

The Constant Function at the Other: A Structural Model of Belief Maintenance in Borderline Personality Disorder

Gökhan Bahtışen

Independent Researcher, Istanbul, Türkiye

Correspondence: gokbahtisen@gmail.com

2026

Author Note

Gökhan Bahtışen is an independent researcher based in Istanbul, Türkiye. There are no conflicts of interest to declare. This is a theoretical paper; no empirical data were collected or analyzed. An AI assistant (Claude, Anthropic) was used during manuscript preparation for literature organization, structural editing, and prose refinement. All theoretical arguments, diagnostic frameworks, predictions, and the cross-disorder architecture are the author's original intellectual contributions. All citations were verified against primary sources by the author.

Correspondence concerning this article should be addressed to Gökhan Bahtışen. Email: gokbahtisen@gmail.com

The Constant Function at the Other: A Structural Model of Belief Maintenance in Borderline Personality Disorder

Note: This paper is a theoretical proposal. It derives predictions from a structural model of belief construction and maintenance (Bahtışen, 2026a)¹ and evaluates them against existing clinical literature on Borderline Personality Disorder. It does not claim to explain BPD. It asks whether existing clinical evidence is consistent with the specific predictions the model generates and identifies where it is not.

Abstract

This paper proposes that the interpersonal patterns characteristic of Borderline Personality Disorder (BPD)—idealization, devaluation, identity disturbance, and abandonment sensitivity—can be understood as expressions of a specific belief-maintenance architecture. Drawing on a structural model of persistent belief-like narrative attributions maintained for regulatory purposes (Bahtışen, 2026a), the paper identifies a three-component system: (1) an invariant attribution structure exculpating the self regardless of surface presentation, (2) the assignment of an impossible regulatory function to another person defined in what it must deliver but not in how, and (3) rapid oscillation between two incompatible narrative frames, each internally consistent but too brief to be tested against reality. The paper introduces the distinction between first-order and second-order constant functions to account for BPD's characteristic instability. Five predictions are derived and evaluated against peer-reviewed literature: splitting as rapid narrative cycling, the partner as externally installed agent, a cortisol-prefrontal impairment pathway, self-blame as cycle fuel rather than exit, and crisis as belief-intensifier. Clinical literature is broadly consistent with all five predictions, though the model's contribution varies from genuine structural novelty to reframing of established insights. The paper addresses equifinality across therapies with different mechanisms,

proposes novel testable predictions—including oscillation as measurable second-order invariance and self-blame intensity as a function of cycle count—and specifies falsification conditions, offering a structural complement to existing etiological accounts of BPD.

Keywords: borderline personality disorder, splitting, constant function, belief updating, exculpatory self-attribution, devaluation, emotion regulation

1. Introduction

1.1 The Structural Model

A recently proposed model of persistent belief-like narrative attributions (Bahtışen, 2026a) identifies a recurring architecture across domains—financial, political, ideological, interpersonal—in which intelligent individuals construct and defend belief structures that diverge from observable reality, even when they possess the cognitive resources to recognize the divergence. A terminological note: "belief" as used throughout this paper refers not to propositional claims or delusions but to narrative attributions used for regulation—the implicit or explicit stories people construct about who is responsible for what, who can be trusted with which functions, and how the world operates. These attributions may never be consciously articulated as beliefs, yet they organize behavior and resist correction in the ways the model describes.

The model's core claim is that such persistent narrative attributions share three structural features. The first is a function-mechanism gap. The belief has a clearly defined function—what it must deliver: meaning, safety, identity, regulation—but a systematically undefined mechanism—how, specifically, the belief delivers it. A person who believes a particular investment will produce wealth can articulate the function (financial security) but cannot

trace the mechanism by which the investment achieves it. A person who believes a political leader will restore justice can articulate the function (a fair society) but cannot specify the process by which one individual accomplishes this. The gap between what the belief promises and how it delivers is not incidental. It is where the architecture lives.

The second feature is agent installation. Into the gap between function and mechanism, the believer installs an agent—a person, force, or principle assigned a role it cannot reliably fulfill through any specified process. The model draws on established findings in cognitive psychology: teleological bias (Kelemen, 1999, 2004), the brain's default tendency to explain events through agents and purposes rather than impersonal mechanisms, and hyperactive agent detection (Epley, Waytz, & Cacioppo, 2007) combine to make agent-based explanations the cognitive path of least resistance. The agent fills the gap. Its presence transforms the undefined mechanism into something that feels like an explanation—even though the gap between function and mechanism has been filled with a person or principle rather than with a traceable process.

The third feature is a neurochemical lock. Once installed, the agent is maintained through a self-reinforcing feedback loop involving the stress and reward systems. The model proposes a six-stage sequential process: (1) agent attribution driven by default cognitive biases; (2) stress-induced sensitization of the mesolimbic dopamine pathway via chronic cortisol elevation; (3) hijacking of the anticipatory reward system, where the brain responds more strongly to anticipated than actual reward; (4) replacement of reality-contact by imagination, where mental rehearsal of the belief activates reward circuits while depleting motivation for reality-testing; (5) a self-reinforcing cortisol-dopamine loop in which the belief structure becomes the dominant source of neurochemical relief; and (6) resistance to correction through emotional flooding that blocks the prefrontal encoding required for belief revision. Each stage draws on independently peer-reviewed findings; the full sequential chain is

advanced as a set of falsifiable hypotheses rather than a validated causal account (Bahtışen, 2026a).

The model also identifies a property it calls the constant function: the invariant output that a persistent belief structure preserves regardless of what input it receives. In a constant function, input varies but output does not. Evidence that confirms the belief is accepted directly; evidence that disconfirms it is processed through narrative revision that preserves the conclusion. The belief system functions, in this sense, like a mathematical constant function: $f(x) = c$, regardless of x .

Together, these features describe an architecture in which a belief resists correction not because the believer is irrational or unintelligent but because the belief has become structurally and neurochemically self-maintaining. The architecture generates a diagnostic question applicable across domains: Is there an agent installed where a mechanism should be?

For clarity, this paper uses "the structural model" to refer to the parent theory (Bahtışen, 2026a) and "the present analysis" to refer to this paper's BPD-specific application and extensions.

1.2 Why Borderline Personality Disorder

If this architecture describes a general process of belief construction and maintenance, a testable implication follows: personality disorders characterized by systematic interpersonal narrative distortions should produce specific, predictable variations in how the architecture manifests. The variations should be structurally required by the model—not added after the fact to accommodate clinical observations.

BPD is the strongest available test case, for three reasons.

First, BPD involves identity-level narrative construction organized around interpersonal relationships—precisely the conditions under which the model's mechanisms should operate most directly. The person's self-concept, emotional regulation, and behavioral patterns are organized around narrative assignments of the other person's role. These assignments have the structural properties the model describes: a defined function (regulate my internal state, complete my identity) served by an undefined mechanism (how, specifically, does another person's presence accomplish this?).

Second, the BPD literature is the richest and most empirically developed among personality disorders characterized by interpersonal disturbance. A theoretical framework tested against thin evidence cannot demonstrate that it generates novel predictions rather than filling gaps with plausible-sounding stories. BPD provides enough published evidence to genuinely evaluate the model's predictions—including enough to reveal where the model merely restates what is already known.

Third, BPD has the most active treatment research of any personality disorder, with multiple evidence-based therapies operating through different stated mechanisms—Dialectical Behavior Therapy (Linehan, 1993), Mentalization-Based Treatment (Bateman & Fonagy, 2004), Schema Therapy (Young, Klosko, & Weishaar, 2003). This creates a natural challenge for the model: if it describes the maintenance mechanism correctly, it should have something specific to say about why treatments work, and it should be vulnerable to the problem posed by treatments that work through apparently different mechanisms.

1.3 What This Paper Does and Does Not Claim

This paper does not claim to explain BPD. The etiology of BPD involves genetic vulnerability, developmental trauma, neurobiological factors, and environmental conditions that the structural model does not address. What this paper does is derive specific predictions

from the structural model's existing architecture and evaluate whether the clinical literature is consistent with, contradicts, or has not yet addressed those predictions. Rather than merely relying on the parent theory's (Bahtışen, 2026a) validity, this paper uses BPD as an extreme clinical stress-test to evaluate whether the proposed architecture can produce independent, testable, and falsifiable predictions in a highly complex domain.

The paper distinguishes explicitly between three types of contributions. Some predictions identify structural patterns that existing frameworks have not formalized—these represent genuine theoretical work. Some predictions provide a neurochemical vocabulary for insights that clinical practice already embodies—these represent translation rather than discovery. And some predictions may prove, on examination, to be restatements of established findings in new terminology—these represent the risk any theoretical framework faces when applied to a well-studied domain. The paper aims to be transparent about which category each prediction falls into, rather than presenting all five as equally novel.

One further clarification: the original structural model (Bahtışen, 2026a) was developed to describe belief persistence in non-clinical domains and did not address personality disorders. Applying it to BPD requires a theoretical extension—specifically, a distinction between first-order and second-order constant functions—that this paper introduces for the first time. Section 2.1 develops this extension and identifies where it departs from the original model.

2. The Architecture: Three Components of BPD's Belief-Maintenance System

2.1 The Constant Function and a Theoretical Extension

As described in Section 1.1, the structural model identifies a property it calls the constant function: the invariant output that a persistent belief structure preserves regardless of what input it receives. In the original model (Bahtışen, 2026a), the constant function is a first-order

property—the output itself does not change. A person who believes they are exceptional will interpret praise as confirmation and criticism as jealousy; the output ("I am exceptional") remains the same. This first-order formulation works cleanly for cases where the belief content is rigid: financial convictions, political commitments, grandiose self-narratives.

BPD does not work this way. The surface outputs oscillate dramatically: the other person is divine, the other person is monstrous; I am helpless, I am righteous; the relationship is salvation, the relationship is destruction. If the model required invariant outputs to identify a constant function, BPD would be a disconfirming case—and a significant one, because BPD is precisely the kind of persistent, narrative-driven, correction-resistant pattern the model claims to describe.

This paper proposes a theoretical extension to resolve this problem: the distinction between first-order and second-order constant functions. This distinction is not present in the original structural model. It is introduced here for the first time, motivated by the specific challenge BPD poses to the framework.

In a first-order constant function, the output itself does not change: input varies, output stays the same. In a second-order constant function, the outputs change freely, but the rule that generates them remains invariant. BPD, on this account, operates a second-order constant function: the assessments of self and other shift dramatically, but the attribution rule that produces them does not.

The rule is: the self is exculpated from causal responsibility, and responsibility for the self's emotional state is assigned to the other. (This paper uses "exculpatory self-attribution" for this pattern to emphasize causal externalization and helplessness rather than necessarily a claim of moral purity or innocence. It strictly denotes who is assigned causal power, though the informal term "innocence" is occasionally retained in illustrative passages for readability.)

This construct must be distinguished from superficially similar ones. Exculpatory self-attribution is not external locus of control, which is a generalized trait-level tendency to attribute outcomes to external forces; the BPD pattern is specific to interpersonal narratives and coexists with internal attribution in other domains. It is not self-serving bias, which preserves positive self-evaluation; BPD patients frequently evaluate themselves negatively while still exculpating themselves from causal responsibility (Section 3.4). It is not defensive attribution bias, which operates on single events; the present construct is a generating rule preserved across incompatible narratives—the same rule produces both idealization and devaluation. And it is not shame-based global self-condemnation, which assigns fault to the self as a type ("I am broken"); the present analysis predicts that such condemnation coexists with the exculpatory rule and functions as fuel for the cycle rather than as a genuine attribution reversal. The differentiator is not that the individual blames others, but that a specific generating rule is maintained across narrative states that would otherwise be contradictory.

During idealization, the other is the Savior whose presence makes the self's suffering bearable—the self is helpless/innocent, the suffering is real, and the other is the answer. During devaluation, the other is the Villain whose betrayal causes the self's suffering—the self is helpless/innocent, the suffering is real, and the other is the cause. Both states are outputs of the same generating rule. The surface changes; the rule does not.

This distinction matters because it determines what falsification looks like and because it makes the constant function concept testable in cases of apparent instability. A first-order constant is falsified when the output changes—if a person with NPD genuinely concludes "I am ordinary," the constant has broken. A second-order constant is falsified not when the outputs change (they are expected to change) but when the generating rule changes—if a person with BPD, during a splitting episode, genuinely assigns fault to self and exculpates the

other at the level of the generating rule, the second-order constant has broken. These are different empirical tests for what appears, on the surface, to be the same theoretical construct. The extension thus preserves the model's falsifiability rather than undermining it—but it does so by adding theoretical machinery that the original model did not contain.

The second-order constant generates a specific, testable prediction: during splitting episodes, regardless of whether the individual's surface self-presentation is "helpless victim," "righteous accuser," or "selfless giver," the underlying attribution should consistently exculpate the self and assign fault to other. If this prediction fails—if BPD splitting episodes produce genuine attribution reversals in which the individual assigns fault to self and exculpates the other at the level of the generating rule—the second-order constant formulation is falsified.

A complication must be acknowledged immediately. BPD involves well-documented self-blame, self-hatred, and self-directed aggression. On the surface, this appears to contradict the claim that the attribution structure exculpates the self. Section 3.4 addresses this directly: the present analysis predicts that BPD self-blame functions as fuel for the cycle rather than as a genuine attribution reversal—the self-blame is incorporated into the narrative of the self as the one who suffers, confirming exculpation through martyrdom rather than assigning responsibility through accountability. This prediction is testable and falsifiable.

2.2 The Agent in the Other Person

That BPD patients assign their relationship partners an impossible regulatory function is well established. Object relations theory describes how the partner is assigned the role of the "all-good" or "all-bad" object (Kernberg, 1967, 1985). Attachment theory characterizes BPD as an "interpersonal hypersensitivity phenotype" in which the individual's entire regulatory architecture is organized around the other person's presence or absence (Gunderson & Lyons-

Ruth, 2008). Stanley and Siever's (2010) neuropeptide model goes further, demonstrating that the partner literally functions as an external neurochemical regulator—opioid, oxytocin, and vasopressin systems mediate a dependency in which the partner's proximity produces safety and their absence produces a state functionally analogous to substance withdrawal. These frameworks describe the clinical pattern with greater precision than the structural model can offer.

What the structural model adds is not a better clinical description but a cross-domain diagnosis. The model's vocabulary—function-mechanism gap, agent installation—reframes what these clinical traditions describe as a specific instance of a general architecture. The partner has been assigned a function (regulate my emotions, complete my identity, make the world safe) with no defined mechanism (how does another person's presence produce internal coherence?). This is the same structural operation the model identifies when a person installs a financial guru as the agent who will deliver wealth without specifying how, or a political leader as the agent who will deliver justice without specifying how. The content differs—financial security, political salvation, emotional regulation—but the architecture is formally analogous: a clearly defined function, a systematically undefined mechanism, and an agent installed in the gap between them.

The claim is structural analogy—shared architectural features that generate shared predictions—not identity of mechanisms across domains. The neurochemistry of interpersonal attachment is not the neurochemistry of financial speculation; the structural model claims only that both involve agent installation into a function-mechanism gap, and that this shared feature predicts shared maintenance dynamics.

The value of this cross-domain connection is that it generates predictions the clinical frameworks do not. If BPD's interpersonal pattern shares an architecture with non-clinical

persistent narrative attributions, the maintenance and exit conditions should also be shared. Specifically: the belief should be hardest to revise during crisis (when cortisol impairs prefrontal function) and easiest to revise during periods of relative safety (when prefrontal capacity is restored). The agent assignment should resist correction not because the patient lacks insight but because the neurochemical lock described in Section 1.1 makes revision physiologically costly. The five predictions in Section 3 follow from this structural homology.

2.3 The Oscillation: Why the Agent Switches Roles

BPD's belief architecture has a structural feature absent from most other applications of this model: the installed agent oscillates between two incompatible roles. During the Savior phase, the partner's capacity to regulate and save is treated as limitless; disconfirming evidence (ordinary limitations, failures of attunement) is attributed to external circumstances. During the Scapegoat phase, the partner's capacity for cruelty is treated as equally limitless; disconfirming evidence (genuine kindness, attempts at repair) is attributed to manipulation. In both phases, disconfirming evidence is absorbed through elastic attribution rather than permitted to update the model.

The speed of this oscillation is structurally necessary. Neither narrative can survive sustained contact with the real person, who will inevitably behave in ways inconsistent with both assignments. Each disconfirmation threatens the current narrative; rather than moving toward an integrated assessment of the partner as finite and mixed, the system switches to the opposite narrative, where the disconfirming evidence becomes confirming. The partner's kindness during a Scapegoat phase triggers a switch to the Savior phase (where the kindness confirms); boundary-setting during a Savior phase triggers a switch to the Scapegoat phase (where boundary-setting confirms).

This generates a novel prediction: the split should occur at the point of maximum divergence between the real person's behavior and the current narrative assignment—not driven by endogenous mood shifts but by the structural impossibility of maintaining an agent assignment against evidence produced by a real human being. This prediction has not, to our knowledge, been directly tested.

2.4 The Complete Architecture

The complete BPD belief-maintenance system, as derived from the structural model, involves three simultaneously operating components:

- Component 1 – The invariant attribution structure: Regardless of surface presentation—helpless, righteous, self-blaming—the deep attribution exculpates the self and assigns responsibility to other. This component is always active. It is the background structure, not a state that alternates.
- Component 2 – The Savior narrative: The other person's presence is the mechanism of regulation, completion, and safety. The function is defined; the mechanism is not. Active during idealization.
- Component 3 – The Scapegoat narrative: The other person's malice or inadequacy is the cause of the self's suffering. The function of this narrative (explaining why the self suffers) is defined; the mechanism (how one person has this much destructive power) is not. Active during devaluation.

Components 2 and 3 alternate. Component 1 runs continuously, providing the structural continuity that makes the oscillation possible without generating the cognitive dissonance that would force genuine revision. The self's exculpation is the thread that connects two incompatible stories about the same person: in both stories, the self is the one to whom things happen, never the one who generates them.

This three-component architecture generates the five predictions evaluated in the following section. The loop operates as a sequential, self-reinforcing process. We term this the BPD Belief-Maintenance Loop:

1. Invariant Baseline: The system begins with an exculpatory self-attribution (the generating rule, or second-order constant).
2. Agent Installation: An agent is installed in the partner (function defined, mechanism undefined).
3. Oscillation: The individual toggles between the Savior narrative and the Scapegoat narrative, driven by behavior-narrative divergence.
4. Collapse: The agent assignment inevitably fails, resulting in an interpersonal crisis.
5. Physiological Response: The crisis activates the stress system.
6. Cognitive Blockade: Stress-induced prefrontal impairment blocks genuine belief updating.
7. Maladaptive Processing: Self-blame generates further shame, which activates the stress system rather than behavioral change (see Section 3.4).
8. Desperation: The individual experiences a critical need for neurochemical relief.
9. Restart: A new agent is installed (or the cycle restarts with the same partner), and the loop repeats.

The loop's self-reinforcing character is driven by the structural impossibility of the initial agent assignment.

3. Five Predictions and Their Evaluation

Each prediction below is derived from the structural model's architecture and evaluated against existing peer-reviewed clinical literature. For each prediction, the paper identifies:

what the model predicts, what the clinical literature shows, what the model adds beyond existing frameworks, and—critically—what it does not add. Where an alternative account explains the same evidence without the structural model, that alternative is acknowledged. Table 1 summarizes the predictions, their novelty status, and their key falsification conditions.

Table 1
Predictions, Novelty Classification, and Falsification Conditions

Prediction	Status	Key falsifier
3.1 Splitting as narrative cycling	Reframing (structural account of known phenomenon)	Oscillation speed uncorrelated with behavior-narrative divergence; correlated instead with cognitive inhibition measures alone.
3.2 Partner as installed agent	Reframing (cross-domain structural homology of known pattern)	BPD interpersonal patterns share no structural features with non-clinical agent-installation; cross-domain predictions fail.
3.3 Cortisol-prefrontal pathway	Reframing (novel specificity within established cascade)	Crisis severity uncorrelated with impossibility of function assigned to partner; correlated only with general dysregulation.
3.4 Self-blame as cycle fuel	Novel	Post-split self-blame shows genuine rule-level reversal (causal specificity + partner exoneration + reparative intent) rather than global self-condemnation.
3.5 Crisis intensifies belief	Translation (neurochemical vocabulary for Linehan's insight)	Interpersonal crises reliably produce narrative revision rather than narrative hardening.
5.1 Splitting trigger = behavior-narrative divergence	Novel	Splits triggered by endogenous mood shifts independent of partner behavior.
5.2 Self-hatred correlates with cycle count	Novel	Self-hatred intensity predicted by objective harm caused, not by number of prior cycles.
5.3 Neuroimaging + NPD comparison	Novel (technically demanding)	Splitting shows purely limbic signature with no narrative-network activation; NPD and BPD show identical attribution patterns.

Note. Predictions 3.1–3.5 are evaluated against existing literature in Section 3; predictions 5.1–5.3 are novel predictions proposed in Section 5.

3.1 Prediction 1: Splitting as Rapid Narrative Cycling

The prediction: The idealization/devaluation cycle in BPD should function as rapid cycling between two internally consistent but mutually incompatible narratives about the same person. Each narrative—the Savior and the Scapegoat—should be subjectively compelling while active. Counter-evidence during idealization should be attributed to external factors; counter-evidence during devaluation should be attributed to the devalued person's malice or manipulation. The speed of oscillation should prevent either narrative from being maintained long enough for reality to falsify it.

Clinical literature: Story, Smith, Moutoussis, and colleagues (2024), using a computational social inference model, found that BPD participants showed significantly higher and more symmetric splitting. Critically, the study found that phases of idealization and devaluation are consolidated by attributing counter-evidence to external factors. When the other person is idealized, their imperfect behavior is attributed to circumstances. When they are devalued, their kind behavior is attributed to manipulation. This is the elastic attribution pattern the structural model identifies in non-clinical belief persistence—the same structural signature, now visible in interpersonal relationships.

Kernberg (1967, 1985, 2015) described splitting as keeping apart "good" and "bad" representations of self and other, rooted in developmental failure to integrate ambivalent experiences of the caregiver. Dammann et al. (2011) found that BPD patients described themselves predominantly as helpful and sensitive while characterizing others as selfish and self-satisfied—a finding that might appear to contradict a self-as-victim account but that is structurally consistent with it. "I give, and they take" is the victim position expressed through

moral self-elevation rather than explicit self-pity. The exculpatory self-attribution is present; it is merely expressed as virtue rather than as helplessness.

Sterna, Fuchs, and Moskalewicz (2025), in a narrative identity study, found that BPD participants described an inability to recognize themselves across time, with shifting traits, values, and memories. Life narratives were overwhelmingly dominated by negative experiences, with positive memories blurred or absent. This selective autobiographical filtering is consistent with the present analysis: the second-order constant—the generating rule that preserves the "I am the wronged one" attribution—selectively retains experiences that confirm it and degrades experiences that would challenge it.

What the model adds: Existing frameworks describe splitting as a defense mechanism (psychodynamic tradition) or a schema-level problem (cognitive tradition). The structural model offers a different account: splitting is rapid cycling between two narrative frames, each internally consistent enough to be subjectively compelling, where the speed of oscillation is not a symptom of affective instability but a structural requirement—neither narrative can survive sustained contact with the real person who occupies the agent role. The oscillation speed is necessary because the agent is a real human being who continuously generates evidence inconsistent with both assignments.

Furthermore, this structural lock is conceptually parallel to the failure of prediction-error updating in Bayesian predictive processing frameworks (e.g., Friston, 2010), where high-precision priors (the rigid narrative assignments) override contradictory sensory or interpersonal evidence. The present model grounds this domain-general predictive failure specifically in the interpersonal agent-assignment architecture.

This is a testable distinction. If splitting is primarily a cognitive inhibition deficit—as proposed by Gagnon, Quansah, Saleh, & Levin (2022), who linked splitting to difficulty

suppressing previously activated negative relational representations—then the speed of oscillation should correlate with measures of cognitive inhibition capacity and should be modifiable by interventions targeting executive function. If the structural model is correct, the speed should correlate with the degree of divergence between the real person's behavior and the current narrative assignment, and should be modifiable by reducing the impossibility of the agent assignment (i.e., by teaching the individual to assign finite, mechanism-specified functions to their partners rather than unlimited, mechanism-unspecified ones). These are different predictions, and distinguishing between them empirically is feasible.

What the model does not add: The individual observations—elastic attribution (Story et al., 2024), moral self-elevation (Dammann et al., 2011)—are not new; the structural model organizes them under a single architecture (see Table 1).

3.2 Prediction 2: The Relationship Partner as Externally Installed Agent

The prediction: Section 2.2 described the established clinical observation that BPD patients assign their partners an impossible regulatory function. The structural model converts this observation into a specific prediction: the partner's perceived capacity should expand and contract based on narrative need rather than reflecting their actual capabilities, and rejection sensitivity should be extreme, because the agent's absence collapses the entire regulatory structure—not merely the relationship but the individual's primary mechanism for emotional regulation.

Clinical literature: Gao, Assink, Cipriani, and Lin (2017), in a meta-analysis, found rejection sensitivity consistently and robustly linked to BPD across clinical and non-clinical samples. The strength of this association (pooled $r = 0.413$) is consistent with the present analysis's prediction that rejection represents not merely interpersonal loss but the collapse of the regulatory architecture itself. Stanley and Siever's (2010) neuropeptide model demonstrates

the neurochemical substrate: the partner's proximity activates opioidergic safety systems, and their absence produces a state functionally analogous to substance withdrawal.

The present analysis adds a specific structural claim to these findings: the inevitable failure of the agent assignment—which occurs whenever the partner behaves as a finite human being with their own needs and limitations—is not a failure of the partner but a structural consequence of assigning an undeliverable function. When the partner fails, the system does not conclude that the assignment was impossible. It concludes that this particular agent was inadequate and either switches to the Scapegoat narrative (devaluation) or begins scanning for a replacement agent (the pattern of rapid attachment to new idealized figures).

Novelty status: The clinical observation is not new (see Table 1). The structural model does not improve on the clinical precision of Kernberg, Gunderson and Lyons-Ruth, or Stanley and Siever. Its contribution is the cross-domain structural homology described in Section 2.2 and the prediction that maintenance and exit conditions should be shared across domains—evaluated in Prediction 5.

3.3 Prediction 3: The Cortisol-Prefrontal Impairment Pathway

The prediction: If BPD involves a chronically active version of the neurochemical loop the structural model describes, BPD should show a specific neural signature: hyperactive threat detection (amygdala), impaired evaluation and revision capacity (prefrontal cortex), and HPA axis dysregulation that makes narrative revision physiologically difficult during interpersonal stress. The system should show signs of chronic activation rather than acute-stress-driven activation, because the belief-maintenance architecture has been running since childhood.

Clinical literature: Drews, Fertuck, Koenig, Kaess, and Arntz (2019), in a meta-analysis published in *Neuroscience & Biobehavioral Reviews*, found that BPD patients show blunted cortisol reactivity to acute psychosocial stress but elevated continuous cortisol output. The

system is chronically activated but paradoxically under-responsive to new acute stressors.

The HPA axis appears to be running at or near capacity as a baseline condition.

Neuroimaging studies confirm the core BPD neural signature: amygdala hyperactivity combined with prefrontal cortex hypoactivity (Schulze, Schmahl, & Niedtfeld, 2016; Krause-Utz, Winter, Niedtfeld, & Schmahl, 2014). Nater et al. (2010) found that BPD patients showed increased subjective stress but attenuated cortisol response—the person experiences the world as maximally threatening while the physiological stress-recovery system is compromised. The developmental connection is well-established: the vast majority of BPD patients report childhood trauma (Zanarini et al., 1997), and early-life adversity produces HPA axis dysregulation.

What the model adds and what it does not: This section requires particular honesty, because the cascading chain the present analysis proposes—early trauma → HPA dysregulation → chronic cortisol → prefrontal impairment → maladaptive coping → more stress (the loop intensifies)—is not new. Linehan's (1993) biosocial model proposes that biological vulnerability interacts with invalidating environments to produce emotional dysregulation, which drives maladaptive coping that generates further crises. Developmental trauma models (e.g., Teicher & Samson, 2016) trace cascading effects from childhood adversity through neurobiological changes to adult psychopathology. The general structure of the chain—stress damages the regulatory system, the damaged system produces more stress—is established in the literature. Neither the structural model nor the present analysis discovered it.

Where the present analysis makes a specific claim that existing cascade models do not is in identifying what kind of coping fills the gap when biological regulation fails, and why that specific coping generates the interpersonal crises that restart the cycle.

Existing models describe the coping that follows prefrontal impairment in general terms: emotional dysregulation, maladaptive behaviors, impulsive actions. The structural model makes a more specific claim: when biological stress-recovery is compromised, the individual defaults to agent-based narrative construction—installing a person, force, or principle in the role of regulator—because the brain's default explanatory mode is agent-based (Kelemen, 1999; Epley et al., 2007) and this default becomes dominant when prefrontal oversight is impaired. The coping is not generically maladaptive. It is structurally specific: a person is assigned the function of emotional regulation without a mechanism for how they will deliver it.

This specificity is what closes the loop in a way existing models do not fully articulate. The biosocial model explains that dysregulation produces crises—but it does not predict why this particular kind of crisis, with this particular pattern. The structural model does: because the agent is a real human being who has been assigned an impossible function, the assignment will fail whenever the person behaves as a finite being with their own needs. Each failure generates an interpersonal catastrophe (the split). Each catastrophe activates the stress system. Each activation further impairs the prefrontal function that would be needed to recognize that the assignment itself—not the agent—is the problem. The loop's self-reinforcing character is driven not by generic dysregulation but by the structural impossibility of the agent assignment.

This is a testable distinction. If the present analysis is correct, the severity of BPD interpersonal crises should correlate with the specificity and impossibility of the regulatory function assigned to the partner—not merely with general emotional dysregulation. A patient who assigns their partner the function "be kind to me sometimes" (specific, deliverable) should show fewer and less severe crises than a patient who assigns the function "make me feel whole" (undefined, undeliverable), controlling for overall emotional dysregulation. This

prediction is not generated by the biosocial model, which does not distinguish between types of interpersonal assignment.

An operationalization challenge must be acknowledged: reliably coding whether a patient's relational expectations are specific-and-deliverable versus undefined-and-undeliverable is not straightforward. One approach would be a structured coding scheme applied to therapy session transcripts, rating each expressed relational expectation on two dimensions: specificity (can the expected behavior be concretely described?) and deliverability (could a reasonable person reliably perform it?). Alternatively, a structured interview assessing patients' relational expectations could generate scorable items. The prediction is clear in principle; the measurement requires development.

Mechanistic boundaries: An important clarification is necessary. The Drews et al. (2019) finding of blunted acute cortisol reactivity alongside elevated chronic output admits multiple mechanistic interpretations. The blunted acute response could reflect chronic sensitization (the HPA axis is overactive and responsive, but already at ceiling), chronic exhaustion (the HPA axis is degraded and unresponsive after years of overactivation), or context-dependent reactivity (the system responds selectively to attachment-specific threats but not to generic psychosocial stressors). These are different endocrine states with different implications for intervention.

The present analysis does not depend on resolving this question. What the structural model requires is a more general claim: impaired capacity for belief updating under interpersonal threat. Cortisol-mediated prefrontal impairment is one candidate pathway—and the best-evidenced one—but the structural prediction (that crises harden narratives rather than disrupting them) holds regardless of whether the impairment is driven by cortisol sensitization, cortisol exhaustion, or an alternative neurobiological mechanism entirely. The

prediction is about the functional outcome (narrative revision fails during interpersonal crisis) rather than about the specific endocrine pathway that produces it.

3.4 Prediction 4: Self-Blame as Fuel, Not Exit

The prediction: BPD self-blame—"I am toxic," "I destroy everything I touch," "I am fundamentally broken"—should function as fuel for the belief-maintenance cycle rather than as genuine attribution reversal. If the second-order constant preserves the generating rule "exculpatory self-attribution, fault to other," then self-blame should be incorporated into that attribution rather than contradicting it: the self-blame becomes evidence of how much the self suffers, further confirming the self-as-victim structure. Specifically, the proposed mechanism is: self-blame generates shame, shame generates cortisol, cortisol impairs prefrontal function, impaired prefrontal function increases dependence on the neurochemical relief that idealization of the next partner provides, and the cycle restarts.

This prediction is counterintuitive and potentially the paper's most genuinely novel contribution. It reframes self-blame not as insight (which would suggest proximity to recovery) nor as simple self-directed hostility (which would suggest failed attribution), but as a structural component of the cycle that maintains the belief system.

A clarification is needed: this is not a redescription of depressive rumination or generic shame spirals. Depressive rumination is content-focused and decelerating—the individual replays events and withdraws. BPD self-blame-as-fuel is accelerating—it generates the shame-cortisol load that drives the individual toward the next idealization target. In depression, shame produces shutdown. In BPD, shame produces desperate re-attachment: the neurochemical relief available from a new Savior agent becomes the only accessible cortisol-reduction pathway. The structural difference is that depressive rumination does not restart a specific interpersonal cycle; BPD self-blame, on the present account, does.

Clinical literature: The prediction finds indirect support from multiple sources. Gold and Kyratsous (2017) proposed an agency approach to BPD identity, emphasizing self-processing deficits rooted in impulsivity and external locus of control. The structural model reinterprets the impulsivity component: impulsive episodes—self-harm, substance use, reckless behavior—cluster at moments when the current narrative about the other person is challenged. When the partner's behavior cannot be processed through either the Savior or the Scapegoat frame, the system destabilizes. Impulsivity, on this account, is the behavioral signature of narrative-frame switching under load, not a separate process.

Sterna et al. (2025) found that BPD life narratives were overwhelmingly negative, with positive memories blurred or absent. The selective retention of negative experience is consistent with the present analysis: positive experiences that do not serve the "I am the one who suffers" attribution are not retained with fidelity, while negative experiences—including self-generated damage—are preserved and elaborated. The accumulated record of self-generated destruction becomes the evidence for "I am broken," which generates shame, which feeds the cycle. The split cycle runs: idealization → devaluation → relationship destruction → post-split clarity → genuine recognition of damage → massive shame → stress-system activation → prefrontal impairment → desperation for neurochemical relief → new idealization target appears → cycle restarts. The chronic baseline self-hatred documented in BPD may represent the accumulated residue of repeated post-split lucid intervals—after enough cycles, the shame residue becomes the resting state between active episodes.

Critically, the self-assessments are not inaccurate. The person's relational architecture does produce destructive outcomes, and they recognize this. The problem is that the recognition is descriptively accurate at a level that generates shame without generating a mechanism for change. The recognition produces stress-system activation (shame, self-punishment) without

producing skill (specific behaviors that could be practiced and changed). This is precisely why Dialectical Behavior Therapy works through skills training rather than insight alone (Linehan, 1993).

What the model adds: No existing BPD framework explicitly models self-blame as a stress-system-activating mechanism that restarts the idealization cycle—that is, as a structural component of the maintenance architecture rather than a symptom or consequence (see Table 1). The psychodynamic, cognitive, and biosocial traditions each treat self-blame differently (internalized conflict, maladaptive schema, dysregulation consequence), but none identifies it as cycle fuel.

This generates a testable prediction: self-blame intensity should correlate with proximity to a recent splitting episode and with the number of prior episodes, rather than with objective harm caused.

Operationalizing the distinction: A reviewer will reasonably object that any self-blame can be reinterpreted as exculpatory-through-suffering, rendering the second-order constant unfalsifiable. The present analysis must specify what would count as a genuine rule-level reversal versus content-level self-condemnation that preserves the generating rule.

The distinction turns on four codable features: causal specificity (does the person identify a specific behavior they performed, or a global trait?), accountability (does the person accept responsibility for the specific causal chain, or locate cause in their brokenness?), reparative intent (does the self-blame generate concrete plans to behave differently, or generate shame that fuels withdrawal or desperate re-attachment?), and partner exoneration (does the person genuinely attribute good faith to the partner's actions, or maintain that the partner also failed?).

Consider two constructed vignettes:

Example A: "I'm a terrible person, I always destroy everything, no one can love someone like me."

This is content-level self-condemnation that preserves the generating rule—the self is blamed as a type (broken, unlovable), not for specific actions, and the implicit frame remains "I am the one who suffers from my own defectiveness."

Example B: "I screamed at him when he set a reasonable boundary. He was right to be upset. I need to learn to tolerate that discomfort without attacking."

This is a candidate for rule-level reversal—specific behavior identified, partner's good faith acknowledged, reparative mechanism proposed.

If BPD patients in post-split intervals consistently produce statements of the first type and rarely of the second, the second-order constant holds. If they regularly produce the second type—genuine causal specificity with partner exoneration and reparative intent—the generating rule has broken and the formulation is falsified.

3.5 Prediction 5: Crisis Intensifies the Belief Structure

The prediction: The worst moment to attempt revision of BPD's belief-maintenance architecture is during interpersonal crisis. Stress-system activation during perceived abandonment impairs the prefrontal function required for narrative updating. Crisis should intensify commitment to the existing narrative structure rather than disrupting it.

Clinical literature: Interpersonal crises trigger the most intense BPD symptoms—suicidal behavior, self-harm, dissociative episodes, rage (Lieb, Zanarini, Schmahl, Linehan, & Bohus, 2004)—precisely when prefrontal function is most compromised. Linehan's (1993) DBT embodies this prediction clinically: distress tolerance must be built before crises, because

learning cannot occur during crisis. The therapeutic sequence—stabilize emotional reactivity first, revise narratives second—maps onto the structural model's predicted exit conditions.

What the model adds: The insight that crisis is the wrong time for learning is Linehan's, not the structural model's. What the model provides is a neurochemical vocabulary for why it works—stress-system activation impairs prefrontal function (Arnsten, 2009), prefrontal function is required for belief revision—and a cross-domain connection: people deepen commitment to failing belief structures at the point of maximum failure across clinical and non-clinical domains alike. The cross-domain connection is a genuine contribution; the clinical principle is not. The equifinality challenge this raises—why therapies with different mechanisms also work—is addressed in Section 4.

4. The Equifinality Problem

The alignment between the structural model's predicted exit sequence and DBT's therapeutic sequence is suggestive but constitutes a retrodiction, not a prediction. The retrodiction becomes problematic when confronted with a well-established finding: DBT (Linehan, 1993), MBT (Bateman & Fonagy, 2004, 2016), and Schema Therapy (Young, Klosko, & Weishaar, 2003) produce broadly comparable outcomes for BPD despite targeting different mechanisms (Storebø et al., 2020). If the model claims belief revision requires stress-system stabilization first, why do therapies that bypass this step also work?

Three interpretations are available. First, all effective therapies may share an unrecognized common factor: stress-system buffering through the therapeutic relationship itself—an attachment figure without personal stakes in the patient's relational outcomes, providing the cortisol reduction that restores prefrontal function regardless of stated technique (Hostinar, Sullivan, & Gunnar, 2014). Second, different therapies may target different stages of the belief-maintenance loop—DBT targeting the neurochemical environment, MBT targeting

agent-based attribution, Schema Therapy targeting the constant function directly—with disruption at any point eventually destabilizing the entire structure. Third, genuine equifinality: multiple independent mechanisms can disrupt the architecture, and stress-system stabilization is one pathway among several.

The paper cannot currently distinguish among these—but it can specify what evidence would. If the common-factor interpretation is primary, therapeutic alliance quality and cortisol trajectory should predict improvement more strongly than technique-adherence measures. If different-entry-points is correct, different modalities should show different early-change signatures: MBT should improve mental-state attribution before distress tolerance; DBT should improve distress tolerance before narrative flexibility. If genuine equifinality holds, early-change signatures should be uncorrelated with outcome. These are discriminating predictions testable through process-outcome research designs already common in psychotherapy research.

5. Novel Predictions

The preceding sections evaluated the model's predictions against existing literature. This section identifies predictions the model generates that existing BPD frameworks do not—predictions that, if tested, would distinguish the structural model from competing accounts rather than merely redescribing their findings.

5.1 The Splitting-Trigger Prediction

Splitting—the switch from idealization to devaluation or vice versa—should be triggered by the point of maximum divergence between the real person's behavior and the current narrative assignment. The partner's boundary-setting during an active Savior phase should trigger the switch to the Scapegoat phase. The partner's genuine kindness during an active

Scapegoat phase should trigger the switch to the Savior phase. The trigger is the behavior-narrative mismatch, not a change in the patient's internal mood state.

This is a different prediction from the affective instability account, which locates the trigger in endogenous mood fluctuation, and from the cognitive inhibition account (Gagnon et al., 2022), which locates the trigger in failure to suppress previously activated representations. The structural model locates the trigger in the relationship between external evidence and internal narrative assignment—a variable that can be measured independently of the patient's mood or cognitive capacity.

A testable design: ecological momentary assessment tracking (a) the partner's behavior (rated independently by the partner or an observer), (b) the patient's current narrative state (idealization vs. devaluation, rated in real time), and (c) the timing of switches. If the model is correct, switches should cluster at moments of maximum behavior-narrative divergence, controlling for the patient's mood state and general affective instability. Obtaining independent partner ratings is methodologically challenging; feasible proxies include partner daily diaries, timestamped text-message sentiment analysis, or therapist-coded session transcripts where the partner's behavior is described.

5.2 Self-Blame Intensity and Cycle Count

The intensity of chronic self-hatred in BPD should correlate with the number of prior splitting cycles rather than with objective measures of interpersonal harm caused. An individual who has completed many cycles—idealization, devaluation, relationship destruction, post-split recognition, shame—should show higher baseline self-hatred than an individual who has completed fewer cycles but caused equivalent interpersonal damage, because the self-hatred represents accumulated shame residue from repeated post-split lucid intervals rather than a proportional response to harm inflicted.

This is testable through retrospective relationship history combined with current self-concept measures. If the model is correct, cycle count should predict self-hatred intensity after controlling for the severity of interpersonal damage. If cycle count adds no predictive power beyond objective harm measures, this prediction fails.

5.3 Further Predictions

Two additional predictions merit brief mention. First, if splitting is fundamentally a narrative-frame switch rather than an affective shift, neuroimaging during splitting events should show activation in narrative and mentalizing networks (medial prefrontal cortex, temporoparietal junction) in addition to limbic regions—a signature distinguishable from general mood shifts. This prediction is technically demanding but would directly adjudicate between the narrative and affective accounts of splitting.

Second, the first-order/second-order constant distinction generates a differential prediction across disorders. NPD (first-order: invariant output) should show high stability in self-evaluation content across contexts but variable intensity. BPD (second-order: invariant generating rule) should show low content stability but high stability in attribution structure—exculpatory self-attribution consistently maintained regardless of surface presentation. This is testable through experience-sampling with attribution coding, and no existing framework generates this specific comparative prediction from a single architectural variable.

6. Limitations

6.1 The Redescription Risk

The most serious challenge to this paper is that the structural model may be an elaborate redescription rather than a genuine explanatory advance. If the model merely restates findings by Kernberg, Gunderson, Linehan, and Stanley and Siever in different vocabulary, it has not

advanced understanding. The paper's defense rests on the predictions in Section 5: if the splitting-trigger prediction, the cycle-count prediction, and the neuroimaging prediction prove testable and distinguishable from predictions generated by existing frameworks, the model is doing genuine theoretical work. If they prove to be reformulations, the model is redescription. The empirical tests will adjudicate.

6.2 The Second-Order Constant Concept

The first-order/second-order distinction (Section 2.1) is a theoretical innovation this paper introduces. The risk is that "constant function" has been defined broadly enough to accommodate both rigidity and instability, rendering it unfalsifiable. The defense is that the two orders generate different falsification conditions, specified in Section 2.1 and operationalized in Section 3.4. If attribution coding reveals genuine rule-level reversals—the self accepting causal responsibility while exculpating the other—the second-order formulation fails.

6.3 Scope of the Architecture

The three-component architecture proposed in Section 2 maps most directly onto BPD's interpersonal oscillation—the idealization/devaluation cycle with specific attachment figures. It does not claim to explain all dimensions of BPD phenomenology. The broader identity disturbance documented in BPD—shifting values, goals, and self-recognition across time (Sterna et al., 2025)—may involve mechanisms beyond what the three-component system describes. Impulsivity, dissociative symptoms, chronic emptiness, and self-harming behaviors each have neurobiological substrates and developmental origins that the model does not address. The paper's claims are limited to the interpersonal belief-maintenance system and should not be interpreted as a comprehensive account of BPD.

6.4 Heterogeneity

BPD is a heterogeneous diagnosis. The present analysis is expected to fit best for presentations dominated by abandonment sensitivity, relational volatility, and idealization/devaluation cycling—the phenotype Gunderson and Lyons-Ruth (2008) termed the "interpersonal hypersensitivity phenotype." It may fit poorly for presentations dominated by dissociative symptoms, where the maintenance mechanism may involve trauma-driven state-switching rather than narrative-driven agent-assignment. It may also fit poorly for presentations where impulsivity is primary and interpersonal patterns are secondary. The architecture described here is a model of one configuration within a heterogeneous diagnosis, not a universal account.

6.5 Methodological and Institutional Scope

The present model was developed outside an institutional clinical research setting. This independent positioning has theoretical benefits—freedom from the commitments of existing clinical schools—but entails objective methodological costs: limited direct access to the clinical populations, neuroimaging facilities, and longitudinal data required to test the model's predictions. The predictions in Section 5 are offered as hypotheses for researchers with the methodological resources to test them. The author's contribution is the theoretical architecture; the empirical evaluation must come from others.

7. Conclusion

This paper derived five predictions from a structural model of belief maintenance and evaluated them against existing clinical literature on Borderline Personality Disorder. The central finding is that BPD's interpersonal patterns—splitting, idealization/devaluation, abandonment sensitivity, and chronic self-blame—are consistent with a specific belief-maintenance architecture: an invariant attribution structure (exculpatory self-attribution,

responsibility to other) maintained by rapid oscillation between two narrative frames about a person who has been assigned a regulatory function no human being can reliably deliver.

The model's contributions are uneven. Some predictions—particularly the reframing of self-blame as a cortisol-generating restart mechanism (Section 3.4) and the splitting-trigger prediction (Section 5.1)—appear to represent genuine theoretical work that existing frameworks do not perform. Other predictions—particularly the DBT alignment (Section 3.5) and the agent-in-the-other formulation (Section 3.2)—represent translation of established clinical insights into the model's vocabulary rather than novel discovery. The paper has attempted to be transparent about which category each contribution falls into.

The equifinality problem (Section 4) is unresolved and represents the most serious challenge to the model's treatment-mechanism claims. The limitations are substantial (Section 6), beginning with the fact that the underlying structural model has not itself been empirically validated.

The intellectual value of the exercise lies not in the conclusions but in the predictions—specific, falsifiable, and derivable from a domain-general architecture rather than disorder-specific theorizing. If the splitting-trigger prediction holds, it suggests that splitting is a response to behavior-narrative mismatch rather than endogenous mood fluctuation. If the cycle-count prediction holds, it reframes chronic self-hatred as accumulated shame residue rather than proportional response to harm. If the neuroimaging prediction holds, it establishes splitting as a narrative event rather than a purely affective one. Each prediction can fail independently, and their failure would specify exactly how the model must be revised.

The diagnostic question the model proposes remains available for clinical application: when a person assigns someone an impossible function—regulate me, complete me, save me—without specifying a mechanism by which any human being could accomplish this, the

structural preconditions for the cycle described in this paper are in place. The question is not whether the person assigned to that role is good enough. The question is whether any person could be.

References

- Arnsten, A. F. T. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience*, 10(6), 410–422.
<https://doi.org/10.1038/nrn2648>
- Bahtışen, G. (2026a). Secondary Universe Theory: A Proposed Integrative Framework for Narrative Belief Construction, Neurochemical Maintenance, and Resistance to Correction. *SSRN Electronic Journal*. <https://papers.ssrn.com/abstract=6276259>
- Bateman, A., & Fonagy, P. (2004). *Psychotherapy for Borderline Personality Disorder: Mentalization-Based Treatment*. Oxford University Press.
- Bateman, A., & Fonagy, P. (2016). *Mentalization-Based Treatment for Personality Disorders: A Practical Guide*. Oxford University Press.
- Dammann, G., Hügli, C., Selinger, J., Gremaud-Heitz, D., Sollberger, D., Wiesbeck, G. A., Küchenhoff, J., & Walter, M. (2011). The self-image in borderline personality disorder: An in-depth qualitative research study. *Journal of Personality Disorders*, 25(4), 517–527. <https://doi.org/10.1521/pedi.2011.25.4.517>
- Drews, E., Fertuck, E. A., Koenig, J., Kaess, M., & Arntz, A. (2019). Hypothalamic-pituitary-adrenal axis functioning in borderline personality disorder: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, 96, 316–334.
<https://doi.org/10.1016/j.neubiorev.2018.11.008>

- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.
<https://doi.org/10.1037/0033-295X.114.4.864>
- Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Gagnon, J., Quansah, J. E., Saleh, G., & Levin, C. (2022). Is splitting related to resistance to proactive interference? A process-oriented study of Kernberg's conceptualization of splitting. *Psychopathology*, 55(6), 345–361. <https://doi.org/10.1159/000525006>
- Gao, S., Assink, M., Cipriani, A., & Lin, K. (2017). Associations between rejection sensitivity and mental health outcomes: A meta-analytic review. *Clinical Psychology Review*, 57, 59–74. <https://doi.org/10.1016/j.cpr.2017.08.007>
- Gold, N., & Kyratsous, M. (2017). Self and identity in borderline personality disorder: Agency and mental time travel. *Journal of Evaluation in Clinical Practice*, 23(5), 1020–1028. <https://doi.org/10.1111/jep.12769>
- Gunderson, J. G., & Lyons-Ruth, K. (2008). BPD's interpersonal hypersensitivity phenotype: A gene-environment-developmental model. *Journal of Personality Disorders*, 22(1), 22–41. <https://doi.org/10.1521/pedi.2008.22.1.22>
- Hostinar, C. E., Sullivan, R. M., & Gunnar, M. R. (2014). Psychobiological mechanisms underlying the social buffering of the hypothalamic-pituitary-adrenocortical axis: A review of animal models and human studies across development. *Psychological Bulletin*, 140(1), 256–282. <https://doi.org/10.1037/a0032671>
- Kelemen, D. (1999). Why are rocks pointy? Children's preference for teleological explanations of the natural world. *Developmental Psychology*, 35(6), 1440–1452.
<https://doi.org/10.1037/0012-1649.35.6.1440>

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological Science*, 15(5), 295–301. <https://doi.org/10.1111/j.0956-7976.2004.00672.x>

Kernberg, O. F. (1967). Borderline personality organization. *Journal of the American Psychoanalytic Association*, 15(3), 641–685.
<https://doi.org/10.1177/000306516701500309>

Kernberg, O. F. (1985). *Borderline Conditions and Pathological Narcissism*. Jason Aronson.

Kernberg, O. F. (2015). Neurobiological correlates of object relations theory: The relationship between neurobiological and psychodynamic development. *International Forum of Psychoanalysis*, 24(1), 38–46.
<https://doi.org/10.1080/0803706X.2014.965005>

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>

Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, 364(9432), 453–461. [https://doi.org/10.1016/S0140-6736\(04\)16770-6](https://doi.org/10.1016/S0140-6736(04)16770-6)

Linehan, M. M. (1993). *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. Guilford Press.

Nater, U. M., Bohus, M., Abbruzzese, E., Ditzen, B., Gaab, J., Kleindienst, N., Ebner-Priemer, U. W., Mauchnik, J., & Ehlert, U. (2010). Increased psychological and attenuated cortisol and alpha-amylase responses to acute psychosocial stress in female patients with borderline personality disorder. *Psychoneuroendocrinology*, 35(10), 1565–1572. <https://doi.org/10.1016/j.psyneuen.2010.06.002>

- 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>
- Stanley, B., & Siever, L. J. (2010). The interpersonal dimension of borderline personality disorder: Toward a neuropeptide model. *American Journal of Psychiatry*, 167(1), 24–39. <https://doi.org/10.1176/appi.ajp.2009.09050744>
- Sterna, A., Fuchs, T., & Moskalewicz, M. (2025). The sense of self and interpersonal functioning in borderline personality disorder: Toward qualitative evidence-based phenomenological conceptualization. *Qualitative Health Research*. <https://doi.org/10.1177/10497323251376224>
- Storebø, O. J., Stoffers-Winterling, J. M., Völlm, B. A., Kongerslev, M. T., Mattivi, J. T., Jørgensen, M. S., Faltinsen, E., Todorovac, A., Sales, C. P., Callesen, H. E., Lieb, K., & Simonsen, E. (2020). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews*, 5, CD012955. <https://doi.org/10.1002/14651858.CD012955.pub2>
- Story, G. W., Smith, R., Moutoussis, M., Berwian, I. M., Nolte, T., Bilek, E., Siegel, J. Z., & Dolan, R. J. (2024). A social inference model of idealization and devaluation. *Psychological Review*, 131(3), 749–780. <https://doi.org/10.1037/rev0000430>
- Teicher, M. H., & Samson, J. A. (2016). Annual research review: Enduring neurobiological effects of childhood abuse and neglect. *Journal of Child Psychology and Psychiatry*, 57(3), 241–266. <https://doi.org/10.1111/jcpp.12507>
- Young, J. E., Klosko, J. S., & Weishaar, M. E. (2003). *Schema Therapy: A Practitioner's Guide*. Guilford Press.

Zanarini, M. C., Williams, A. A., Lewis, R. E., Reich, R. B., Vera, S. C., Marino, M. F.,
Levin, A., Yong, L., & Frankenburg, F. R. (1997). Reported pathological childhood
experiences associated with the development of borderline personality disorder.
American Journal of Psychiatry, 154(8), 1101–1106.
<https://doi.org/10.1176/ajp.154.8.1101>

¹ While the parent model is currently available as an independent preprint, the present paper is designed to be evaluable entirely on the merits of its own BPD-specific predictions. The validity of these predictions and their alignment with clinical evidence can be assessed independently of the parent theory's formal publication status.