.com/10.1007/s00115-003-1622-2

8. Retz-Junginger P, Retz W, Blocher D, Weijers HG, Trott GE, Wender PH, et al. Wender Utah Rating Scale (WURS-k) Die deutsche Kurzform zur retrospektiven Erfassung des hyperkinetischen Syndroms bei Erwachsenen: Die deutsche Kurzform zur retrospektiven Erfassung des hyperkinetischen Syndroms bei Erwachsenen. Nervenarzt. 2002 Sep;73(9):830–8.

9. Steyer R, Schwenkmezger P, Notz P, Eid M. The multidimensional mental state questionnaire: Manual. Göttingen: Hogrefe; 1997.

10. Borkenau P, Ostendorf F. NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (NEO-FFI; 2. neu normierte und vollständig überarb. Aufl.). Göttingen: Hogrefe; 2007.

11. Bach B, Kerber A, Aluja A, Bastiaens T, Keeley JW, Claes L, et al. International Assessment of DSM-5 and ICD-11 Personality Disorder Traits: Toward a Common Nosology in DSM-5.1. Psychopathology. 2020;53(3–4):179–88.

12. Zimmermann J, Altenstein D, Krieger T, Holtforth MG, Pretsch J, Alexopoulos J, et al. The Structure and Correlates of Self-Reported DSM-5 Maladaptive Personality Traits: Findings From Two German-Speaking Samples. J Personal Disord. 2014 Aug;28(4):518–40.

13. Rohrmann S, Hodapp V, Schnell K, Tibubos AN, Schwenkmezger P, Spielberger CD. Das State-Trait-Ärgerausdrucks-Inventar-2 (STAXI-2). Deutschsprachige Adaptation des State-Trait Anger Expression Inventory-2 (STAXI-2) von Charles D. Spielberger [State-Trait Anger-Expression-Inventory-2 (STAXI-2). German Adaptation of the State-Trait-Anger-Expression-Inventory-2 (STAXI-2) of Charles D. Spielberger]. Bern: Hans Huber; 2013.

14. Preuss UW, Rujescu D, Giegling I, Watzke S, Koller G, Zetzsche T, et al. Psychometrische Evaluation der deutschsprachigen Version der Barratt-Impulsiveness-Skala. Nervenarzt. 2008 Mar;79(3):305–19.

15. Hofmann SG, Kashdan TB. The Affective Style Questionnaire: Development and Psychometric Properties. J Psychopathol Behav Assess. 2010 Jun;32(2):255–63.

16. Pruessner L, Holt DV, Gölz R, Sevcenko N, Hofmann SG, Backenstrass M. Psychometrische Eigenschaften der deutschsprachigen Version des Interpersonal Emotion Regulation Questionnaire. Diagnostica. 2020 Jan;66(1):62–73.

17. Baudson TG, Preckel F. mini-q: Intelligenzscreening in drei Minuten. Diagnostica. 2016 Jul;62(3):182–97.

18. Damovsky F, Zettl M, Zimmermann J, Herbold W, Curtius T, Buecker S, et al. Das Persönlichkeitsinventar für ICD-11: Reliabilität und Validität der deutschen Version in einer klinischen und nicht-klinischen Stichprobe [Internet]. Open Science Framework; 2021 Sep [cited 2022 May 17]. Available from: https://osf.io/cm3qv

19. Oltmanns JR, Widiger TA. A self-report measure for the ICD-11 dimensional trait model proposal: The Personality Inventory for ICD-11. Psychol Assess. 2018 Feb;30(2):154–69.

20. Spitzer C, Müller S, Kerber A, Hutsebaut J, Brähler E, Zimmermann J. Die deutsche Version der Level of Personality Functioning Scale-Brief Form 2.0 (LPFS-BF): Faktorenstruktur, konvergente Validität und Normwerte in der Allgemeinbevölkerung. PPmP - Psychother · Psychosom · Med Psychol. 2021 Jul;71(07):284–93.

21. Weekers LC, Hutsebaut J, Kamphuis JH. The Level of Personality Functioning Scale-Brief Form 2.0: Update of a brief instrument for assessing level of personality functioning: The Level of Personality Functioning Scale - Brief Form 2.0. Personal Ment Health. 2019 Feb;13(1):3–14.

22. Kröger C, Kosfelder J. IES-27: Skala zur Erfassung der Impulsivität und emotionalen Dysregulation der Borderline-Persönlichkeitsstörung. Göttingen: Hogrefe; 2011.

23. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007 May;39(2):175–91.

24. Korn CW, La Rosée L, Heekeren HR, Roepke S. Social feedback processing in borderline personality disorder. Psychol Med. 2016 Feb;46(3):575–87.

25. Leiner DJ. SoSci Survey (Version 3.1.06) [Internet]. 2019. Available from: https://www.soscisurvey.de

26. Murphy RO, Ackermann KA, Handgraaf M. Measuring Social Value Orientation. SSRN Electron J [Internet]. 2011 [cited 2019 Jul 23]; Available from: http://www.ssrn.com/abstract=1804189

27. Doppelhofer LM, Hurlemann R, Bach DR, Korn CW. Social motives in a patient with bilateral selective amygdala lesions: Shift in prosocial motivation but not in social value orientation. Neuropsychologia. 2021 Nov;162:108016.

28. Fehr E, Schmidt KM. A Theory of Fairness, Competition, and Cooperation. Q J Econ. 1999 Aug 1;114(3):817–68.

29. Bolton GE, Ockenfels A. ERC: A Theory of Equity, Reciprocity, and Competition. Am Econ Rev. 2000 Mar;90(1):166–93.

30. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978;6(2):461–4.

31. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. NeuroImage. 2009 Jul;46(4):1004–17.

32. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies — Revisited. NeuroImage. 2014 Jan;84:971–85.

33. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/

34. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R [Internet]. Cham: Springer International Publishing; 2015 [cited 2024 Oct 8]. (Use R!). Available from: https://link.springer.com/10.1007/978-3-319-21416-0

35. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-On Guide [Internet]. 1st ed. Boca Raton: Chapman and Hall/CRC; 2021 [cited 2024 Oct 8]. Available from: https://www.taylorfrancis.com/books/9781003107347

36. Hepp J, Hilbig BE, Kieslich PJ, Herzog J, Lis S, Schmahl C, et al. Borderline Personality and the Detection of Angry Faces. Felmingham K, editor. PLOS ONE. 2016 Mar 31;11(3):e0152947.

37. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003 Sep 15;22(17):2693–710.

38. Paule RC, Mandel J. Consensus Values and Weighting Factors. J Res Natl Bur Stand. 1982 Sep;87(5):377.

39. De Panfilis C, Schito G, Generali I, Gozzi LA, Ossola P, Marchesi C, et al. Emotions at the border: Increased punishment behavior during fair interpersonal exchanges in borderline personality disorder. J Abnorm Psychol. 2019 Feb;128(2):162–72.

40. Graumann L, Kulakova E, Cho AB, Deuter CE, Wolf OT, Schell J, et al. Elevated testosterone and prosocial behavior in female patients with borderline personality disorder independent of social exclusion. Psychoneuroendocrinology. 2025 Jan;171:107232.

41. Thielmann I, Hilbig BE, Niedtfeld I. Willing to Give but Not to Forgive: Borderline Personality Features and Cooperative Behavior. J Personal Disord. 2014 Dec;28(6):778–95.

Study IV – Supplementary material

Supplementary material

Supplementary methods

Economic models for the JPE task.

(1) Individualism model: $U = \alpha * \$Self$

In the individualism model, utility (U) depends linearly on the payoff for self ($Self).

(2) Prosocial model: $U = \alpha * \$Self + \beta * \$Other$

In the prosocial model (McClintock, 1972), utility (U) is a weighted sum of the linear payoffs for self ($Self) and other ($Other).

(3) Altruism model: $U = \beta * \$Other$

In the altruism model, utility (U) solely depends on the payoff for other ($Other).

(4) Joint-gain model: $U = \alpha * (\$Self + \$Other)$

The joint gain-model considers the total welfare of self and other but does not distinguish between payoffs for self and other.

(5) Inequality model: $U = \alpha * abs(\$Self - \$Other)$

The inequality model considers the absolute difference between the payoffs for self and other.

(6) Inequality & Joint-gain model:

$$U = \alpha * abs(\$Self - \$Other) + \beta * (\$Self + \$Other)$$

The inequality & Joint-gain model is a linear combination of the inequality model and the joint-gain model.

(7) Fehr-Schmidt model:

$$U = \alpha * \$Self + \beta * \max(\$Other - \$Self, 0) + \gamma * \max(\$Self - \$Other, 0)$$

The Fehr-Schmidt (Fehr & Schmidt, 1999) model separately considers the payoff for the self, disadvantageous inequality (self gets less than other), and advantageous inequality (other gets less than self). Here, beta represents disadvantageous inequality aversion and gamma represents advantageous inequality aversion. If the difference between $Other and $Self is a negative value, the term *max($Other - $Self,0)* sets the result to zero (and keeps the result if it is a positive value). The same holds for the difference between $Self and $Other ($Self - $Other). Both, disadvantageous and advantageous inequality aversion are supposed to affect the utility negatively. Parameters were not restricted to enable the possibility of altruistic and competitive motives. This model is equivalent to the model used by Haruno and Frith (2010) but has a different parameterization (i.e., this parameterization distinguishes

between payoffs to self and other but collapses advantageous and disadvantageous inequality aversion into absolute inequality aversion; see Appendix in the Supplementary Material). The Fehr-Schmidt model is also equivalent to a formulation that includes a regressor for joint-gain (i.e., $\$Self + \$Other$) instead of a regressor for pure self-gain (i.e., $\$Self$).

(8) ERC (Equity, Reciprocity, Competition) model:

$$U = \alpha * \$Self + \frac{\beta}{2} * \left(\frac{\$Self}{\$Self + \$Other} - \frac{1}{2}\right)^2$$

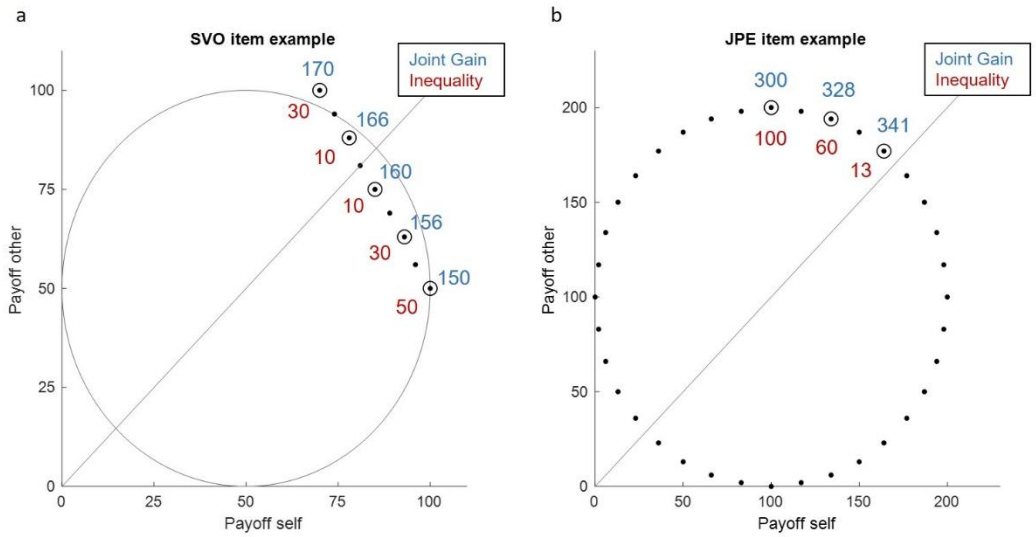
The ERC model (Bolton & Ockenfels, 2000) considers (quadratic and normalized) inequality (but does not explicitly distinguish between advantageous and disadvantageous inequality).

Model fitting and model selection. For model fitting, we excluded participants who showed no variance in the JPE task at all (i.e., they selected the same evaluation option on each trial; see Supplementary Table 1). Regression models were fitted using the MATLAB function *regress*, which minimizes the sum of squared errors (also known as residual sum of squares, RSS). Parameter estimates were unbounded.

For model selection, we used the Bayesian information criterion (BIC). BIC depends on model fit (calculated as RSS) and includes a penalty for model complexity (that depends on the number of parameters, k), according to the following standard formula (Burnham & Anderson, 2004; Hurvich & Tsai, 1990), where n is the number of trials per participant:

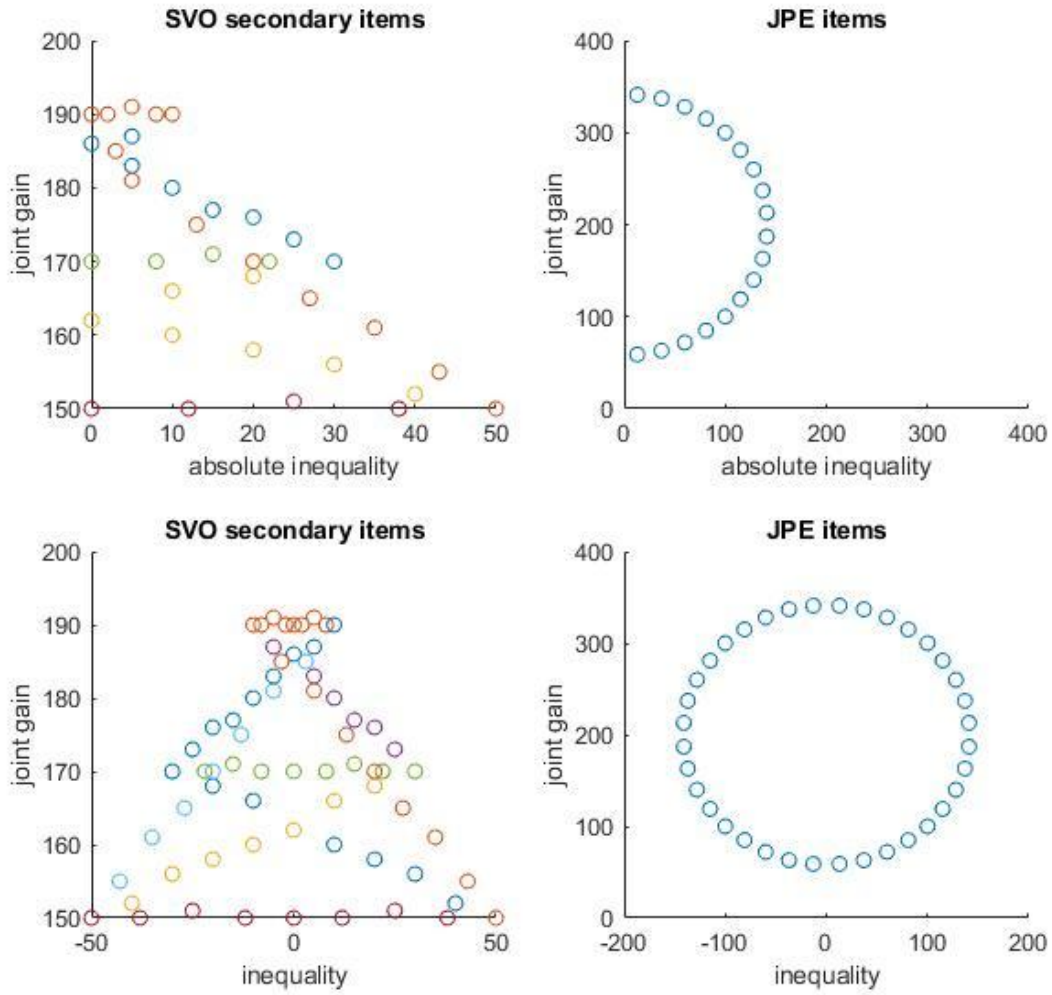
$$BIC = n * \ln\left(\frac{RSS}{n}\right) + k * \ln(n)$$

First, we performed fixed-effects analyses, under the assumption that every participant uses the same model: BIC values are summed over participants to obtain log-group Bayes factors. The model with the lowest log-group Bayes factors is preferred. Second, we performed random-effects analyses under the assumption that different participants may use different models. We used the function *spm_BMS* in the MATLAB toolbox SPM (https://www.fil.ion.ucl.ac.uk/spm/) to calculate protected exceedance probabilities, which measure how likely it is that any given model is more frequent than all other models in the population.



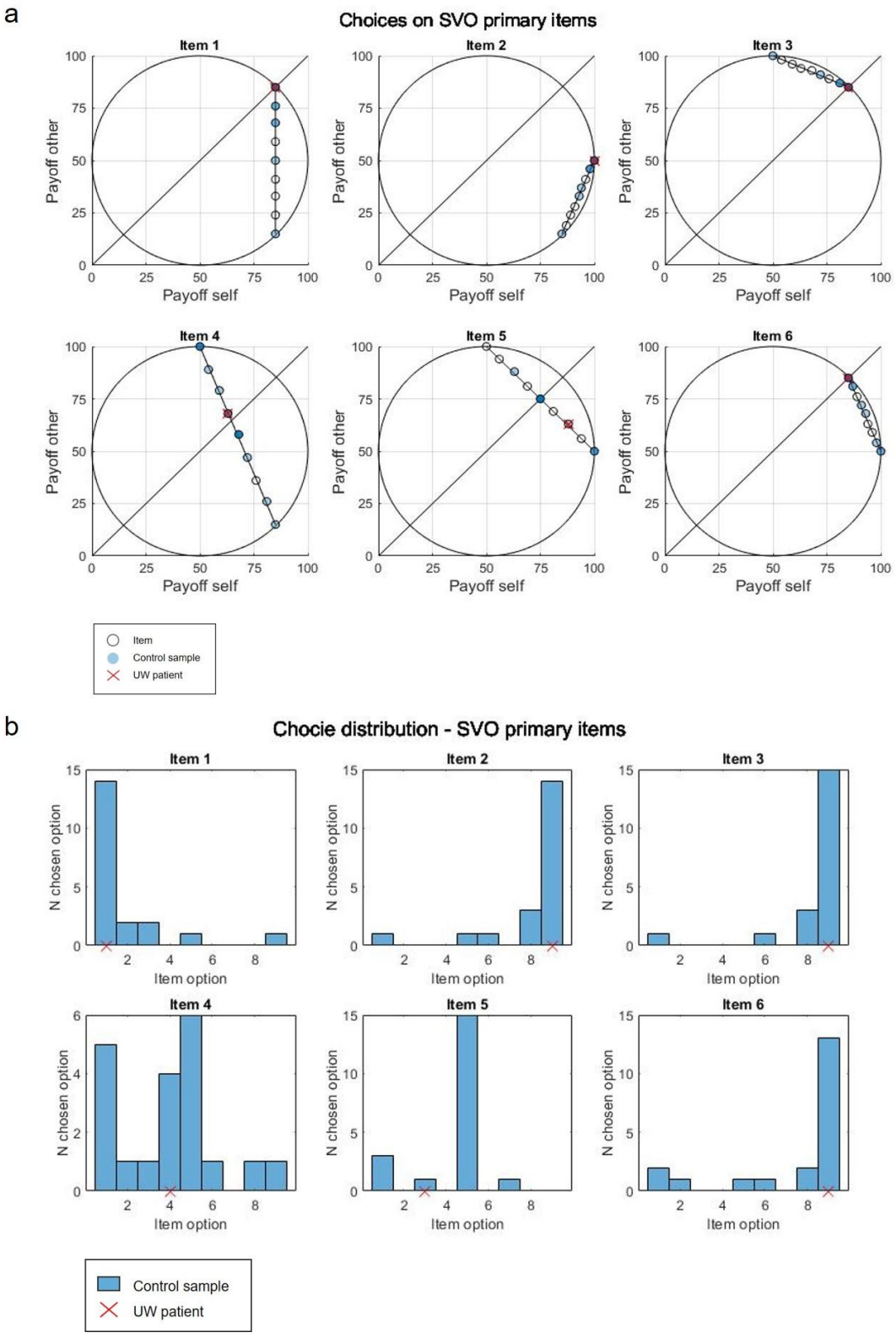
Supplementary Figure 1. Example items of the SVO task and the JPE task.

a. Example item of one of the secondary items of the SVO task and corresponding joint gain and inequality values. Black dots visualize the 9 options of the item. Circles around the dots indicate the example option corresponding to the numbers. **b.** Example items of the JPE task and corresponding joint gain and inequality values. Black dots display all 36 items of the JPE task (see Supplementary Table 2 for a list of these 36 items). Circles around the dots indicate the example item corresponding to the numbers **a. & b.** Blue values above the dots indicate the joint gain, red values below the dots indicate the inequality between payoffs for self and other for the corresponding option/item. Secondary items of the SVO comprise different combinations (higher/lower) of joint gain and inequality. See Figure 1 for a visualization of all SVO items. See Supplementary Figure 2 for the depiction of the SVO and JPE items in terms of (absolute) inequality and joint gain. See Supplementary Figures 3 & 4 for choice distributions across the SVO items.



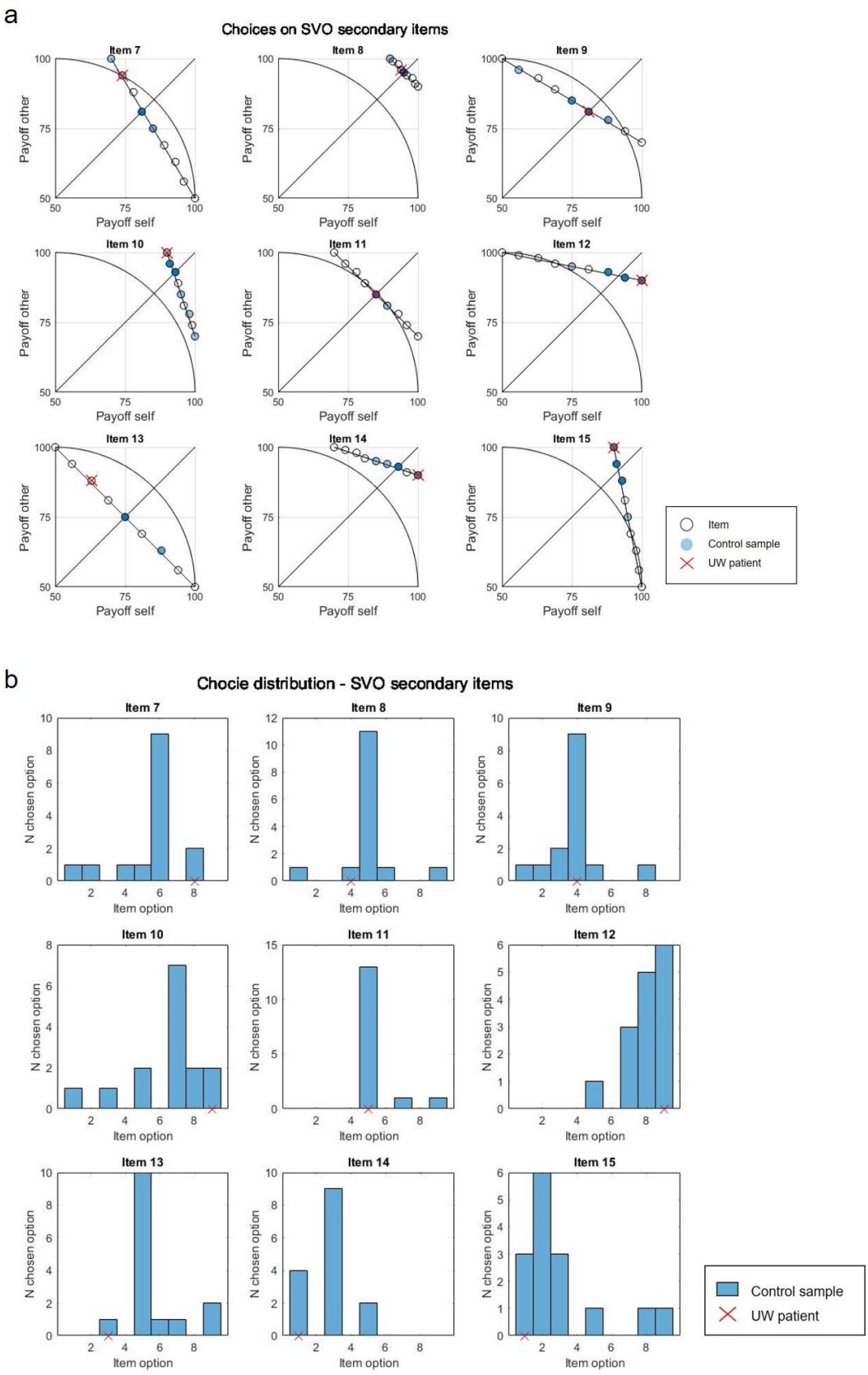
Supplementary Figure 2. Items of the SVO task and the JPE task in terms of (absolute) inequality and joint gain.

Left column: The nine options of each of the nine SVO secondary items are depicted. Right column: The 36 items of the JPE task are depicted. These 36 items are positioned on a circle in the coordinate system with payoff for self and payoff for other as axes (see Supplementary Figure 1). In the coordinate system of Supplementary Figure 1, the diagonal identity line (with a slope of 1 and an intercept of 0, i.e., at an angle of 45°) corresponds to inequality of 0 between payoffs for self and other (see also Figure 1c). The orthogonal diagonal line (with a slope of -1 and an intercept of 0, i.e., at an angle of 135°) corresponds to joint gain (see also Figure 1d). That is, the axes for the JPE items can be rotated by an angle of 45° to switch from axes in terms of self and other payoffs (Supplementary Figure 1) to axes in terms of inequality and joint gain.



Supplementary Figure 3. Choice distributions on the SVO primary items.

a. All nine options for the six primary items are depicted. The numbers of participants in the control sample who chose the respective options are color-coded. b. For better visualization, these numbers are shown as histograms.



Supplementary Figure 4. Choice distributions on the SVO secondary items.

Same logic as in Supplementary Figure 3 for the SVO secondary items.

Supplementary Table 1. Sample sizes according to the used exclusion criteria for different analyses.

Criteria	Descriptions of exclusion or inclusion criteria	Control	Online subgroup	Online sample	Student
(1) Tested participants	No participants excluded	23 (100)	97 (100)	628 (100)	41 (100)
(2) Transitive in SVO task primary items	Participants with intransitive choices in the primary items of SVO task excluded; this exclusion criterion was used for all analyses	20 (86.96)	91 (93.81)	603 (96.02)	40 (97.56)
(3) Not deterministic in JPE task	Participants with a choice variance of 0 in the JPE task were excluded because they made deterministic choices; this exclusion criterion was used for modeling the JPE data	19 (82.61)	86 (88.66)	586 (93.31)	40 (97.56)
(4) Prosocials in SVO primary items	Participants with a prosocial SVO scores according to primary items of SVO task; this inclusion criterion was used for calculating the proportion of prosocials in Table 1	16 (69.57)	74 (76.29)	474 (75.48)	32 (78.05)
(5) Prosocials in SVO primary and secondary items	Participants with prosocial SVO scores according to both the primary and the secondary items of the SVO task; this inclusion criterion was used for all analyses and plots that describe prosocial motivation; the numbers correspond to the proportion of prosocials in Table 2	15 (65.22)	66 (68.04)	432 (68.79)	26 (63.41)

Participants are given as N (% of tested participants)

Supplementary Table 2. Items of the JPE task

Item number	Payoff self	Payoff other
1	36	177
2	23	164
3	13	150
4	6	134
5	2	117
6	0	100
7	66	6
8	83	2
9	100	0
10	117	2
11	134	6
12	150	13
13	198	117
14	194	134
15	187	150
16	177	164
17	164	177
18	150	187
19	2	83
20	6	66
21	13	50
22	23	36
23	36	23
24	50	13

25	134	194
26	117	198
27	100	200
28	83	198
29	66	194
30	50	187
31	164	23
32	177	36
33	187	50
34	194	66
35	198	83
36	200	100

Supplementary Table 4. Model comparisons in the JPE task according to random-effects analyses

Model	Control sample	Online subgroup	Online sample	Student sample
Individualism	0	0	0	0
Prosocial	0.003	0	0	0
Altruism	0	0	0	0
Joint-gain	0	0	0	0
Inequality	0	0	0	0
Inequality & Joint-gain	0.984	0.931	1	0.806
Fehr-Schmidt	0.005	0.069	0	0.072
Equity, Reciprocity, Competition (ERC)	0.007	0	0	0.122

Higher protected exceedance probabilities indicate better model fit according to random-effects analyses, which assume that different participants may use different models. Protected exceedance probabilities are rounded to three decimal places.

Supplementary Table 5. Model comparisons in the JPE task according to number of participants for which the model performed best in terms of BIC

Model	UW patient	Control sample	Online subgroup	Online sample	Student sample
Individualism	0	0	7	51	0
Prosocial	0	2	10	81	2
Altruism	0	0	2	12	0
Joint-gain	0	0	2	18	0
Inequality	0	1	2	22	2
Inequality & Joint-gain	1	11	34	238	15
Fehr-Schmidt	0	2	21	109	9
Equity, Reciprocity, Competition (ERC)	0	3	8	55	12
Total N	1	19	86	586	40

Appendix

Here, we show that the Fehr-Schmidt model is equivalent to the model used by Haruno and Frith (2010) but with a different parameterization. Two regression models are equivalent if a linear combination of the regressors of the first model results in the regressors of the second model. Below, we first list the two models again:

Fehr-Schmidt model:

(1) $U_{FS} = \alpha_{FS} * \$Self + \beta_{FS} * \max(\$Other - \$Self, 0) + \gamma_{FS} * \max(\$Self - \$Other, 0)$

Haruno-Frith model:

(2) $U_{HF} = \alpha_{HF} * \$Self + \beta_{HF} * \$Other + \gamma_{HF} * \text{abs}(\$Self - \$Other)$

Now, we want to show that $U_{FS} = U_{HF}$ for any scalars $\$Self$ and $\$Other$. That is, we want to show that the parameters of the Fehr-Schmidt model can be expressed by the parameters of the Haruno-Frith model (without any dependence on the regressors). Because the models contain the "max" and the "abs" operators, we treat $\$Self \geq \$Other$ and $\$Self < \$Other$ separately.

For $\$Self \geq \$Other$, the Fehr-Schmidt model (1) is

$U_{FS} = \alpha_{FS} * \$Self + \beta_{FS} * 0 + \gamma_{FS} * (\$Self - \$Other)$

For $\$Self \geq \$Other$, the interior of the first "max" operator equates to zero. Rearranging gives

(3) $U_{FS} = (\alpha_{FS} + \gamma_{FS}) * \$Self - \gamma_{FS} * \$Other$

For $\$Self \geq \$Other$, the Haruno-Frith model (2) is

$U_{HF} = \alpha_{HF} * \$Self + \beta_{HF} * \$Other + \gamma_{HF} * (\$Self - \$Other)$

That is, the "abs" operator can be omitted (because $\$Self - \$Other \geq 0$). Rearranging gives

(4) $U_{HF} = (\alpha_{HF} + \gamma_{HF}) * \$Self + (\beta_{HF} - \gamma_{HF}) * \$Other$

Thus, the two models can be reduced to expressions that only contain $\$Self$ and $\$Other$ as regressors. Fitting a regression with these two regressors to the same data will result in identical parameter estimates regardless of the notation used in (3) or (4). Therefore, we can equate the parameter estimate for $\$Self$ in the two formulations (3) and (4).

(5) $\alpha_{FS} + \gamma_{FS} = \alpha_{HF} + \gamma_{HF}$

Analogously, we can equate the parameter estimate for $\$Other$ in the two formulations (3) and (4).

(6) $-\gamma_{FS} = \beta_{HF} - \gamma_{HF}$

Solving for the parameters of the Fehr-Schmidt model in (5) and (6) yields

$\alpha_{FS} = \alpha_{HF} + \beta_{HF}$ and

$\gamma_{FS} = -\beta_{HF} + \gamma_{HF}$

Now, we consider the other case in the same way. For $\$Self < \$Other$, the Fehr-Schmidt model (1) is

$U_{FS} = \alpha_{FS} * \$Self + \beta_{FS} * (\$Other - \$Self) + \gamma_{FS} * 0$

For $\$Self < \$Other$, the interior of the second "max" operator equates to zero. Rearranging gives

(7) $U_{FS} = (\alpha_{FS} - \beta_{FS}) * \$Self + \beta_{FS} * \$Other$

For $\$Self < \$Other$, the Haruno-Frith model is

$U_{HF} = \alpha_{HF} * \$Self + \beta_{HF} * \$Other + \gamma_{HF} * (\$Other - \$Self)$

For $\$Self < \$Other$, the interior of the "abs" operator has to be multiplied by -1 (because $\$Self - \$Other < 0$). Rearranging gives

(8) $U_{HF} = (\alpha_{HF} - \gamma_{HF}) * \$Self + (\beta_{HF} + \gamma_{HF}) * \$Other$

Again, the two models can be reduced to expressions that only contain $\$Self$ and $\$Other$ as regressors. We equate the parameter estimate for $\$Self$ in the two formulations (7) and (8).

(9) $\alpha_{FS} - \beta_{FS} = \alpha_{HF} - \gamma_{HF}$

We equate the parameter estimate for $\$Other$ in (7) and (8).

(10) $\beta_{FS} = \beta_{HF} + \gamma_{HF}$

Solving for the parameters of the Fehr-Schmidt model in (9) and (10) yields

$\alpha_{FS} = \alpha_{HF} + \beta_{HF}$ and

$\beta_{FS} = \beta_{HF} + \gamma_{HF}$

Thus, we can express the parameters of the Fehr-Schmidt model solely in terms of the Haruno-Frith model. For the two cases, we obtain the same equivalence for α_{FS} and unique equivalences for β_{FS} and γ_{FS}. The two models are equivalent.

An analogous derivation can be made for a model that includes a regressor for joint-gain (i.e. , $\$Self + \$Other$) instead of a regressor for pure self-gain (i.e., $\$Self$).

References

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Hurvich, C. M., & Tsai, C. (1990). The Impact of Model Selection on Inference in Linear Regression. *The American Statistician*, *44*(3), 214–217. https://doi.org/10.1080/00031305.1990.10475722

McClintock, C. G. (1972). Social motivation—A set of propositions. *Behavioral Science*, *17*(5), 438–454. https://doi.org/10.1002/bs.3830170505