Commentary

Biological Psychiatry: CNNI

Don't Ever Leave Me, You Disgusting Monster: Computational Insights Into Moral Inference Updating in Borderline Personality Disorder

Joshua W. Buckholtz

Borderline personality disorder (BPD) is a serious mental illness characterized by volatility in mood, social attachments, and self-concept (1). A core feature of BPD is instability in interpersonal relationships. Individuals with BPD have intense and chaotic attachments to others, characterized by an often cyclical pattern of idealization followed by devaluation ("splitting"). The shift between these two extremes is often abrupt and seemingly out of proportion to the eliciting event. Splitting is often accompanied by intensely dysphoric emotional states, which in turn drive highly impulsive and typically maladaptive behaviors (e.g., self-injury, substance abuse, and reckless spending) (1). Interpersonal disturbances are responsible for much of the distress and impairment and are key targets for psychotherapeutic interventions.

Yet despite the severe subjective distress, functional impairment, and economic burden imposed by interpersonal symptoms in BPD (2), their underlying cognitive and neurobiological mechanisms are only beginning to be identified. In the current issue of *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, Siegel *et al.* (3) make a significant contribution to our understanding of these mechanisms by using an innovative computational cognition approach to understanding how patients with BPD generate and update moral inferences about other agents.

There is a wealth of evidence showing that patients with BPD exhibit deficits in multiple domains of impulse control (4–6) and show heightened responses to stress and threat along with a reduced capacity for effective emotion regulation (7–9). This work accords well with the clinical picture of BPD—indeed, the diagnostic criteria for BPD emphasize affectively driven disinhibition as a core feature. The relationship between such deficits and social and interpersonal symptoms, however, was less clear. The use of economic decision-making paradigms, as in seminal work by King-Casas *et al.* (10), shed new important light on the mechanisms underlying social and interpersonal deficits by quantifying alterations in trust-related behavior during dyadic interactions. This line of research, which showed enhanced sensitivity to trust violations coupled with a blunted response to behavioral signals indicating an effort to repair trust, provided a basis for quantifying social dysfunction in BPD. Yet, as noted above, the dynamic of idealization and devaluation—central to the nomological net of BPD and a critical driver of maladaptive behavior in the disorder—remained relatively understudied. Siegel *et al.* (3) make a critical contribution to our understanding of this feature of BPD.

The authors used an ingenious paradigm that permits an objective quantification of moral inference: how an individual uses information about the observed actions of an agent to judge their moral character. Adaptive moral inference requires the veridical assessment of social signals. However, such signals are often nonstationary; as an agent's behavior changes over time, observers must flexibly incorporate new signals into their model of that agent's moral character. Agents that appear morally "good" or trustworthy can reveal themselves over time to be otherwise, and agents that initially present as morally "bad" or untrustworthy can redeem themselves by altering their choice preferences. The task used by Siegel *et al.* (3) operationalizes this by having participants observe the behavior of two agents, each of whom is faced with a choice, on each trial, to inflict a painful electric shock on a third party in exchange for varying sums of money. The two agents differed in their moral preferences: the good agent only chose to shock when offered large sums of money, while the bad agent was willing to shock at much lower rates of compensation. The authors use a Bayesian modeling framework to describe how participants integrate prior beliefs about an agent with new information about that agent's behavior. Less certain priors are associated with higher learning rates, such that new information about an agent's behavior is weighted more heavily. In previous work, the authors found evidence of greater subjective uncertainty for bad compared with good agents; this subjective evaluation tracked with a computationally derived measure of belief volatility for morally bad agents. It has been proposed that this enhanced belief volatility for agents that are initially assessed as morally bad permits more flexibility in belief updating, which could in turn facilitate forgiveness when interpersonal trust is compromised.

Strikingly, Siegel *et al.* (3) report disrupted belief updating for bad versus good agents in patients with BPD. Specifically, BPD was associated with slower updating of beliefs about the bad agent and faster updating of beliefs about the good agent. In other words, individuals with BPD held more certain beliefs about agents who were initially assessed as morally untrustworthy; these beliefs were more rigid, in that they were less influenced by otherwise trust-enhancing signals from that agent. Conversely, such participants were less certain about the moral character of trustworthy agents and were more amenable to updating their beliefs about such agents. Greater symptom severity was linked to less flexible updating in patients with BPD. Notably, participants with BPD held more pessimistic (though equally certain) expectations compared

SEE CORRESPONDING ARTICLE ON PAGE 1134

https://doi.org/10.1016/j.bpsc.2020.10.012

ISSN: 2451-9022 Biological Psychiatry: Cognitive Neuroscience and Neuroimaging December 2020; 5:1075–1076 www.sobp.org/BPCNNI

with individuals without BPD when assessed before performing the experimental task. Together, these findings suggest that previous expectations about moral character may more strongly influence behavior in individuals with BPD, who appear relatively impaired when using new information about an agent's behavior to flexibly update moral inference in order to override prior beliefs.

A particular strength of this study is the authors' demonstration that this ostensible computational phenotype for inflexible moral inference is sensitive to intervention. Using a sample of patients undergoing a psychosocial treatment for BPD (in this case, democratic therapeutic community treatment), treatment was associated with more uncertain beliefs and faster belief updating for bad agents. This is especially striking given that patients with BPD who were treated and patients with BPD who were untreated showed similar moral expectations; treatment appeared to selectively enhance the ability to flexibly incorporate new information about agents who were initially assessed as morally bad. On the whole, this work shows that a key element of human cooperation—the ability to flexibly adapt one's perception of the moral character of another when faced with changing behavioral signals—is compromised in BPD. This rigidity may explain, in part, why individuals with BPD are so prone to rupture interpersonal relationships based on small, ultimately nonpredictive perceived transgressions and why they have difficulty repairing these ruptures once they have occurred.

A few issues merit consideration. First, patients with BPD often enter a cyclical pattern of idealization and devaluation. While the present data provide a computational cognitive mechanism for devaluation, it remains unclear why such patients may eventually come to idealize an individual they have abruptly determined is morally untrustworthy. This may pertain to the structure of the task: while in this study participants were asked to make and update moral inferences based on an agent's actions toward a third party, the behavior of individuals with BPD is typically motivated by actions directed at them (i.e., second-party moral inference). It is possible that the differences in the mechanisms underlying moral inference updating differ in third-party versus second-party contexts, especially if that second party is an individual to whom the patient with BPD has developed (or desires to develop) an attachment. Second, whereas the current work uses a between-subjects design to assess the sensitivity of these computational phenotypes to intervention, it is possible that selection biases could account for at least some of the observed effects. Future work using a within-subject design to assess change over time as a function of treatment would provide exceptionally compelling evidence for treatment sensitivity. Further, such work—especially if performed in a sample undergoing a standardized treatment, such as dialectical behavior therapy—could selectively map the treatment domains and particular skills that are most effective for helping patients more adaptively use social signals to guide moral inference and choice behavior. Answering these questions will help refine our understanding of the important, novel pathomechanism for social dysfunction in BPD discovered by Siegel *et al.* (3).

Acknowledgments and Disclosures

The author reports no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Psychology and Center for Brain Science, Harvard University, Cambridge, and the Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts.

Address correspondence to Joshua W. Buckholtz, Ph.D., at joshuabuckholtz@fas.harvard.edu.

Received Oct 26, 2020; accepted Oct 27, 2020.

References

1. Gunderson JG, Herpertz SC, Skodol AE, Torgersen S, Zanarini MC (2018): Borderline personality disorder. Nat Rev Dis Primers 4:18029.
2. Zanarini MC, Frankenburg FR (2007): The essential nature of borderline psychopathology. J Pers Disord 21:518–535.
3. Siegel JZ, Curwell-Parry O, Pearce S, Saunders KEA, Crockett MJ (2020): A computational phenotype of disrupted moral inference in borderline personality disorder. Biol Psychiatry Cogn Neurosci Neuroimaging 5:1134–1141.
4. Nigg JT, Silk KR, Stavro G, Miller T (2005): Disinhibition and borderline personality disorder. Dev Psychopathol 17:1129–1149.
5. Linhartova P, Latalova A, Bartecek R, Sirucek J, Theiner P, Ejova A, *et al.* (2020): Impulsivity in patients with borderline personality disorder: A comprehensive profile compared with healthy people and patients with ADHD. Psychol Med 50:1829–1838.
6. Martino F, Gammino L, Sanza M, Berardi D, Pacetti M, Sanniti A, *et al.* (2020): Impulsiveness and emotional dysregulation as stable features in borderline personality disorder outpatients over time. J Nerv Ment Dis 208:715–720.
7. Daros AR, Williams GE (2019): A meta-analysis and systematic review of emotion-regulation strategies in borderline personality disorder. Harv Rev Psychiatry 27:217–232.
8. Haliczer LA, Woods SE, Dixon-Gordon KL (2020): Emotion regulation difficulties and interpersonal conflict in borderline personality disorder [published online ahead of print Jul 16]. Personal Disord.
9. Southward MW, Cheavens JS (2020): Quality or quantity? A multistudy analysis of emotion regulation skills deficits associated with borderline personality disorder. Personal Disord 11:24–35.
10. King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR (2008): The rupture and repair of cooperation in borderline personality disorder. Science 321:806–810.