Interpersonal dysfunction in borderline personality: a decision neuroscience perspective

Michael N. Hallquist¹,

Nathan T. Hall¹,

Alison M. Schreiber¹,

& Alexandre Y. Dombrovski²

¹Department of Psychology, Penn State University

²Department of Psychiatry, University of Pittsburgh

Corresponding author: Michael Hallquist, Department of Psychology, Penn State University, 140 Moore Building, University Park, PA 16802. Email: michael.hallquist@psu.edu.

This work was supported by the National Institutes of Health (R01 MH100095, R01 MH085651 to AYD; K01 MH090791 to MNH; R01 MH048463 to AYD and MNH).

Conflicts of interest: None.

Accepted for publication in Current Opinion in Psychology
Abstract

Borderline personality disorder (BPD) is characterized by disadvantageous decisions that are often expressed in close relationships and associated with intense negative emotions. Although functional neuroimaging studies of BPD have described regions associated with altered social cognition and emotion processing, these correlates do not inform an understanding of how brain activity leads to maladaptive choices. Drawing on recent research, we argue that formal models of decision-making are crucial to elaborating theories of BPD that bridge psychological constructs, behavior, and neural systems. We propose that maladaptive interactions between Pavlovian and instrumental influences play a crucial role in the expression of interpersonal problems. Finally, we articulate specific hypotheses about how clinical features of BPD may map onto neural systems that implement separable decision processes.
Introduction

Borderline personality disorder (BPD) is defined by emotion dysregulation, interpersonal dysfunction, and impulsivity [1]. Many problematic behaviors in BPD (e.g., suicide attempts) occur in response to negative interpersonal events such as perceived abandonment [2,3]. Clinical theory and experimental data indicate that individuals with BPD are interpersonally hypersensitive, often distorting social cues, forming extreme opinions of others, and making negative attributions about others’ actions and even facial expressions [4–6]. These biases are linked with alterations in brain function, including heightened amygdala responses to negative emotional stimuli [7,8] and greater dorsal anterior cingulate cortex (dACC; BA 24/32) and medial prefrontal cortex (mPFC; BA 10) activity to social rejection [9]. More generally, regions involved in emotion regulation and reward processing are affected, including subgenual anterior cingulate cortex (sgACC; BA 25), medial orbitofrontal cortex/ventromedial prefrontal cortex (mOFC/vmPFC; BA 10-14 [10]), and the amygdala [11].

This literature has informed the perspective that interpersonal hypersensitivity and emotion dysregulation in BPD reflect the predominance of ascending limbic influences over networks supporting social cognition and self-regulation [12]. However, the notion that emotions interfere with self-regulation and deliberation in BPD [13,14] provides few specific hypotheses about the neural underpinnings of salient features such as aggression [15] or rejection sensitivity [16]. In part, this reflects the absence of a testable theory linking BPD symptoms, underlying neurocognitive processes, and neural representation (see Box 1). Moreover, several other mental disorders (e.g., depression) prominently
feature abnormalities in frontolimbic circuits [17], further undermining the neurobiological specificity of this perspective.

We argue that these limitations can be overcome by developing formal models of decision-making [18] in BPD and testing them using the tools of modern cognitive neuroscience [19,20] (Box 2). We focus specifically on reinforcement learning (Box 3), which has a rich tradition of developing algorithms in silico [21] and validating them at the level of behavior and neural representation [22–24]. The past two decades of behavioral and computational neuroscience research have provided overwhelming evidence that adaptive and maladaptive decisions can be understood as reflecting the contribution of at least three computational systems: goal-directed, habitual, and Pavlovian (see Box 3) [18]. Furthermore, a coarse anatomical distinction between the primitive limbic system and sophisticated neocortical networks has given way to the perspective that cortical-striatal-limbic networks [25] carry decision-relevant signals from each of these systems [26]. We propose that the interpersonal hallmarks of BPD reflect: 1) an abnormally heightened motivation for interpersonal affiliation combined with a vulnerability to shifts in motivation; and 2) a predominance of Pavlovian (i.e., stimulus-outcome) influences on behavior that dominate goal-directed decision-making in stressful and emotionally arousing contexts.

**Disadvantageous Decisions in Borderline Personality**

Clinicians in training are often vexed by the extreme and seemingly irrational decisions that patients with BPD make. For example, why would a patient who talked at length about hating his girlfriend attempt suicide a week later when she broke up with him? Clinical theories of BPD have offered psychodynamic and cognitive-behavioral
explanations. For example, object relations theory frames contradictory behaviors in BPD in terms of split object representations in which attractive and aversive aspects of oneself and others remain unintegrated, with strong affect promoting transitions between extremes [27]. Conversely, dialectical behavior therapy emphasizes the functional role of extreme behaviors (e.g., rage), which may meet immediate interpersonal desires (e.g., emotional intimacy) at the expense of long-term relationships [28]. Several therapeutic approaches to BPD [27–29] involve improving metacognition (especially representation of others’ states of mind) and providing direct feedback to patients about maladaptive decisions [30].

Disadvantageous decisions in BPD often occur in the context of strong negative emotions, a facet of impulsivity called negative urgency [31]. Such decisions may reflect an emotion-dependent preference for immediate outcomes over delayed outcomes [32]. For example, an individual coping with marital conflict may assign greater subjective value to substance use than problem-solving to preserve the relationship. However, the notion that negative emotions promote short-sighted decisions was challenged by a recent meta-analysis of decision-making tasks [33]: individuals with BPD had an exaggerated preference for immediate rewards, but this was not enhanced by stress [34] or social exclusion [35]. Moreover, this preference was associated with a lack of premeditation, but not negative urgency. Individuals with BPD also tend to make poor decisions on the Iowa Gambling Task (IGT), a widely used paradigm that involves learning to choose safer card decks over high risk-high reward decks. This deficit may reflect insensitivity to losses in BPD [36], but this literature is inconsistent [37] and the IGT is likely a sensitive, not specific, measure of decision deficits in psychopathology [38].
What is a Formal Model of Decision-Making?

Although clinical theories provide a framework for understanding and remediating decision-making in BPD, they do not offer a mathematical account of how decision processes result in maladaptive choices. Likewise, empirical research to date has largely focused on summary statistics of decision quality (e.g., proportion of risky bets on a game) [33]. Over the past few decades, scientists have developed formal models of decision-making that integrate behavioral economics [39], psychology [40], and artificial intelligence [21]. The purpose of a formal model is to test falsifiable hypotheses about how decisions are made, typically by observing an evolving series of choices in experiments that approximate real-world challenges. Using a set of related equations, formal models specify what information is tracked by the individual and how it is integrated to choose an action in a specific situation.

Most formal models of decision-making can be understood within a Bayesian decision theory framework [41], which has five major components: 1) a specific state of the environment; 2) a set of actions that can be chosen in the current state; 3) subjective value estimates for each action quantifying its net costs and benefits given the current state; 4) a policy for selecting an optimal action given current value estimates; and 5) a procedure for updating value estimates in light of additional experience or shifts in the environment [18]. Crucially, subjective representations of the environment incorporate objective factors, internal factors, and prior beliefs and experience [42]. For example, imagine a man waiting at a restaurant for his partner to arrive (objective). Knowing that the partner often runs late (prior experience), he is prepared and starts reading a novel. After about 30 minutes, however, he begins to feel lonely (an internal state in which emotional intimacy has high
Decision neuroscience perspective on borderline personality

value). When the partner arrives, he immediately begins to complain about feeling unloved (greater response vigor in a 'deprived' state) and does not notice the partner's agitation. Only after calming down does he learn that the partner was in a minor car accident and had to exchange insurance information before driving to dinner.

Decision makers rarely, if ever, have perfect knowledge about the environment, instead deciding based on heuristic estimates subject to cognitive constraints [43]. Formal models of decision-making offer the potential to identify whether maladaptive decisions in BPD reflect a short-sighted policy (e.g., acting to preserve immediate proximity to a partner at the cost of losing the relationship), imprecise or biased subjective value estimates (e.g., believing that emotional intimacy is valuable regardless of relationship context), or overgeneralized representations that fail to incorporate new evidence (e.g., expecting others to be hostile based on prior conflicts).

**How can formal models of decision-making help to dissect clinical features of BPD?**

Our approach is motivated by the observation that individuals with BPD struggle to develop long-term relationships characterized by reciprocity, trust, and secure attachment [29,44]. Instead, relationships are often intense and dysfunctional, which may reflect maladaptive responses to others’ mental states [45,46], a poor understanding of appropriate behaviors across social domains [47], and unstable representations of others [27,48]. Indeed, ecological momentary assessment (EMA) studies have found increased positive feedback among interpersonal perception, affect, and behavior in BPD. For example, perceiving others as cold or quarrelsome triggers elevated negative affect and quarrelsome behavior in individuals with BPD, which elicits quarrelsome behavior from
others [49]. Using more nuanced time-lagged analyses, Scott and colleagues [50] recently found that the tendency for perceived rejection to evoke aggression was exacerbated by negative emotions (especially anger) in young women with borderline symptoms.

Altogether, clinical theory and both cross-sectional and EMA studies provide three important leads that inform a decision neuroscience approach to BPD. First, individuals with BPD are often insensitive to adverse consequences of their own behaviors. For example, during social economic exchanges, individuals with BPD fail to cooperate to maximize their gains [51] and make inconsistent choices [52]. Second, coping strategies (e.g., self-injury) and responses to interpersonal discord (e.g., overt hostility) are often incongruent with individuals’ longer-term goals [53]. Third, interpersonal perceptions and ensuing responses (e.g., perceived rejection eliciting aggression) are often rigidly coupled, especially during times of distress [50].

We propose that maladaptive, inflexible interpersonal behaviors reflect excessive influence of the Pavlovian system (stimulus-outcome) on social decisions. More specifically, conditioned responses (CRs) to social stimuli are likely a key component of disruptive interpersonal behaviors in BPD, from quarrelsomeness to self-destructive acts. The inflexibility of CRs and their sensitivity to internal state rather than consequences enables feedback loops, leading to escalating conflicts. Interpersonal problems are likely to emerge in situations where the motivational state (including emotional context) promotes actions with potentially aversive relationship consequences. For example, if one is angry with a seemingly distant partner, the congruent Pavlovian response is to start an argument (i.e., approach). This will often lead the partner to withdraw. Individuals with BPD may learn to associate the partner (the conditioned stimulus) with the experience of abandonment
(outcome) such that fights paradoxically escalate as the partner becomes more aloof in response (cf. [49,50]). Such dynamics are common early in therapy, when patients initially fail to understand how their actions affect others, focusing instead on how others affect them [29].

By contrast, adaptive social learning requires one to track others’ internal states [46,54] and the results of one’s actions in order to learn from interpersonal successes and mistakes. In the BPD literature, this has been called ‘mentalizing’ [29]; the learning theory homologue is goal-directed learning, which is defined as choosing an action by mentally traversing a transition structure to select a currently desired outcome (i.e., action-outcome) [55]. Returning to the example above, if one’s goal is to maintain a mutually rewarding relationship, one could learn from repeated experience that fighting (goal-incongruent action) pushes the partner away (outcome) and restrain oneself (goal-congruent action). Such learning could involve emotion reappraisal (e.g., viewing the partner as preoccupied, but not emotionally distant) in order to promote goal-congruent decisions [56]. Goal-directed learners are sensitive to action-outcome contingencies [18,42] and update their preferences in light of specific feedback.

**The role of Pavlovian-to-instrumental transfer (PIT) in BPD**

Integrating clinical, experimental, and neuroscience evidence, we propose that PIT is a key mechanism through which the Pavlovian system dominates decisions in BPD, leading to maladaptive choices that are sensitive to one’s motivational state and relatively insensitive to context and action-outcome contingencies. A comprehensive account of learning system alterations in BPD is beyond the scope of this brief review. We note, however, that stress disrupts goal-directed learning and enhances Pavlovian and habitual
tendencies [57–61]. Furthermore, uncontrollable stress during development has disruptive long-term effects on stress reactivity and learning from new experience [62–64]. This accords with evidence that individuals with BPD experience relationships as stressful and unpredictable, which may relate to early experiences with caregivers who are erratic to the point of abusiveness [65].

Research over the past decade has revealed behavioral and biological distinctions between two forms of PIT, outcome-specific and general [66]. Outcome-specific PIT describes the tendency of a conditioned stimulus (S₁) associated with an appetitive outcome (O₁) to invigorate actions (A₁) that lead to that outcome. For example, the smell of freshly baked cookies may lead one to seek them out. Likewise, aversive cues can promote behaviors to avoid undesirable outcomes [67]. One account of specific PIT is that a Pavlovian cue activates the representation of its associated outcome (i.e., S₁·O₁), which in turn promotes outcome-congruent actions (A₁·O₁) [68]. In contrast, general PIT describes how a cue (S₁) can promote actions (A₂…n) that are not specifically related to the cue and lead to a number of different outcomes (O₂…n) [69]. For example, the presence of appealing foods at a party may elicit other reward-seeking behaviors such as drinking wine or socializing. Thus, outcome-specific PIT reflects the role of Pavlovian cues in promoting actions that are likely to achieve a particular goal. Conversely, in general PIT, cues signal the availability of rewards and punishments in the broader environment, thereby altering the value of instrumental actions [66].

Crucially, general PIT depends on motivational state, whereas outcome-specific PIT does not [70]. Extending the concept of polarized affect states and representations of self and other from object relations theory [27], we view intense, changing emotions in BPD
Decision neuroscience perspective on borderline personality

(i.e., affective instability [71]) as motivational shifts that amplify the perceived value of emotion-congruent outcomes [32,56] and enhance Pavlovian influences via general PIT. Accordingly, an individual who experiences a strong 'hunger' to affiliate with others may assign high value to various social outcomes. In BPD, affiliative motivation often results in decisions that may have aversive consequences (e.g., pursuing a sexual relationship with a therapist) or that are inappropriate to the context (e.g., expressing intense jealousy about a friend spending time with others) [47]. Such behaviors fail to satisfy the motivational demand and often lead to aversive outcomes (e.g., a therapist rejecting the patient's sexual advances).

Furthermore, the intensity of motivational states in BPD (i.e., states in which some outcomes have very high perceived value) likely has a general invigorating effect on behavior that may interfere with long-term goals [72]. Viewed through the lens of learning theory, this may reflect three related effects: 1) the enhancement of Pavlovian-congruent actions via general PIT, irrespective of goal; 2) a reduction in the overall threshold for executing actions [73]; and 3) the tendency to execute habitual actions more vigorously regardless of the value of the outcome [42]. For example, intense emotional distress following an argument may lead an individual with BPD to pursue casual sex or to have a drug relapse even if these are incongruent with a goal of preserving a long-term relationship. Altogether, the persistence of extreme motivational states in BPD suggests a difficulty mapping interpersonal contingencies in a goal-directed fashion or an overreliance Pavlovian influences that may not align with situational demands.

**Neural substrates of Pavlovian-instrumental interactions**

We predict that maladaptive behaviors in BPD involve specific alterations in the
Decision neuroscience perspective on borderline personality
cortico-limbic circuitry underlying Pavlovian-instrumental interactions (detailed in Figure 1). We propose that the phylogenetically old pathway from the central nucleus of the amygdala (CNA) and ventral tegmental area (VTA) to the core of the nucleus accumbens (NAcc-C) underpins emotion-driven Pavlovian responses, as in general PIT [74].

Furthermore, general PIT tends to invigorate habitual responses [66], which are not sensitive to the value of the resulting outcome and thus may contribute to problems such as reactive aggression or substance abuse. Invigoration of goal-congruent behaviors by environmental stimuli, on the other hand, is likely mediated by the phylogenetically newer pathway from the basolateral amygdala (BLA) to the shell of the nucleus accumbens (NAcc-S) and by the recurrent projections of the ventromedial prefrontal (vmPFC) and medial orbitofrontal cortex (mOFC) into the NAcc-S, as in outcome-specific PIT [69,75]. Finally, we believe that projections from the medial prefrontal cortex (mPFC) to the striatum may facilitate the suppression of Pavlovian influences on instrumental behavior [76] and play a key role in the shift from reactive to deliberative behaviors in treatments for BPD.

**Toward a decision neuroscience account of BPD**

This new approach recasts a frontolimbic account of BPD in terms of the interplay among computational modes, and it is solidly grounded in experimental psychology, neuroscience, and computer science [21,26]. More than simply describing behavior, formal models of Pavlovian-instrumental interactions [77,78] enable scientists to understand individual differences in latent decision processes and their neural signatures. These models have yielded novel behavioral and neurobiological insights into alcohol dependence [79], schizophrenia [80], and depression [81]. Furthermore, recent advances in high-resolution functional neuroimaging [82] offer the potential to study subregions of the
amygdala and striatum that are differentially implicated in general and outcome-specific PIT [83]. Thus, formal models of decision-making support theory-guided accounts of behavior during fMRI experiments, providing a crucial bridge between the complex clinical features of BPD and the latent neurobehavioral systems that underpin them [20,84]. We believe that a decision neuroscience approach that experimentally dissects Pavlovian-instrumental interactions in the context of social decisions [85] will advance our understanding of BPD.
Box 1. Levels of analysis in borderline personality research

[Text to appear adjacent to figure] Empirical research on the modern construct of borderline personality disorder (BPD) emerged from the neo-Kraepelinian descriptive psychiatry movement of the 1970s [86]. Drawing on psychodynamic theories of BPD [27], scientists articulated a view of BPD as a clinical syndrome with specific symptoms (e.g., intense, unstable relationships) that could be diagnosed reliably by clinicians [87]. Although there have been a few revisions to the diagnostic criteria for BPD over the past four decades, the DSM-5 [88] retained the definition of the DSM-IV, requiring that five of nine criteria be present for a diagnosis.

Problems with boundary overlap and co-occurrence among personality disorders (PDs) [89] have motivated an alternative view that reframes borderline personality in terms of dimensional variation on a number of psychological constructs such as impulsivity [90]. Indeed, proponents of a dimensional view question the diagnostic utility of BPD [91]. Furthermore, in order to develop an empirically based taxonomy of psychopathology, it is essential to consider borderline personality vis-à-vis other PDs and mental disorders. For example, in analyses of the meta-structure of psychopathology, Eaton and colleagues [92] found that borderline personality had both internalizing and externalizing aspects. These two superordinate dimensions explain covariation among many mental disorders [93] and may be nested under a general psychopathology (aka ‘p’) liability factor [94]. Indeed, Sharp and colleagues [95] found that BPD symptoms were largely explained by a general personality pathology factor.

Debates about the latent structure and ontological status of borderline personality arise in part from the observational nature of interview and self-report methods.
Experimentation is essential to elaborate a mechanistic account of cognitive processes and neural abnormalities that underlie the behavioral manifestations of borderline personality. We focus specifically on disadvantageous decision-making such as desperately seeking intimacy with a partner even when it causes the partner to withdraw emotionally. Early experimental research on decision-making used paradigms such as the Iowa Gambling Task (IGT) to compare summary statistics (e.g., number of risky bets) between patient and control groups. This work identified abnormalities across various mental disorders, but no disorder specificity [38]. Crucially, summary statistics provide limited insight into the cognitive processes underlying each behavioral output during the experiment (e.g., a single decision to gamble). As detailed in the main text, formal models of decision-making, such as reinforcement learning (Box 3) offer a generative account of how underlying cognitive processes lead to specific choices.

Formal models provide a crucial bridge linking latent decision processes with the dynamics of underlying neural systems [20] measured at the level of neuronal populations (e.g., fMRI or EEG). This is in contrast to the oversimplified view of a one-to-one correspondence between psychological constructs and neural systems (e.g., the neural correlates of neuroticism), which leaps across multiple levels of analysis and has not substantially advanced theories of personality disorders [96,97]. Formal models instantiate testable hypotheses about the latent processes underlying decision-making that are proximate to specific behavioral outputs. The structure of such models mirrors the organization of neural populations into input-output units that are responsible for processing specific information. While we see value in studying borderline personality at many levels of analysis, we believe that formal models decision-making paired with
functional neuroimaging can provide new insights into BPD [for a cogent discussion of the importance of behavior in neuroscience, see 98].
Box 2. Using formal models to elucidate cognitive representation and neural systems

Over the past three decades, cognitive neuroscientists have used functional neuroimaging to describe neural activity elicited by stimuli presented sequentially during an experiment. In traditional fMRI paradigms, regional activation to one type of stimulus (e.g., faces) can be compared to another stimulus (e.g., houses) or to no task (also called the resting baseline). For example, a region involved in reward processing may exhibit greater activity to the receipt of a reward than at rest. Such analyses identify regions that are sensitive to differences in experimental conditions (e.g., faces compared to houses), but they do not necessarily elucidate the underlying cognitive architecture or its neural implementation.

By contrast, formal, algorithmic models of behavior seek to predict what an individual will do. In such models, equations and quantitative parameters specify an explicit mapping between latent cognitive processes and task-relevant behavior. Individual differences in parameter values may reflect intermediate phenotypes that are associated with broader psychological constructs.

The use of formal models for bridging concurrent behavioral and neural data has been termed model-based cognitive neuroscience [99]. In model-based fMRI analyses, time-varying estimates of a cognitive process are convolved with a hemodynamic response function to derive the predicted BOLD activity for that process. One can then extend standard voxel-wise general linear model analyses to identify regions whose activity scales parametrically with the model-based regressors. Such regions are likely to be involved in the representation of a cognitive process instantiated in a theoretically motivated model. Importantly, a model may be falsified if it provides a poor account of behavior or
demonstrates flaws in generative simulations. Moreover, a model may describe behavior well, but its components may not be associated with neural activity (i.e., a failure to bridge levels of analysis).

The structure of formal models depends on the scientific question — indeed, a model describes information processing in the context of a specific task. For example, many models of value-based decision-making specify how individuals update their estimates of subjective value based on surprising outcomes (i.e., prediction errors; see Box 3), whereas models of category learning may focus on differentiating prototypes from exemplars [100]. For a thorough exposition of model-based cognitive neuroscience we refer readers to [19].

The growing field of computational psychiatry uses formal models and model-based neuroimaging analyses to characterize psychopathology in terms of latent cognitive processes that underlie symptoms and maladaptive behaviors [101]. Computational psychiatry may also elucidate meaningful clinical phenotypes that are based purely on theoretically motivated formal models rather than expert consensus [102]. Altogether, the marriage of formal models of behavior with functional neuroimaging enables more specific tests of cognitive representation, thereby constraining and clarifying the interpretation of neural activity.
Model-Based Cognitive Neuroscience

Decision-making paradigm

Traditional fMRI Data Analysis
- Cognitive Representation: Occurrence versus non-occurrence of stimulus presentation
- Design Matrix
- Prediction Error vs. Feedback
- Brain Activity Associated with Feedback

Model-based fMRI Data Analysis
- Cognitive Representation: $V_{t+1} = V_t + \alpha(R_t - V_t)$
- Design Matrix
- Prediction Error vs. Feedback
- Brain Activity Associated with Prediction Error

Warmer colors indicate increased activation during feedback. Cooler colors indicate decreased activation during feedback. Thresholded at $|z| > 3.79$
In traditional and model-based fMRI analyses, participants complete an experimental task designed to elicit a cognitive process and/or behavior of interest. In this example, on each trial, a dot revolves around a circle over the course of four seconds [103]. Responding at any point during the revolution terminates the trial and leads to a unique probabilistic reward displayed during a feedback phase lasting 900ms. Thus, participants must learn when to respond in order to obtain larger rewards.

In general linear model (GLM) analyses of fMRI data, the design matrix encodes the onset, duration, and magnitude of stimuli presented during the experiment. To estimate the predicted blood-oxygen-level-dependent (BOLD) response (i.e., the conventional index of neural activity in fMRI), each regressor in the design matrix is convolved with a canonical hemodynamic response function (convolution with HRF).

In traditional analyses (left column), feedback events are binary (design matrix) and reflect whether participants were viewing the feedback stimulus at that time. The major difference in model-based fMRI analyses is that the design matrix represents a cognitive process estimated according to a formal model of behavior (cognitive representation). Here, we illustrate the magnitude of prediction errors in six trials from a single subject, fit according to a delta rule learning model (see Box 3). Thus, the brain activity associated with the model-based signal describes the response of functional regions to the magnitude of prediction errors, whereas regions associated with the feedback signal reflect overall activity without a specific hypothesis of what cognitive operations are involved processing feedback.

Box 3. Reinforcement learning: A primer
Reinforcement learning (RL) is an algorithmic account of how agents learn to make decisions on the basis of rewards and punishments (to be brief, we refer to rewards below). Each decision point is a state marked by cues (stimuli) that may be predictive of rewards, as well as available actions. In Pavlovian learning (stimulus-outcome), if a stimulus (e.g., light) is present in a given state and a reward is subsequently delivered, the agent learns the expected reward value of the stimulus. In instrumental learning (action-outcome), by selecting an action, the agent can observe its outcome and update the expected value of that action. In both cases, agents seek to minimize the discrepancy between expected and obtained outcomes. This discrepancy is called the prediction error (PE), a key learning signal. Prediction errors are positive if the outcome is greater than expected and vice versa. Most RL models update expected value estimates on the basis of PEs according to a form of delta rule learning [21,24]:

$$V_{t+1} = V_t + \alpha (R_t - V_t)$$

where $V$ is the expected value of a stimulus or action, $R$ is reward, $(R_t - V_t)$ is the PE in the current learning episode, and $t$ is time (across learning episodes). The impact of each PE on the current value estimate is determined by the learning rate parameter $\alpha$, which conventionally varies between zero and one. The gradual evolution of expected value across multiple learning episodes gives rise to a learning curve in which an agent’s estimate converges toward the underlying (unknown) reward contingency. Human imaging studies map expected value representations to the ventromedial prefrontal cortex (vmPFC), whereas PE signals are found in the ventral and dorsal striatum, which receive projections from the dopaminergic midbrain, as well as the thalamus [see 104 for more on the correlates of RL in the human brain].
Some important behaviors fall outside of this simple stimulus-outcome or action-outcome learning framework. For example, in Pavlovian learning, although no proper actions are involved, the organism emits appetitive or aversive conditioned responses (CRs) in response to predictive cues (i.e., conditioned stimuli). From an evolutionary standpoint, CRs are genetically determined adaptive behaviors, such as freezing in response to a threat or licking in response to food cues [105]. In the case of instrumental learning, when the value of an action in a given state is over-trained, habits emerge, where cues signifying a specific state elicit an automatic action (i.e., stimulus-response pairings). For example, one may fail to exit the highway to meet friends at a restaurant because the exit falls along the usual route from home to work. Finally, as detailed in the main text, important interactions between these learning modes have been highlighted by recent research on Pavlovian-to-instrumental transfer (PIT; for a review, see [66]).
Figure 1. Mapping clinical manifestations of borderline personality in terms of learning theory and neural systems

<table>
<thead>
<tr>
<th>Clinical phenomenon in BPD</th>
<th>Learning Theory Construct</th>
<th>Neural substrates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-down control over maladaptive behavior, possibly including shifts from reactive to deliberative decisions following treatment</td>
<td>Resistance to Pavlovian influences</td>
<td>Medial prefrontal cortex</td>
</tr>
<tr>
<td>Difficulty learning from negative consequences of interpersonal mistakes</td>
<td>Goal-directed learning</td>
<td>Dorsolateral prefrontal cortex</td>
</tr>
<tr>
<td>Limited mentalizing</td>
<td></td>
<td>Dorsomedial striatum</td>
</tr>
<tr>
<td>Goal-incongruent emotion-driven behaviors:</td>
<td></td>
<td>Dorsolateral striatum</td>
</tr>
<tr>
<td>- Interpersonal aggression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Self-injury</td>
<td>Non-specific invigoration of instrumental responses by Pavlovian values (general PIT)</td>
<td>Nucleus accumbens core</td>
</tr>
<tr>
<td>- Impulsive substance use</td>
<td></td>
<td>Central nucleus of amygdala</td>
</tr>
<tr>
<td>Goal-congruent emotion-driven behaviors:</td>
<td></td>
<td>Ventral tegmental area</td>
</tr>
<tr>
<td>- Interpersonal manipulation</td>
<td>Enhancement of goal-congruent instrumental responses by Pavlovian values (specific PIT)</td>
<td>Nucleus accumbens shell</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basolateral amygdala</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medial orbitofrontal cortex</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ventromedial prefrontal cortex</td>
</tr>
</tbody>
</table>

Figure 1 Caption. Brain images were adapted from the Allen Institute Human Brain atlas ([http://atlas.brain-map.org/atlas?atlas=265297126](http://atlas.brain-map.org/atlas?atlas=265297126)) [106].
Annotated Bibliography

* [11] Schulze 2016: Meta-analysis of several functional imaging studies of emotion processing in BPD that finds hyperactivity of the amygdala and posterior cingulate cortex to negative emotional stimuli, but only in unmedicated samples.

** [18] Huys 2015: Provides an outstanding introduction to the formalism of Bayesian decision theory and how it can help to dissect maladaptive decisions in psychopathology.

* [19] Turner 2017: An outstanding review that explores different approaches for synthesizing neural data with formal models of behavior. Covers both the theoretical differences among approaches and the empirical findings that each has yielded.

** [20] Love 2015: A conceptual paper that argues for the importance of formal, algorithmic models in understanding how the brain implements specific cognitive processes that are expressed in behavior (i.e., bridging these levels of analysis).

** [33] Paret 2017: Meta-analysis of behavioral decision studies of BPD that describes key effects for delay discounting, reversal learning, and Iowa Gambling Task (IGT). Across studies, individuals with BPD displayed a preference for immediate rewards independent of acute stress, and they performed more poorly on the IGT.

* [34] Krause-Utz 2016: Describes the effect of acute stress on several behavioral measures of impulsivity in BPD. The BPD group exhibited difficulty withholding in appropriate responses on a speeded choice task under stress.

* [38] Mukherjee 2014: A meta-analysis of value-based decision tasks in psychopathology. Focuses especially on the Iowa Gambling Task, finding that decision deficits are common across mental disorders. Underscores the importance of employing decision battery assessments that account for specific cognitive processes in order to differentiate among disorders.

* [46] Miano 2017: An observational study of conversation dynamics in romantic couples that found greater empathic accuracy to relationship-threatening content in BPD, suggesting a promising latent process underlying disturbed interpersonal functioning.

* [48] Beeney 2016: Synthesizes functional MRI and behavioral data to demonstrate that inconsistency in representations of self and others mediates hyperactivity of social cognitive regions such as mPFC in BPD.

* [60] de Berker 2016. The authors demonstrate that an acute stressor (socially evaluated cold pressor task) impairs the ability to learn an instrumental action, regardless of the valence of the outcome. This is consistent with the possibility that aversive Pavlovian influences impair instrumental approach, likely a form of Pavlovian-to-instrumental transfer.

** [80] Morris 2015: An empirical study of Pavlovian-to-instrumental transfer and goal-directed learning in schizophrenia. Patients persisted in choosing a devalued outcome (food) despite rating it more negatively, suggesting problems updating action-outcome representations. This deficit was associated with reduced caudate activity during decisions.

** [81] Huys 2016: An empirical study of Pavlovian-to-instrumental transfer that identified reduced ability of Pavlovian cues to invigorate congruent actions (e.g., positive cue–approach behavior) in currently depressed individuals. Greater congruence of PIT effects predicted the degree of symptom recovery at 4-6-month follow-up.

* [104] Chase 2015: Meta-analysis of human imaging studies that mapped prediction error and expected value signals using reinforcement learning, accompanied by a brief overview of the implementation of reinforcement learning in functional imaging.

* Of special interest
** Of outstanding interest
References


