Journal of Psychopathology and Clinical Science

A Generative Model of Personality Disorder as a Relational Disorder

Orestis Zavlis, Michael Moutoussis, Peter Fonagy, and Giles Story

Online First Publication, July 24, 2025. https://dx.doi.org/10.1037/abn0001010

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Psychopathology and Clinical Science

ISSN: 2769-7541

https://doi.org/10.1037/abn0001010

A Generative Model of Personality Disorder as a Relational Disorder

Orestis Zavlis[1], Michael Moutoussis[2], Peter Fonagy[1], and Giles Story[3]

[1] Unit of Psychoanalysis, Department of Psychology and Language Sciences, University College London

[2] Department of Imaging Neuroscience, Institute of Neurology, University College London

[3] Department of Imaging Neuroscience, Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London

Existing models of personality disorder are statistical models: dispersion patterns of personality facets. Although useful in describing personality differences, such models fall short in terms of explaining those differences. Generative models can address these explanatory gaps by explicating the mechanisms that generate descriptive pathologies. In this article, we aim to move beyond the former descriptive models and toward the latter explanatory ones. To do so, we formalize personality pathology using a generative model with four key properties. First, it is probabilistic: it outlines how humans leverage uncertainty to make sense of their own and others' ways of being. Second, it is relational: it posits that personality pathology is about poor ways of experiencing and relating to the self and others. Third, it is hierarchical: it accounts for the multiplicity of self- and other-states (in the here-and-now) and traits (in the long run). Finally, it is dynamic: it outlines how these properties evolve over time, accounting for the development of personality. By simulating data from this model, we demonstrate how it can account for the generation, maintenance, and treatment of various personality problems (from borderline instability to narcissistic grandiosity) by formalizing them as relational problems: problems with navigating relationships. We thus discuss how our model could be used to address recent debates on what is central to personality pathology by clarifying the distinction between description (what personality "is") and explanation (what personality "does"). We conclude the article with a tutorial on our model and suggestions for future research.

General Scientific Summary

This study suggests that personality disorder can be more formally defined as a relational disorder because it is mainly about how patients understand and relate to themselves as well as others.

Keywords: generative model, mentalizing, relationships, personality pathology

Supplemental materials: https://doi.org/10.1037/abn0001010.supp

> Rarely and under certain exceptional conditions is individual psychology in a position to disregard relations to others. In the individual's mental life someone else is invariably involved: as a model, as an object, as a helper, as an opponent. So, from the very first individual psychology is at the same time social psychology as well. (S. Freud, 1921, p. 69)

Christopher J. Hopwood served as action editor.

Orestis Zavlis https://orcid.org/0000-0002-3985-6346

Orestis Zavlis served as lead for conceptualization, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing–original draft, and writing–review and editing. Michael Moutoussis served as lead for supervision and served in a supporting role for conceptualization, investigation, methodology, project administration, and writing–review and editing. Peter Fonagy served as lead for supervision and served in a supporting role for conceptualization, investigation, project administration, validation, and writing–review and editing. Giles Story served as lead for software and supervision, contributed equally to methodology, and served in a supporting role for conceptualization, formal analysis, investigation, project administration, validation, and writing–review and editing.

The data are available at https://github.com/OrestisZavlis/RelationalDisorder.

The experimental materials are available at https://github.com/OrestisZavlis/RelationalDisorder.

Correspondence concerning this article should be addressed to Orestis Zavlis, Unit of Psychoanalysis, Department of Psychology and Language Sciences, University College London, Office 512, 1–19 Torrington Place, London WC1E 7HB, United Kingdom. Email: orestis.zavlis.23@ucl.ac.uk

We have come a long way since the old categorical days of personality pathologies. Stigmatizing categories like “hysteric” and “schizoid” personalities have been replaced with continuous traits, upon discovering that problems of “personality” are not typologies that only some people have but dimensions on which every human can be ranked (Hopwood et al., 2018). Likewise, rigorous statistical analyses have established that what is central to personality pathology is what has been theorized for centuries: ways of relating to oneself and others (Berrios, 1993; Hopwood et al., 2013; Pincus et al., 2020; Wright et al., 2022; Zavlis, 2024b). Finally, recent perspectives have highlighted that these problems cannot be merely understood at the patient level (“self-in-relation-to-others”) but need to also be considered at the societal level (“others-in-relation-to-self”; Rodriguez-Seijas et al., 2023). In light of these advances, therefore, our field has almost certainly moved forward.

At the same time, however, our field faces a risk of stagnating methodologically—and, by extension, as we argue, theoretically. Historically, research on personality psychopathology has been heavily atheoretical and data-driven in nature (John & Srivastava, 1999). Such research has focused on factor-analytic methodologies as a way of exploring statistical regularities of “socially undesirable” adjectives (see Coker et al., 2002 for an early analysis). To be sure, these methodologies are descriptively invaluable, because they have allowed us to distill those statistical regularities into elegant descriptive hierarchies (such as the five-factor structure and its maladaptive variant; Widiger & Costa, 2012). At the same time, however, these methodologies are limited theoretically, because they cannot reveal the mechanisms that underlie personality and its putative pathology (Cervone, 2005; Mischel, 2004; Revelle, 1995; Wright & Hopwood, 2016).

To illustrate, consider the construct “personality functioning,” which has been developed with the aim of providing a more explanatory, rather than a descriptive, way of viewing personality pathology by casting it as an “intrapsychic system that *drives* trait manifestation” (Sharp, 2022, p. 317; Sharp & Wall, 2021). Although theoretically notable and consistent with rich clinical work, this conception of personality functioning has yet to be realized empirically. Indeed, at the time of writing, personality functioning is typically operationalized as a unidimensional factor model that comprises items relating to identity, agency, intimacy, and empathy (Sharp & Wall, 2021). Moreover, with a few notable exceptions (e.g., Haehner et al., 2024; Kerber et al., 2024; Ringwald et al., 2021; Roche, 2018), this construct has only been examined cross-sectionally and with methods that preclude the aforementioned dynamic interpretation (Hopwood, 2025). This mismatch between theory and methods likely explains the persistence of recent debates around the usefulness of “personality functioning” over “personality traits” in defining personality pathology: the two constructs are indistinguishable at the descriptive level, implying that any putative theoretical differences cannot be revealed by descriptive models (see Morey, 2019; Zavlis & Fonagy, 2025).

Importantly, these methodological problems are not constrained to the recent construct of personality functioning. Instead, they remarkably extend to various other constructs that have been devised throughout the years to explain, rather than simply describe, personality pathology. Notable examples here entail mentalizing, attachment styles, defense mechanisms, and schematic self–other beliefs (see Bender et al., 2013). Although these concepts are rich theoretically and clinically, they tend to be poor empirically because they are almost always reduced to factor models that cannot adequately capture their nuances (e.g., their context-driven or dynamic aspects). As a consequence, these constructs are at risk of stagnating because they cannot, in their current descriptive and static forms, produce evidence that uniquely supports their more intricate causal assumptions (see Fried, 2020; Haslbeck et al., 2022; Robinaugh et al., 2021 for commentaries on this problem).

Our argument in this article is that much of the stagnation in the field of personality psychopathology can be traced back to our reliance on data-driven statistical approaches that cannot assess the more intricate theoretical conjectures that we want them to assess. To elaborate, statistical models are descriptive models because their data-generating processes are not specific enough to provide deep theoretical insights (Lewandowsky & Farrell, 2011). Consider, for example, machine learning models: although these models are useful in terms of predicting new observations from past observations, they remain agnostic as to the processes that causally underpin their predicted responses (i.e., black-box problem; Bishop & Nasrabadi, 2006). The same limitation applies to other data-driven models like factor models and network models: Although these models are invaluable in terms of revealing robust phenomena in psychological data (e.g., the phenomenon that borderline symptoms are the strongest reflections of personality pathology; Sharp et al., 2015), they cannot illuminate the mechanisms that explain these phenomena (i.e., why is borderline personality so central to personality psychopathology?).

Generative models could address these explanatory problems. Generative models are so-called because they are specifically handcrafted to instantiate unique data-generating processes that are theoretically meaningful (Lewandowsky & Farrell, 2011). For example, reinforcement learning models can be used to instantiate the hypothesis that emotions arise from mismatches between what people “expect” and what they “get” in life (Emanuel & Eldar, 2022). Likewise, probabilistic models can be used to examine whether people rely more on their prior beliefs or unfolding observations when explaining social behavior (Barnby et al., 2023). Such generative models advance traditional statistical approaches by not simply describing psychopathologies, but rather by elucidating the (within-person) mechanisms that might generate those psychopathologies (Adams et al., 2015; Friston et al., 2014; Huys et al., 2016; Montague et al., 2012; Zavlis, 2024a).

In this article, we aim to move beyond the former descriptive models of personality problems and toward the latter generative models that might explain those problems (Zavlis & Fonagy, 2025). To do so, we formalize personality pathology using a generative model that has four key properties: it is probabilistic, relational, hierarchical, and dynamic. Each of these properties was installed in our modeling framework based on substantive theory (see next section). Our generative model was thus handcrafted with the unique aim of explicating one possible mechanism by which personality pathology emerges, persists, and ultimately ameliorates. In the next section, we outline this data-generating mechanism before turning to the explanations and predictions it furnishes.

Generative Framework

In this section, we outline the four key aspects of our generative model (i.e., probabilistic, mentalizing, hierarchical, and dynamic aspects), as well as the theoretical reasoning that underpins them. In doing so, we focus on a verbal exposition of our model to facilitate its dissemination. For a more detailed exposition (including mathematical and coding details), readers are referred to the online supplemental materials, GitHub repository (https://github.com/OrestisZavlis/RelationalDisorder), and previous publications on this topic (see Moutoussis et al., 2018; Moutoussis, Fearon, et al., 2014; Moutoussis, Trujillo-Barreto, et al., 2014; Story et al., 2024).

Probabilistic Inference

The first axiom of our framework is that human beings are meaning-making beings: they continually strive to make sense of the world by (consciously or unconsciously) inferring the hidden causes of human behavior. This axiom has its origins in ancient philosophies (like Kant who emphasized that perception is fundamentally imaginative; Clark, 2013) and was recently embraced by computational neuroscientists who suggest that humans perceive the world in an active, not passive, manner (Friston, 2010; Friston et al., 2014). From this perspective, humans are not passive receivers of sensory information (such as the visual observation that your partner is not replying to your texts). Instead, humans are active constructors of the meaning behind those observations (e.g., that your partner might be busy chatting with others). In that sense, human beings do not wander through life with no fantasies of meaning; instead, human beings are creatures of meaning: they continually strive to make sense of the world by fantasizing about the hidden causes of everyday observations.

Conceptually, this process of fantasizing or meaning-making is known under various terms, including the "appraisal," "construal," "evaluation," or "perception" of live events (Arnold, 1960; Friston et al., 2012, 2014; Heider, 1983; Lazarus, 1966; Rauthmann et al., 2014). Although differing slightly in their theoretical nuances, these concepts converge in suggesting that "raw" and objectively "meaningless" sensorial experiences gain meaning (i.e., cognitive and emotional qualities) once they are filtered through a person's subjective belief system (Zavlis, Bentall, et al., 2024). For example, the "raw" and otherwise "meaningless" observation that a therapist checks the clock may gain a completely different meaning by two patients who have different belief systems: while one patient may perceive it as an act of politeness ("My therapist is considerate and aims to ensure that our session does not run over to the next session"), another may construe it as an act of rudeness ("My therapist is bored of me and wishes the session to be soon complete"). Throughout the article, we will refer to this meaning-making process as the "construal" of life events.

Formally, this construal process can be cast as a form of probabilistic inference: that is, the process of combining prior beliefs with empirical observations in order to make sense of the world. This form of inference is predicated on beliefs: that is, memory units that may be present from birth (i.e., "temperament") or acquired later (i.e., "learning"; Smith et al., 2022). From a probabilistic perspective, beliefs are probability distributions that outline states of the world—for example, a paranoid belief will assign a high probability to a threatening state of the world (Adams et al., 2013; Barnby et al., 2023). Such probabilistic beliefs can be conscious (explicit) or unconscious (implicit) and endorsed with a certain level of precision (i.e., conviction) that is reflected in a distribution's variation (Adams et al., 2015). Specifically, when the distribution is highly concentrated around the most likely value (i.e., mean or expectation), then it reflects a more confident, rigid, and emotional belief (e.g., "I am sure I am not interesting"; Figure 1B). Conversely, when the distribution is highly dispersed (i.e., many states of the world are relatively equally likely), then it reflects a more uncertain, flexible, and apathetic belief (e.g., "I may or may not be interesting"; Figure 1A).

Bayes theorem outlines how such beliefs combine with empirical evidence to result in particular construals of life events:

$$P(\text{cause}|\text{observation}) \propto P(\text{cause})P(\text{observation}|\text{cause}). \qquad (1)$$

In simple terms, Bayes theorem defines the construal of an observation (e.g., $o =$ "therapist checking the clock") as a "posterior" belief, $P(c|o)$: the probability that a given causal factor, c, was responsible for that observation. This posterior belief is predicated on two terms: one's prior expectation of a causal factor ($P(c)$, "I think I am not interesting") and whether that expectation fits with their current sensorial observation ($P(o|c)$, "How likely is it that my therapist checked the clock, assuming that I am not interesting?"). Bayesian inference combines these two pieces of information (known as the prior and likelihood, respectively) to yield a posterior belief: an updated estimate of the state of the world in light of your most recent social observation: $P(c = \text{I am boring} \mid o = \textit{therapist};$ checks clock). Importantly, Bayesian inference weighs prior beliefs and incoming information according to their precision (π) (inverse of variance) when computing the posterior.[1] If a prior is more precise relative to incoming information, then the posterior will be biased toward that prior (i.e., a person will be "stubborn" and unlikely to change their mind that they are not interesting; Figure 1B). Conversely, if a prior is less precise relative to incoming information, then the posterior will shift to accommodate incoming information (i.e., a person will be "open" to changing their mind that they are not interesting; Figure 1C).

How are these belief-updating processes relevant to personality disorder? We argue that they are relevant because they can be leveraged to explain how personality disorders emerge and persist because of maladaptive beliefs (for related belief-based models in other mental disorders, see Adams et al., 2013; Fradkin et al., 2020; Maisto et al., 2021). For instance, as we outline later, rigidity in interpersonal beliefs may explain inflexible behavior whereas plasticity in intrapersonal beliefs may explain instability in self-perception. Before we turn to these patterns, however, the precise architecture of our generative model must first be explicated. We turn to this architecture in the next section where we outline how our modeling framework places a key process at the heart of personality problems: the inability to construe (or "mentalize") the self and others.

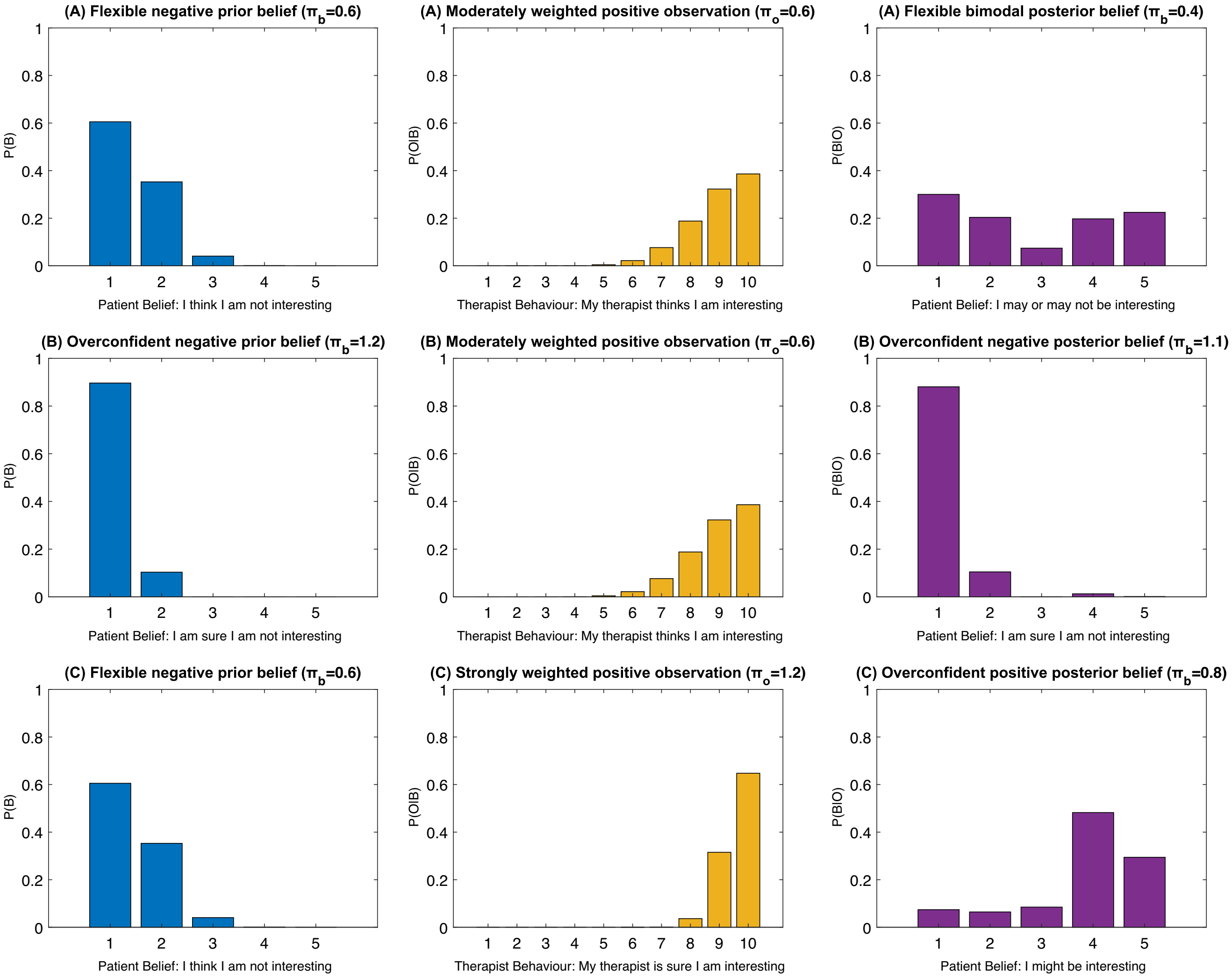
Mental Inference

To elaborate, the second axiom of our framework specifies the first by suggesting that personality and its pathology are specifically about how humans construe themselves and others as intentional agents. This focus on self versus other is predicated on a wealth of evidence, suggesting that personality functioning is typified by two dialectic capacities: the capacity to establish an integrated and worthwhile relationship with oneself and the capacity to establish meaningful and long-lasting relationships with others (see Bender et al., 2013; Blatt, 2008; Fonagy et al., 2018; Hopwood, 2018; Mikulincer & Shaver, 2010; Ryan et al., 2015; Wright et al.,

[1] This applies to unimodal normally distributed beliefs.

Figure 1

Examples of Belief Updating Using the Scenario of a Patient Who Believes That They Are Not Interesting but Encounters a Therapist Who Reassures Them That They Are Interesting



Note. (A) The patient updates their belief in a flexible manner because they equally weigh ($\pi_c = \pi_o = 0.6$) their prior ("I think I am not interesting") and their social observation ("My therapist thinks I am interesting") to create a flexible belief suggesting that they might be interesting (see ratings 1–2 and 4–5). (B) The patient finds it difficult to update their belief since they weigh their prior much more than their social observation ($\pi_c = 1.2 > \pi_o = 0.6$), yielding an overconfident posterior belief that they are not at all interesting (see ratings 1 and 2). (C) The patient updates their belief substantially since they weigh their positive observation much more than their prior expectation ($\pi_c = 0.6 < \pi_o = 1.2$), yielding an overconfident posterior belief that they might be interesting (see ratings 4 and 5). See the online article for the color version of this figure.

2023; Zavlis, 2024b). These motives are known under various terms, including autonomy versus surrender (Angyal, 1951), agency versus communion (Bakan, 1966), dominance versus warmth (Pincus, 2005), and power versus love (S. Freud, 1930). Importantly, these "meta" ideas are not specific to isolated theories but rather span various academic disciplines (from philosophy to anthropology and neuroscience), suggesting that this self–other dialectic offers a truly pantheoretical perspective for understanding personality and its pathology (see Luyten & Blatt, 2011 for a comprehensive review).

Our model builds upon this pantheoretical perspective by explicating one of its key capacities: namely, the capacity to "mentalize" which can be defined as the capacity to view oneself and others in intentional terms (i.e., in terms of "thoughts, feelings, and wishes"; Fonagy, 1991; Fonagy & Luyten, 2018; Fonagy et al., 2018). This axiom of our model is predicated on a wealth of developmental, clinical, and neuroscientific evidence suggesting that personality disorders are typified by maladaptive ways of experiencing the self and others (see Fonagy & Luyten, 2018; Hopwood, 2018; Hopwood et al., 2013; Luyten & Fonagy, 2015; Luyten, Campbell, & Fonagy, 2020; Moutoussis, Fearon, et al., 2014; Moutoussis, Trujillo-Barreto, et al., 2014; Pincus, 2005; Story et al., 2024; Wright et al., 2022, 2023; Zavlis, 2024b). Moreover, this axiom is consistent with recent diagnostic approaches that have re-conceptualized dimensional personality disorder primarily in terms of "underlying structural impairments" (i.e., "beliefs about the self and others") rather than in terms of symptomatic consequences (such as "maladaptive behaviors"; Hutsebaut & Bender, 2024, p. 412). Our model embraces these approaches and suggests that the main problems that patients with personality diagnoses experience arise from impairments in one key personality process: the process of mentalizing.

Mentalizing here is formalized as a probabilistic inference that explains social behaviors, $o_1 = [1, 10]$, based on two mental states: the mental states of the self, s_1, and the mental states of others, s_2: $P(o_1|s_1, s_2)$ (Figure 2). Mental states take five values that denote the benevolence of one's intentions: $s = \{$extremely bad, bad, neutral, good, extremely good$\}$. Our model formalizes these intentional states as hidden causes that map onto behavior with a certain likelihood: for instance, if my friend behaves in a certain way, $o_2 =$ being rude, then there is an increased likelihood that they have a bad intention toward me: $P(o_2 = \text{rude}|s_2 = \text{bad intention})$. Mentalizing, in this sense, can be cast as a process of probabilistic inference: a process of working backward to infer the mental states that are most likely to have caused one's observed behavior.

During this probabilistic inference, several things can go awry, resulting in ways of experiencing the self and others that are pathognomonic of personality disorders. For example, a tendency to focus on others and explain their behavior in predominantly negative terms may lead to a paranoid personality pattern (McWilliams, 2011, p. 214). Conversely, a tendency to direct this negativity to oneself may lead to a depressive personality pattern: the "guilty" or "self-defeating" self (McWilliams, 2011, p. 267). In that sense, as we argue later, personality disorders are fundamentally relational disorders: maladaptive ways of experiencing and relating to oneself and others.

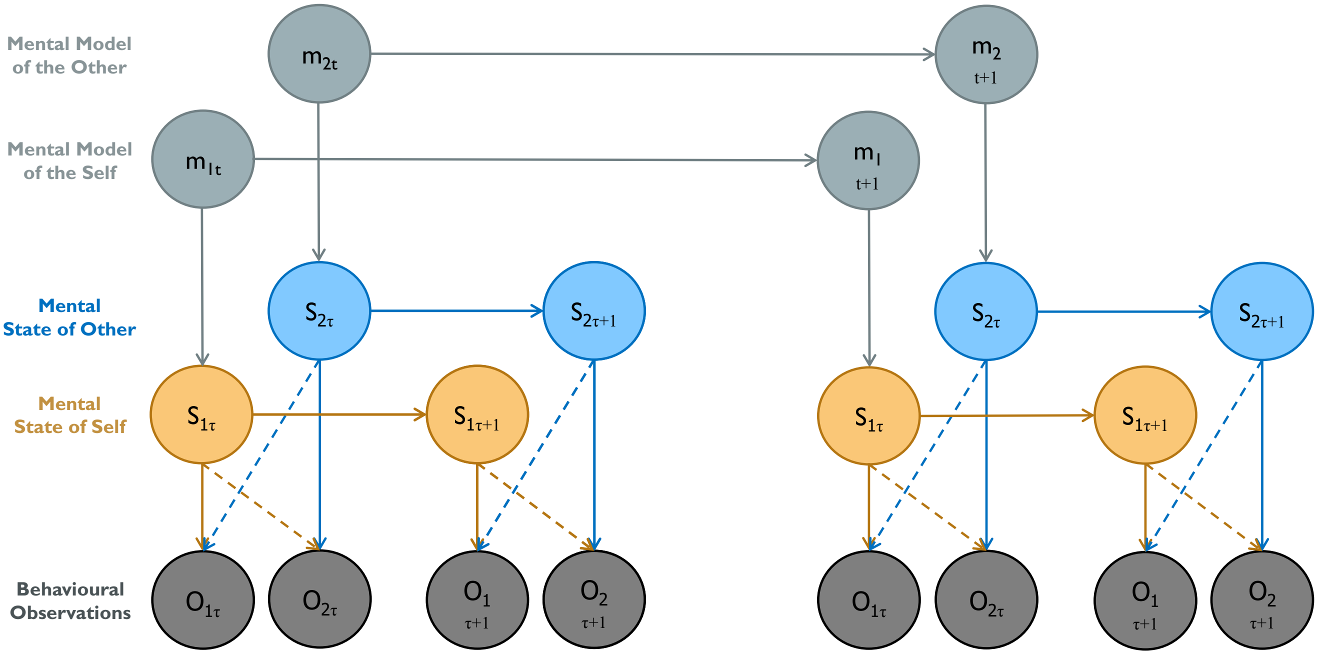
Although intuitively simple when cast in this manner, a key question that arises when it comes to these mentalizing problems is: where do their maladaptive mental states come from? In the next section, we address this question by arguing that mental states come from internal working models.

Hierarchical Inference

To elaborate, the third key aspect of our model posits that mental states are rooted in deeply held beliefs about ourselves and others. This axiom is predicated on a large corpus of developmental evidence suggesting that the human capacity to fantasize about disparate mental states is based on mental concepts that were acquired during early relational experiences (see Fogel, 1993; Fonagy et al., 2018; Luyten et al., 2020; Luyten, Campbell, & Fonagy, 2020). Importantly, this early psychoanalytic perspective is further supported by an independent line of neuroscientific research which has illustrated that these concepts emerge from the integration of multimodal sensory information (e.g., bodily and affective cues from caregivers) that culminate a brain network (known as the "default-mode" or "mentalizing" network) that is responsible for the re-presentation of both the self and others as agents (e.g., see Atzil et al., 2018; Barrett & Satpute, 2013; Buckner et al., 2008;

Figure 2

Generative Model of Personality Disorder as a Relational Disorder



Note. Under this model, a personality disorder is generated from maladaptive mental inferences: that is, maladaptive ways of combining internal information about the self and others (i.e., internal mental models, m_t) with external information about the self and others (i.e., external behavioral evidence, o_τ) to infer the mental states of the self and others, s_τ. Personality disorder is therefore conceptualized as a dynamic relational disorder because it continually engages in this dialectic process of self-relatedness (i.e., self-mentalizing) and other-relatedness (i.e., other-mentalizing). See the online article for the color version of this figure.

Cozolino, 2014; Fotopoulou & Tsakiris, 2017; Saxbe et al., 2020; Siegel, 2012). Together, this large corpus of evidence suggests that our very own sense of self and otherness is intimately predicated on mental concepts that were acquired during early social development.

In psychology, these mental concepts are known under various terms, including internal working models (Bowlby, 1969), mental representations of the self and others (Klein, 1930), relational schemas (Baldwin, 1992), nuclear scripts (Tomkins, 1978), and archetypes (Jung, 1919). Different theorists emphasize different functional characteristics of these mental concepts, with some noting that they encode expectations of reliability in others and worth in oneself (Mikulincer & Shaver, 2010), others emphasizing that they distill interpersonal motives of affiliation (hostility vs. friendliness) and control (dominance vs. submissiveness; Wright et al., 2023), and yet others suggesting that they represent universal patterns of human perception and behavior (e.g., archetypal patterns of "mistrust"; Knox, 2003). Despite differing in their nuances, these ideas converge in suggesting that mental concepts encode an internal representation of the "self-in-relation-to-others" (Bender et al., 2013, p. 332) which drives humans to experience themselves and others in particular manners.

Here, we focus on one such experience: namely, the view of the self as a "worthy agent" and the view of the other as a "trustworthy agent." Although rather simple (see the online supplemental materials for a more complex representation), this unidimensional conception of the self and other is consistent with the mentalizing theory of psychopathology which emphasizes that self- and other-mentalizing are fundamentally based on mental concepts relating to (personal) "worthiness" and (epistemic) "trustworthiness" (Fonagy & Luyten, 2009; Fonagy et al., 2015, 2019).

Formally, these mental concepts can be understood as hyper-priors: that is, beliefs about beliefs. Technically, this implies that mental states are generated from mental concepts (hereon referred to as internal working models): $P(s_1|m_1)$. Thus, internal working models of the self and others (m_1 and m_2, respectively) generate mental states of the self and others (s_1 and s_2, respectively) which in turn generate observed behaviors from the self and others (o_1 and o_2, respectively; see Figure 2). This generative process is known as "hierarchical inference" because it outlines how higher-order beliefs (here, mental concepts) encode deeper (sometimes, unconscious) expectations about lower-order beliefs (here, mental states).

In this article, we explore five such expectations, that is, working models of the world: $m = \{\text{integrated, bad, good, split, empty}\}$. Each working model here encompasses a probability distribution that encodes an expectation of mental states in the self and others. First, an "integrated" working model encodes the nuanced and flexible expectation that, in general, the self is "worthy" and others are "trustworthy": $P(s_{\tau=0}|m_t = \text{integrated}) \sim \mathcal{N}(s;\ \mu = 3,\ \sigma^2 = 1/\pi_s = 1)$. Second, a "bad" working model encodes the extreme and rigid expectation that the self is "unworthy" and others are "untrustworthy": $P(s_{\tau=0}|m_t = \text{bad}) \sim \mathcal{N}(s;\ \mu = 1,\ \sigma^2 = 1/\pi_s = 1/1.4)$. Third, a "good" working model encodes the extreme and rigid expectation that the self is "exceptional" and others are "ideal": $P(s_{\tau=0}|m_t = \text{bad}) \sim \mathcal{N}(s;\ \mu = 5,\ \sigma^2 = 1/\pi_s = 1/1.4)$. Fourth, a "split" working model encodes the extreme and rigid expectation that the self is either "worthless" or "exceptional" and others are either "untrustworthy" or "ideal" (i.e., combined good and bad mental models). Finally, an "empty" working model encodes no expectation of the self or the other: that is, under this model, all mental states are equally likely: $P(s_{\tau=0}|m_t = \text{empty}) = 1/N = 1/5$.

Our argument in this article is that these working model distributions can explain the maladaptive ways by which people with personality diagnoses experience themselves and others (for instance, by promoting overly rigid or overly plastic ways of viewing the self and others; Story et al., 2024). To fully understand these patterns, however, their development over time must also be explicated. We turn to this temporal aspect of our model in the next and final section.

Dynamic Inference

The final key aspect of our model suggests that personality disorders are not static over time (as traditionally assumed) but rather dynamic: that is, they shift across time and space. Theoretically, this dynamic view of personality is based on contemporary network and cybernetic perspectives that conceptualize personality as a self-regulating system that continually transacts with the environment to convert sensorial inputs (e.g., social information) into personality outputs (e.g., social behavior; see Cramer et al., 2012; DeYoung, 2015; Read & Miller, 2002; Read et al., 2010; Safron & DeYoung, 2021; Shoda, 2007; Shoda et al., 2002). Empirically, this enactive view of personality is supported by longitudinal evidence showing that personality (pathology) is characterized by both personality structures (i.e., static traits that describe personality) and personality functions (i.e., dynamic mechanisms that explain how personality manifests across time and space; Wright & Hopwood, 2016; Wright & Kaurin, 2020). Our model is consistent with this evidence and suggests that one key mechanism of personality pathology (namely, inferences of the mind) is similarly dynamic.

Specifically, our model suggests that mental inferences are dynamic because they can change not only in the short term (i.e., mental states shifting dynamically in the here-and-now) but also in the long term (i.e., mental models shifting over longer periods of time; see Story et al., 2024). Evidence for the short-timescale updating comes from work on situational dynamics showing that personality processes (like mental inferences) can be adaptive in particular contexts (such as under supportive and cooperative relational contexts), but abruptly become maladaptive in other contexts (such as stressful or conflictual relational contexts; Rauthmann & Sherman, 2019; Rauthmann et al., 2014). Evidence for the long-timescale updating comes from work at the interface of social, clinical, and personality psychology showing that personality dispositions (from temperamental traits to working models) can shift in response to social information to become either reinforced ("I know I am boring since even my new friends don't like me") or dismantled ("I might not be boring because my new friends like me"; see Back & Vazire, 2015; Back et al., 2011; Fraley & Roisman, 2019). Our model embraces this large corpus of evidence and suggests that although mental inferences may have their roots in early temperament and attachment experiences, they can nevertheless change over time with new relational experiences (e.g., Luyten, Campbell, & Fonagy, 2020).

To formalize this developmental aspect of personality, we enable mental states and mental models to change over time based on contextual evidence. Under this scheme, mental states and mental models are (Dirichlet) priors that get updated at two timescales: τ (short) and t (long), respectively. We refer to these timescales as inference and learning respectively. The short timescale, τ, encoded in a transition matrix, represents how particular mental states can spontaneously shift from one moment to the next—for instance, quickly changing your mind that your therapist had a bad intention when checking the clock. By contrast, the long timescale, t, represents the relatively long learning of how particular mental models account for particular mental states—for instance, the gradual learning that the therapist may hold a positive or optimistic model of their patient because they continually exhibit benevolent intentions toward them.

A core argument of our article is that these social inference and learning processes could explain the emergence and maintenance of personality disorders. For example, a negative social history (featuring abusive relationships) may explain paranoid ways of being (such as the tendency to be mistrustful in relationships; Fonagy et al., 2015). Conversely, a chaotic social history (featuring disorganized relationships) may explain unstable ways of being (such as the tendency to be both mistrustful and credulous in relationships; Fonagy et al., 2019). In that sense, although personality disorders may be sometimes based on extreme temperaments, what might actually explain their emergence is poor social development (a pattern that we explore later).

A Formal Definition of Personality Disorder

To summarize, we have outlined a probabilistic, mentalizing, hierarchical, and dynamic model of personality disorder. This model suggests that personality disorder is generated from maladaptive ways of experiencing and relating to the self and others. This perspective is supported by evidence showing that mentalizing impairments map onto general aspects of personality disorder (capturing its general dysfunction), while traits map onto specific aspects of personality disorder (capturing its stylistic expression; Kerber et al., 2024; Nysaeter et al., 2023; Wendt et al., 2023). Accordingly, given that our goal is to provide a generative model of dimensional personality disorder, we focus on the general mechanism of mental inference because it might account for transdiagnostic variance in all personality disorders (as well as perhaps other mental disorders; Luyten et al., 2020). Although a limitation of this approach is that our model cannot account for more specific aspects of categorical personality disorders (for instance, specific maladaptive goals and behaviors that are unique to individual personality conditions), as we explore later on it nonetheless provides a principled framework for formally understanding the generation, maintenance, and treatment of dimensional personality disorder (by formally defining it as a relational disorder).

Simulations

In this section, we present three key simulations that outline how personality disorder is generated (from maladaptive ways of inferring the self and others), how it is developed and maintained (from maladaptive ways of learning the self and others), and how it can be ameliorated (from re-learning, or revitalizing, the self and others).[2]

Simulation 1: Inferring the Self and Others

For our first simulation, we explore how personality pathology could be generated from maladaptive mentalizing: formally, a way of inferring mental states that focuses too much on either "irrelevant" working models or "irrelevant" social evidence (noisy observations). Here, behavioral observations are emitted from a single actor, who can be interpreted to be "the self" (in which case the simulated agent mentalizes themselves) or "the other" (in which case the simulated agent mentalizes someone else). In either case, observed behaviors are drawn from a univariate normal distribution, $o \sim N(\mu = 5,\ \sigma = 4)$, embodying the ground truth that the actor who emits them has, on average, a moderate intention. In that sense, our simulated agents are expected to infer, on average, moderate intentions, despite sometimes observing overly poor ($o = 1$) or overly good ($o = 10$) behaviors.

We explore how agents make sense of these varying observations by leveraging five possible working models: the integrated, negative, positive, split, and empty working models. As mentioned earlier, these models can be formalized as probability distributions that denote the likelihood of invoking certain mental states to explain human behavior. Here, we explore the effects of five distributional configurations (normal, left-skewed, right-skewed, bimodal, and flat) showing how they could respectively formalize integrated, negative, positive, split, and empty ways of being. Crucially, for these simulations, we disallow learning (i.e., updating of working models) in order to purely explore the mental inferences from each working model.

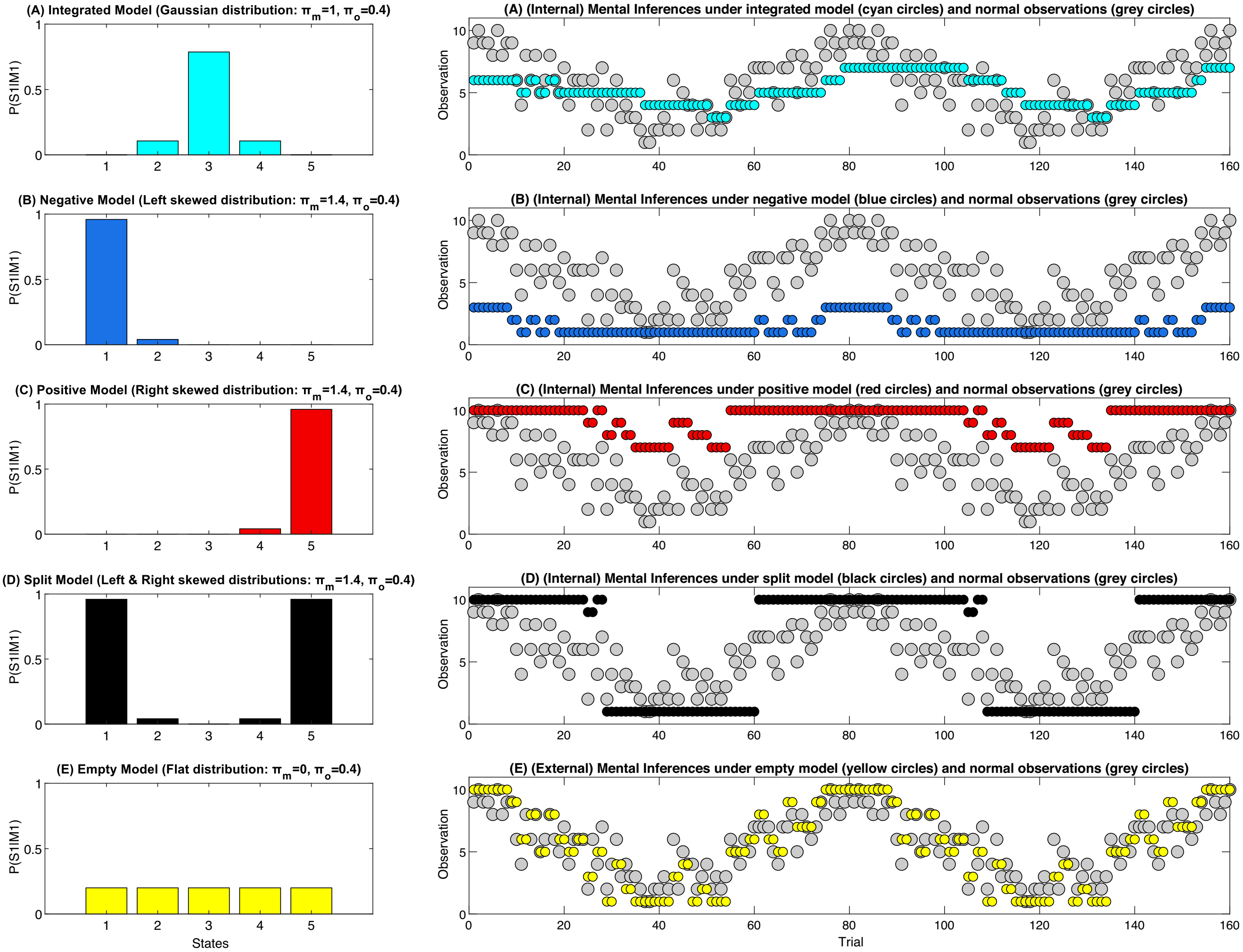
To begin with, we explore the mental inferences of an agent who embodies a normally distributed working model: $P(s|m) = N(\mu = 3, \sigma^2 = 1)$ (see Figure 3A). This model favors an initial mental inference of neutrality, $p(s_1 = 3 \mid m_1 = \text{integrated})$, with some uncertainty ($\sigma = 1/\pi_m = 1$) around this estimate. When an agent with this working model is faced with the observations outlined earlier, they mostly infer moderate intentions (in accordance with the true intention underlying the simulated observations). These patterns are visualized in Figure 3A, from which we can see that while observations (gray circles) fluctuate greatly, the posterior expectations of future behavior (cyan circles) remain around the average values ($o \sim 3 - 7$).

Based on previous work (see Story et al., 2024), we suggest that this coherent pattern of inference reflects what psychoanalysts have termed an "integrated" (or "balanced") sense of the self and others. For instance, self and ego psychologists have suggested that an "integrated" sense of self is characterized by a view of oneself as a worthy person despite sometimes exhibiting suboptimal behavior (Erikson, 1950; A. Freud, 1936; Goldberg, 1990; Hartmann, 1958; Mitchell, 1991). Likewise, object-relational and attachment theorists have outlined similar states of "object constancy" and "attachment security," which embody the expectation that close others are on the whole "good" and "loving" despite sometimes exhibiting hurtful ways of being (e.g., "my mother has left the room but has not abandoned me"; Ainsworth et al., 1978; Bowlby, 1969; Fairbairn, 1954; Klein, 1923; Winnicott, 1953). Finally, although other definitions of this "integrated" way of being exist (e.g., Beck et al., 2016; Fogel, 1993; Siegel, 2012; Young et al., 2006), they all converge in emphasizing the capacity to experience oneself and others in a coherent and balanced manner across time and contexts. Our

[2] We wish to highlight that our simulations provide a highly rarefied account of personality pathology. We hope that readers with clinical or lived experiences will bear with us by tolerating a necessary degree of simplification.

Figure 3

Internal Mentalizing Which Relies More on Internal Mental Models Rather Than External Social Evidence When Inferring Mental States: $\pi_m > \pi_o$



Note. (A) The integrated mental model yields balanced mental states with moderate uncertainty (variation) because it is characterized by a Gaussian hyper-prior that is endorsed with moderate levels of precision (or conviction) ($\pi_m = 1 > \ \pi_o = 0.4$). (B) A negative mental model yields negative mental states with low uncertainty because it is typified by a left-skewed hyper-prior that is endorsed with high levels of precision ($\pi_m = 1.4 > \ \pi_o = 0.4$). (C) A positive mental model is the mirror-opposite of the negative one and results in positive mental states with low uncertainty ($\pi_m = 1.4 > \ \pi_o = 0.4$). (D) The split mental model combines the positive and negative models to yield polarized (i.e., all-positive and all-negative) mental states with low uncertainty ($\pi_m = 1.4 > \ \pi_o = 0.4$). (E) The empty mental model presents the case of external rather than internal mentalizing: it is based on a flat (empty) hyper-prior that assigns the same probability to all mental states letting them be fully determined by social evidence: $\pi_m = 0 < \pi_o = 0.4$. See the online article for the color version of this figure.

model formalizes these approaches by casting the "integrated" self as a "regularized" inference: that is, an inference that promotes mental states that explain human behavior in stable and balanced ways (i.e., "shades of gray").

This integrated form of mental inference stands in stark contrast to the following forms of inference, which promote maladaptive ways of viewing oneself and others that are pathognomonic of personality disorders. For instance, the second form of mental inference is predicated on a left-skewed distribution, which favors negative mental states: $P(s|m) = N(\mu = 1,\ \sigma^2 = 1/\pi_m)$ where $\pi_m = 1.4$ (Figure 3B). Unlike integrated agents, those with this working model tend to view everything in a pessimistic way: their inferences are skewed toward negative values ($o \sim 1 - 3$) even when they are based on extremely positive observed values (e.g., first 20 trials).

We suggest that when these inferences are directed toward others, they reflect paranoid personality patterns. Indeed, much like our model depicts, individuals with paranoid personalities tend to attribute negative intentional states to others (McWilliams, 2011, p. 214). For instance, such individuals are more likely to blame others for their misfortunes (Murphy et al., 2018), misconstrue neutral and even benign events as hostile (Trotta et al., 2021), and exhibit conspiratorial thinking (Greenburgh & Raihani, 2022). Our model formalizes these paranoid patterns by suggesting that they are predicated on skewed models of others—models that distort incoming observations by casting them in a negative light (Story et al., 2024).

It is possible that this negativity is directed toward the self rather than others. In such circumstances, the result would be a "depressive" or "guilty" self that is pathognomic of internalizing personality pathology (Barlow et al., 2021). Such negative inferences toward the self would provide a principled way of viewing neuroticism (the hallmark feature of internalizing psychopathology), as well as traditional notions of the "depressive personality" which is typified by self-defeating psychopathology (Barlow et al., 2021; Beck, 1961; McWilliams, 2011).

Moving on to the third model, we observe the exact opposite patterns. In particular, under this right-skewed model, our simulated agent favors positive, rather than negative, mental states: $P(s|m) = N(5, 1/\pi_m)$ where $\pi_m = 1.4$ (Figure 3C). Thus, unlike the previous pessimistic agent, this agent views the world through rose-tinted glasses: their inferences are skewed toward positive values ($o \sim 7 - 10$) even when they are based on negative observed values (e.g., trials 30–50).

We suggest that these patterns could reflect grandiose ways of viewing the self. Indeed, in accordance with this profile, grandiose (or malignant) narcissism can be largely understood as a disorder of "the self": the self is elevated above all others, resulting in arrogant and performative (rather than authentic) ways of being (Grapsas et al., 2020). Notably, similar patterns have been noted in the so-called "hypomanic" personality, a more trait-like and less severe form of bipolarity that is closely aligned with narcissism (Nagel et al., 2023), reflecting the clinical phenomenon that "sustained periods of grandiosity may be associated with hypomanic mood" (American Psychiatric Association, 2013, p. 671). Our model formalizes these patterns by casting grandiose and hypomanic personalities as forms of inflated self-inference: an inference that consistently yields positive mental and affective states about oneself.

Of course, in everyday clinical practice, most narcissistically oriented patients are not purely grandiose (and hypomanic) but rather present with more vulnerable (and neurotic) features. Our fourth working model addresses this pattern by combining the previous two models (left- and right-skewed ones) to create a bimodal distribution (Figure 3D). Under this "split" distribution, our agent consistently draws either extremely positive or extremely negative mental inferences, alternating between them when recent life events indicate sufficient evidence to switch polarities. Notably, this switch in polarities can occur either slowly (i.e., the agent faces several minor life events that need to accumulate to switch a polarity; see first switch from idealization to devaluation in Figure 3D) or quickly (i.e., the agent faces a subjectively extreme life event that is enough on its own to abruptly switch polarities; see second switch from devaluation to idealization; see also Story et al., 2024).

These polarized inferences are remarkably consistent with the classical psychoanalytic notion of "splitting" (aka "dichotomous thinking"), which includes the tendency to view oneself or others in either "all good" or "all bad" manners (S. Freud, 1938; Klein, 1946). When applied to the self, this splitting dynamic may explain why narcissistically oriented patients can be both grandiose (when recent life observations allow them to maintain an idealized sense of self) and vulnerable (when faced with observations that challenge their idealized notions of themselves; see Pincus & Lukowitsky, 2010 for an extended discussion). Alternatively, when applied to others, this splitting pattern explains the tendency of patients with (borderline) personality disorder to idealize close others but abruptly devalue them once they observe suboptimal behavior from them (Gunderson, 2007). Our model formalizes these split ways of being by casting them as a form of bimodal inference: an inference that is predicated on fragmented mental models which have their origins in attachment histories that are marked by relational inconsistencies (e.g., the self is treated as an object of both love and hate; Ball & Links, 2009; Kernberg, 1967; Luyten et al., 2020; see the Simulation 2: Learning the Self and Others section).

Finally, we reach our last model which is paradoxically an empty model: that is, a model with a flat prior that assigns the same probability to all mental states: $P(s_{\tau=0}|m_t = \text{empty}) = 1/N = 1/5$. When we equip our agent with this empty model of the world, we find that they are a complete function of their emerging observations (see Figure 3E). Specifically, the agent changes their beliefs about themselves/others based solely on their most recent experiences. In that sense, the agent has no working model and is continually anticipating future behavior based purely on their most recent social observation.

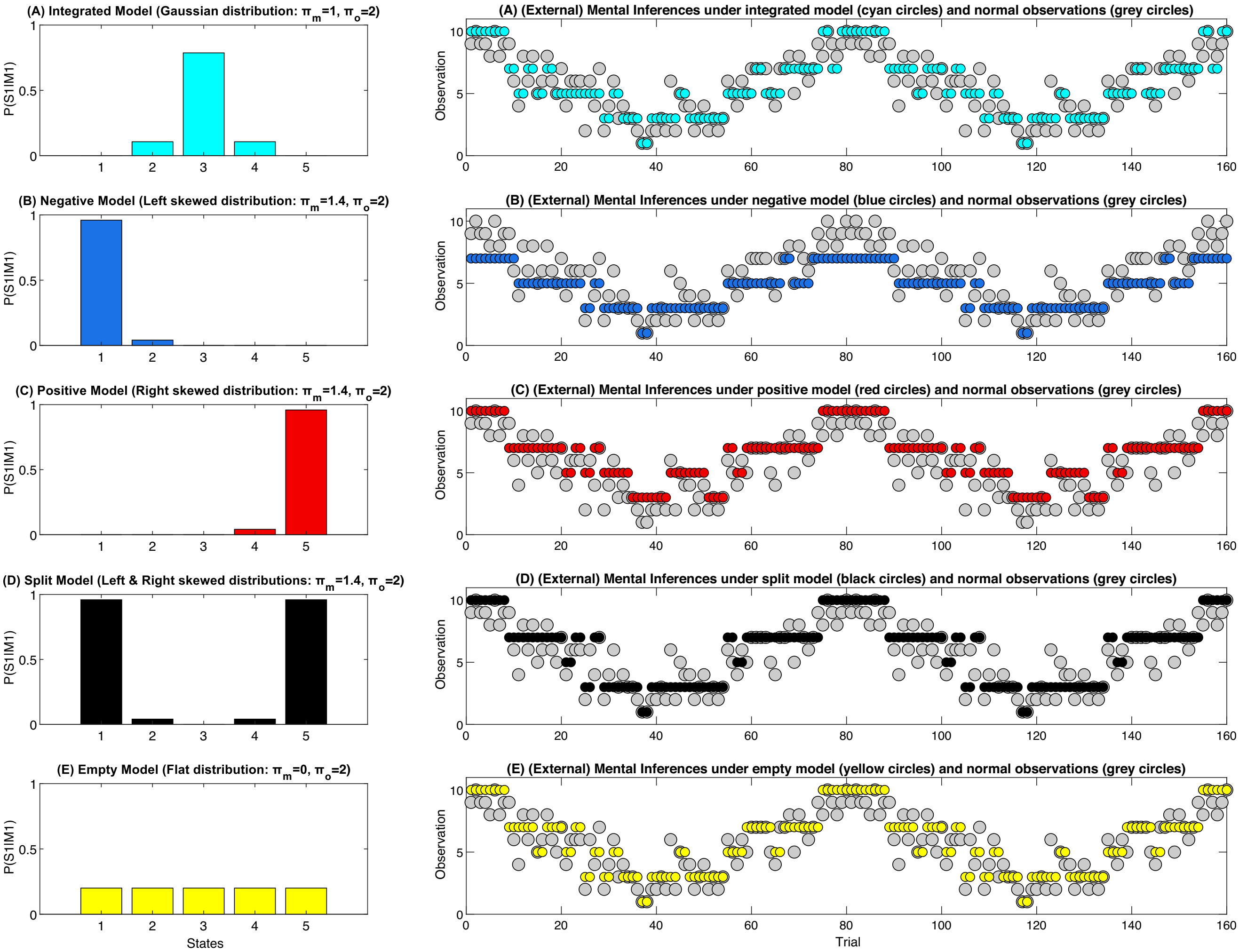
We propose that this plastic way of being reflects the unstable sense of self that typifies borderline personality disorder (Deutsch, 1942). Indeed, consistent with our model, individuals with "borderline" traits tend to frequently change themselves (including their appearance, friends, and belief systems) according to their most recent experiences (Kaufman & Meddaoui, 2021). Notably, evidence suggests that these unstable ways of viewing oneself are rooted in unstable rearing environments which promote the tendency to quickly adapt (and overidentify with) changing life circumstances (Luyten, Campbell, & Fonagy, 2020). Our model formalizes these plastic ways of being by casting them as a form of "external" inference: a mental inference that focuses too much on the immediate moment, rendering patients "prisoners of their present" (Rigoli, 2022) and ultimately "strangers to themselves" (Fuchs, 2007).

Interestingly, and somewhat counterintuitively, this form of "unstable" and "externally focused" inference can occur regardless of one's internal working model. That is, even a person with an integrated model of themselves/others can be relationally and emotionally unstable when they strongly weigh their social observations (see Rigoli, 2022). To illustrate this pattern, we repeat the previous simulations but this time we alter a key model parameter: the precision the agents place on their social observations relative to their models of the world ($\pi_0 > \pi_m$). Increasing this parameter ($\pi_0 = 2.0$) renders agents more sensitive to social experiences and yields unstable dynamics that are strikingly similar across all agents (albeit slightly attenuated in those with more rigid and polarized working models; see Figure 4).

We argue that these patterns of instability reflect a trait of "interpersonal hypersensitivity" that might be a transdiagnostic feature of all personality psychopathology (Gunderson & Lyons-Ruth, 2008). Indeed, a wealth of evidence on the circumplex of interpersonal sensitivities has illustrated that people with personality difficulties tend to be more sensitive to others' behaviors, particularly when those behaviors differ from their own (e.g., an overly cold person being sensitive to an overly warm person; Asan & Pincus, 2023; Cain et al., 2017; Dowgwillo & Pincus, 2017; Dowgwillo et al., 2018; Hopwood & Good, 2019; Hopwood et al., 2011). Moreover, factor-analytic evidence has indicated that the general factor of

Figure 4

External Mentalizing Which Relies More on External Social Evidence Rather Than Internal Mental Models When Inferring Mental States: $\pi_o > \pi_m$



Note. (A) The integrated mental model yields mental inferences that are strikingly similar to the mental inferences from the empty mental model because they similarly weigh social observations a lot more than internal mental models: $\pi_o = 2 > \pi_m = 1$. (B) Negative, (C) positive, and (D) split mental models also yield mental inferences that are similar to the ones from the empty mental model but also slightly different (skewed toward negative, positive, and polarized values, respectively) because they weigh social observations more, but not a lot more, than internal mental models: $\pi_o = 2 > \pi_m = 1.4$. (E) The empty mental model represents the most extreme case of external mental inference: namely, inferring mental states based purely on observable evidence without relying on an internal mental model: $\pi_o = 2 > \pi_m = 0$. See the online article for the color version of this figure.

personality (pathology) is reflected most strongly by symptoms relating to social sensitivities, including the tendency to be reactive, unstable, and insecure in social settings (Sharp et al., 2015; Wright et al., 2022). Finally, clinical and experimental evidence has demonstrated that people with personality difficulties tend to overweigh social stimuli during probabilistic reasoning, pointing to a specific sensitivity in the processing of social information (Rigoli, 2022; Zavlis, Story, et al., 2024). Our model is consistent with this evidence and formalizes interpersonal sensitivities as a form of "external" inference: a perceptual inference that overweighs social observations (relative to one's internal model of the world).

In that sense, our model delineates two kinds of mental inferences. The first kind places more emphasis on mental models (π_m) and is thus more "internal" and transferential: the self and the other are viewed in more rigid ways and their behavior is explained primarily in terms of one's internal working models. By contrast, the second kind places more emphasis on social observations (π_o) and is thus more "external" and ephemeral: the self and the other are viewed as empty vessels, ready to take any form (the lover, the villain, etc) based on their most recently emitted behaviors. These patterns are consistent with contemporary theorizing on internal versus external mentalizing (see Luyten et al., 2020) and provide a principled way of conceptualizing them in terms of probabilistic inference: mental inferences are impaired when they focus too much on rigid or irrelevant working models; or when they focus too much on ephemeral external evidence.

This point on internal versus external mental inference concludes our section on how humans infer themselves and others. In this section, we have explored the main dynamics of our model, showing how they can formalize archetypal personality patterns. In the next section, we expand on these matters by showing how personality problems get established and maintained during social development.

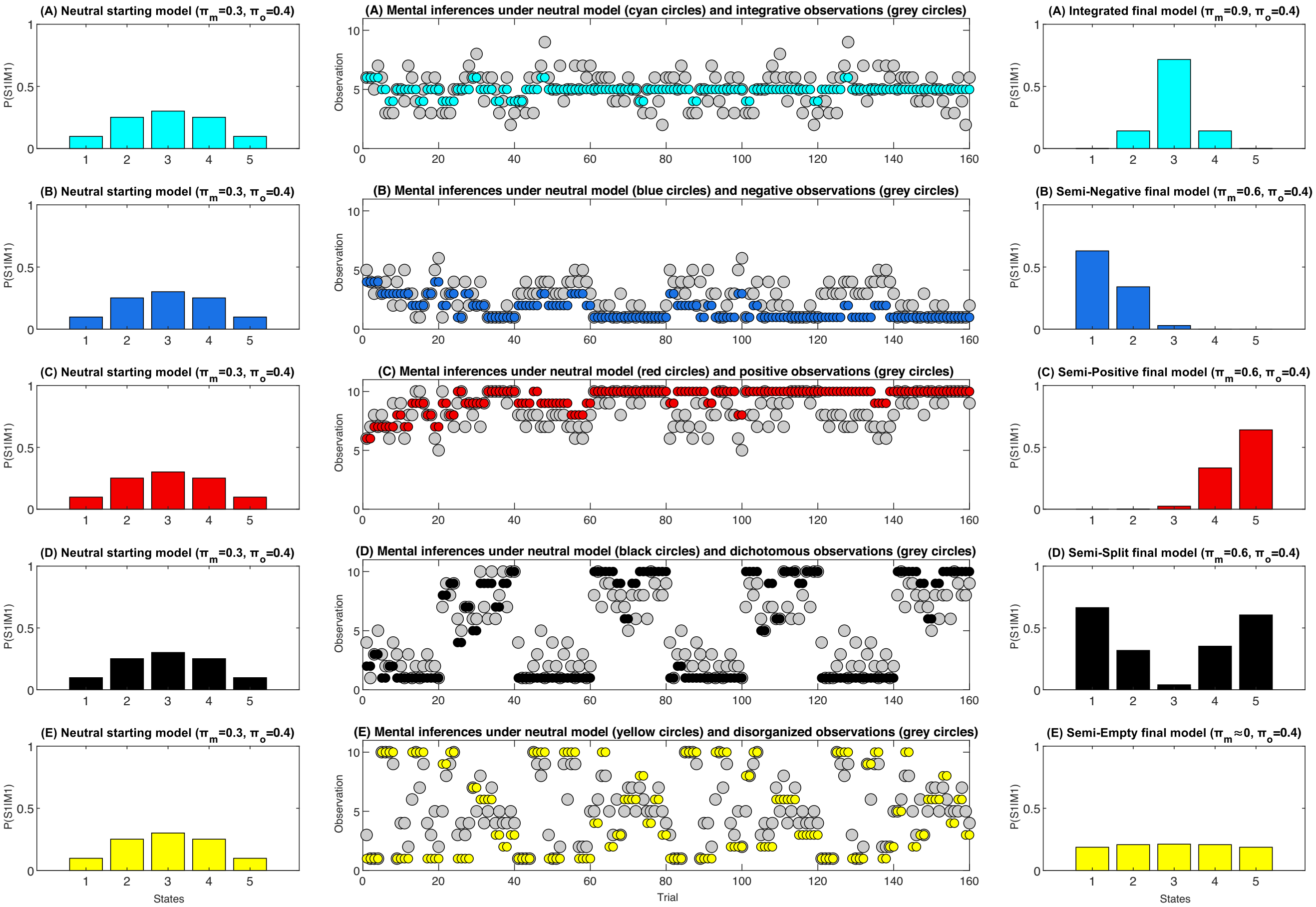
Simulation 2: Learning the Self and Others

In this section, we extend the previous set of simulations by showing how personality problems get established (and maintained) during social development. To do so, we simulate how a first agent (baby) internalizes observations from themselves, o_1, and another agent (caregiver), o_2, to learn themselves and others. Overall, we explore two possible "learning" routes: one based on a neutral starting condition (suggesting no psychiatric predisposition) and another based on skewed or plastic starting conditions (suggesting different types of psychiatric predisposition).

First, in the absence of a specific predisposition, we posit that a baby inherits a normally distributed model of the world that is typified by high variation (σ), aka uncertainty, because of brain plasticity: $P(s_{\tau=0}|m_t = \text{normal}) \sim N(\mu = 3,\ \sigma^2 = 1/\pi_m)$ where $\pi_m = 0.3$ (see Figure 5 left column). Given the neutral starting point of this model, maladaptive development is necessarily explained by maladaptive social experience (nurture). In particular, the development of skewed models of the self and others is predicated on skewed social experiences (i.e., mainly positive, mainly negative, or both); while the development of plastic models is predicated on disorganized social experiences (i.e., experiences under which all mental states are equally likely). These patterns are visualized in Figure 5, which demonstrates how neutral models develop into (A) integrated, (B) seminegative, (C) semipositive, (D) semisplit,

Figure 5

Development of Internal Working Models in Babies With No Psychiatric Predisposition



Note. The left panel illustrates babies' starting (inherited) working models which are all neutral and associated with integrative (normally distributed) behaviors. Middle panel illustrates mental inferences (colored circles) which are predicated on these normally distributed behaviors from the self (not visualized for clarity) and five kinds of behaviors from others (gray circles): integrative (normally distributed), (B) negative (left-skewed), (C) positive (right-skewed), (D) dichotomous (both negative and positive), and (E) disorganized (equal probability of being positive, neutral, or negative) behaviors. The right panel shows that the inherited neutral working models become (A) integrated (based on integrative self–other observations), (B) seminegative (based on negative other observations), (C) semipositive (based on positive other observations), (D) semisplit (based on dichotomous other observations), or (E) semiempty (based on disorganized other observations). See the online article for the color version of this figure.

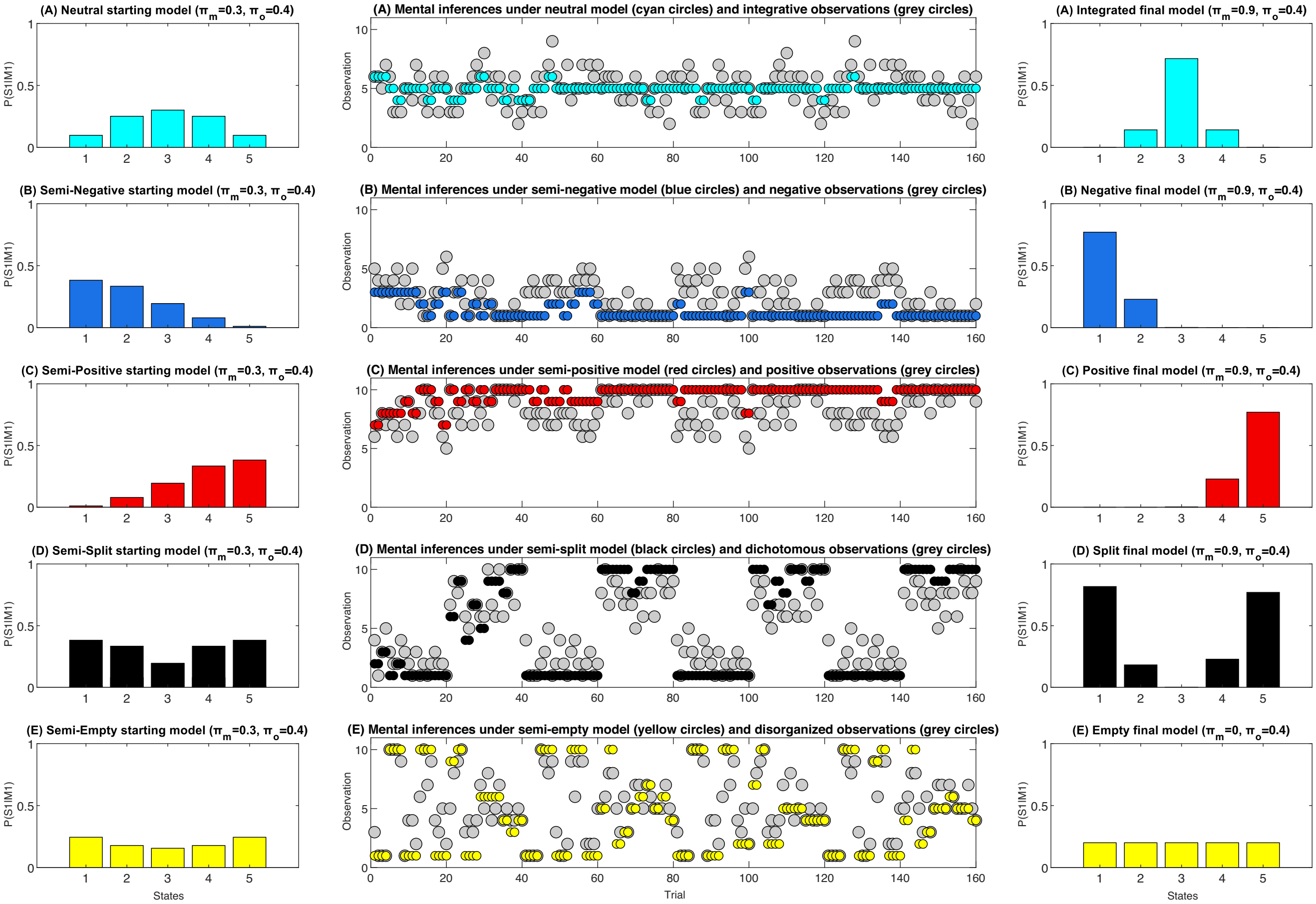
or (E) semiempty models based on (A) integrative, (B) negative, (C) positive, (D) dichotomous, and (E) disorganized experiences.

Although somewhat intuitive, we believe these patterns are noteworthy because they formalize established social effects on personality development. For example, our model formalizes well-known effects suggesting that depressive versus narcissistic working models arise, respectively, out of overly negative (e.g., caregiver abuse; Luyten et al., 2020) versus overly positive (e.g., caregiver overevaluation; Brummelman et al., 2015) experiences. Moreover, our model clarifies historically elusive concepts like "split" and "empty" working models by formally outlining how they could respectively develop from polarizing experiences (that result in a fragmented inner experience because of the presence of extreme "objects"; Kernberg, 1967) and disorganized experiences (that result in an experience of inner emptiness because of the absence of consistent "objects"; McWilliams, 2011, Chapter 2). In that sense, social development is cast as a type of probabilistic learning: that is, the gradual internalizing of mental inferences within internal working models (Fonagy et al., 2018).

Our second developmental route extends these patterns by suggesting that the development of maladaptive working models is not only "passive" (i.e., a baby simply "receives" adversity), but can also be "active": that is, a baby might be born with an interpersonal style that renders them more likely to experience themselves and others in a poor manner. For example, a baby with a skewed working model (overly positive, overly negative, or split) may invoke skewed reactions from others, which will in turn reinforce the baby's already skewed inferences. Likewise, a baby with a flat working model may display disorganized behavioral patterns, which could trigger mixed responses from others, reinforcing chaotic mental inferences. These patterns are graphically presented in Figure 6, from which we can see that initially maladaptive working

Figure 6

Development of Internal Working Models in Babies With Either (A) a Neutral Predisposition or (B–E) Various Psychiatric Predispositions



Note. The left panel illustrates babies' starting (inherited) working models which are (A) neutral, (B) seminegative, (C) semipositive, (D) semisplit, and (E) semiempty and are associated with (A) integrative (normally distributed), (B) negative (left-skewed), (C) positive (right-skewed), (D) dichotomous (both negative and positive), and (E) disorganized (equal probability of being positive, neutral, or negative) behaviors. The middle panel illustrates mental inferences (colored circles) which are based on these behaviors from the self (not visualized for clarity), as well as similar (Pearson's $r \sim 1$) behaviors from others (gray circles), triggered by those from the self. The right panel demonstrates that inherited working models become (A) integrated (based on integrative self–other observations), (B) negative (based on negative self–other observations), (C) positive (based on positive self–other observations), (D) split (based on dichotomous self–other observations), or (E) empty (based on disorganized self–other observations). See the online article for the color version of this figure.

models become more maladaptive than those from Figure 5 because of a specific interactional effect: the baby tends to emit particular behaviors, $o1$, which trigger corresponding reactions from their caregiver, $o2 \approx o1$, which end up reinforcing the baby's predisposed working model.

We argue that these patterns highlight the importance of interpersonal dynamics in establishing and maintaining personality styles. For example, the idea that personality styles get reinforced (or dismantled) via interpersonal transactions is noted by several interpersonal theorists, including those who emphasize person-environment feedback loops (Hopwood, Wright, & Bleidorn, 2022), those who highlight "if-then" behavioral modes (e.g., "If my caregiver is not around, I have to cry to get attention"; Mischel & Shoda, 1995), and those who underscore the law of interpersonal complementarity (e.g., that hostility from the self tends to trigger hostility in others; Pincus, 2005; Wright et al., 2023). Interestingly, these ideas help explain not only the development but also the maintenance of personality: Once maladaptive working models are established, they will result in rigid ways of relating to the self and others (i.e., low plasticity) that will likely trigger corresponding reactions from others (i.e., complementarity), culminating a cycle of social reinforcement (Back et al., 2011; Fonagy & Luyten, 2018; Hopwood, 2018; Sullivan, 1953; Wright et al., 2023). Importantly, these approaches have clear clinical implications (as reflected in our simulation), such as the prediction that clinicians should avoid "reacting in kind" toward their patients (e.g., becoming cold or hostile in response to cold or hostile patients) because such reactions run the risk of reinforcing (rather than revitalizing) their patients' maladaptive personality styles (Cain et al., 2024). Our model is remarkably consistent with these approaches and formalizes their view of "personal" disorders as "interpersonal" disorders: that is, disorders that are established and maintained socially through maladaptive ways of relating (see Hopwood, 2024; Lilienfeld et al., 2019; Wright et al., 2022).

This conception of "personality disorders" as "interpersonal disorders" concludes our section on how humans learn about themselves and others. In this section, we have shown how maladaptive social experiences are internalized to culminate maladaptive ways of viewing the self and others. In the next and final section, we illustrate how people can overcome these problems by revitalizing the way they view themselves and others.

Simulation 3: Revitalizing the Self and Others

In this final section, we explore how patients could overcome their relational problems by revitalizing how they view themselves and others. To do so, we simulate how agents with five different working models internalize mental inferences within therapeutic contexts to re-learn themselves and others. For this simulation, mental inferences are predicated on two pieces of evidence: first, the patient's behavior (which is biased in the direction of their working model distribution, as shown before); second, the therapist's behavior (which is considered unbiased as it only reflects normally distributed behaviors that are uncorrelated with those from their patients).

Figure 7 illustrates the patients' mental inferences under these circumstances. From this figure, we can see that the patient's initial inferences are biased: they are skewed toward values promoted by their internal working models. Over time, however, these biased inferences become ameliorated because patients learn how to view themselves and others in more integrated manners. By the end of treatment, maladaptive working models have changed considerably, indicating that patients have revitalized substantially how they mentally re-present themselves and others. Finally, and most importantly, the now-revitalized models are not only adaptive in this therapeutic context, but are also adaptive in other social contexts, as shown in Figure 8 which illustrates flexible mental inferences but in a different social environment (specifically, the patient's wider social environment, i.e., typified by normally distributed social experiences).

We suggest that these treatment patterns elucidate two mechanisms for alleviating personality psychopathology: a specific treatment mechanism and a general treatment mechanism. The specific treatment mechanism implies that rigid personality styles subside by decreasing the precision (certainty) of mental states, but that plastic personality styles subside by increasing the precision (certainty) of mental states. These patterns are reflected in the difference between the precision parameters at the beginning (Figure 7 left column) versus the end (Figure 7 right column) of psychotherapy. Specifically, rigid personality styles (reflected in the negative, positive, and split working models) start with high precision (certainty) in mental states $\pi_m = 1.2$ but drop to $\pi_m = 0.8$ after therapy (Figure 7B–7D); whereas plastic personality styles (reflected in the extreme case of the empty working model and the moderate case of the neutral working model) start with low precision in mental states $\pi_m = 0 - 0.4$ but rise to $\pi_m = 0.4 - 1.0$ after therapy (Figure 7A and 7E).

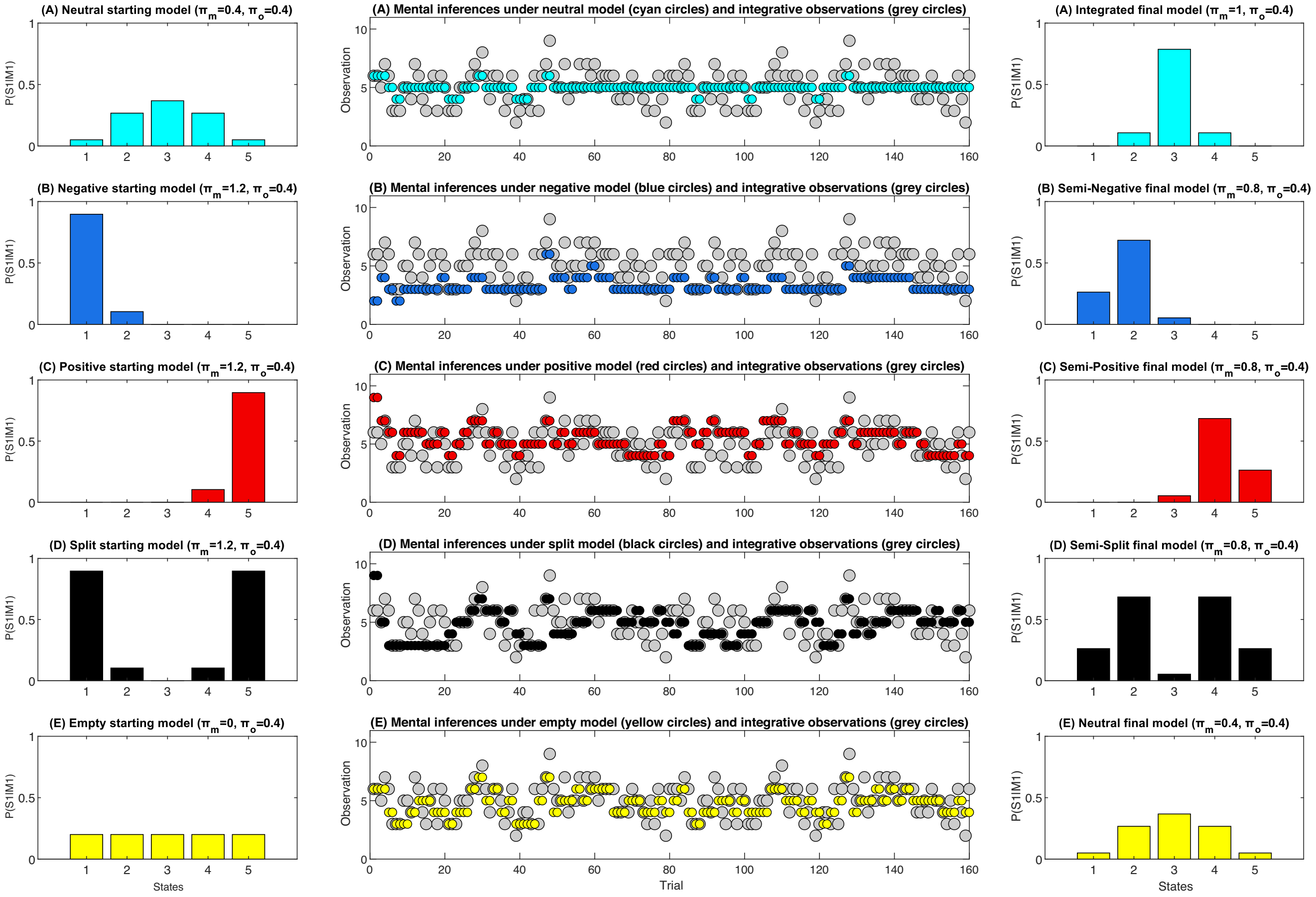
These patterns are consistent with evidence on the treatment of rigid versus plastic personality problems. Specifically, evidence from metacognitive therapies suggests that rigid ways of being (e.g., viewing others in polarized manners) subside when people become more "nuanced" and "uncertain" in their mentalizing (Bateman & Fonagy, 2016; Kernberg et al., 2008; Young et al., 2006). Conversely, evidence from cognitive-behavioral therapies suggests that plastic ways of being (e.g., having no identity or direction in life) ameliorate when people adopt more "precision" and "certainty" in their decision-making (Beck et al., 2016; Hayes et al., 2011; Speed et al., 2018). Our model formalizes these personality dynamics of rigidity versus plasticity (cf. DeYoung, 2015) and offers a clear and formal way of testing their treatment empirically—specifically, by tracking the precision of working model distributions across the course of psychotherapeutic interventions.

Beyond these specific treatment mechanisms, our model further highlights a general treatment mechanism for personality pathology: namely, a mechanism of relational psychotherapy that suggests that the treatment of personality disorder should be centered around emotionally meaningful relationships that provide evidence against maladaptive working models, revitalizing how patients view themselves and others (Figure 7). This aspect of our model is supported by a wealth of clinical trial research suggesting that this type of "relational practice" is currently the gold-standard treatment for personality disorder (Bateman & Fonagy, 2000; Cain et al., 2024; Trevillion et al., 2022, p. 22). Our model formalizes relational psychotherapy as a form of social learning: that is, the provision of integrative relational experiences, which ameliorate maladaptive working models and thereby enable patients to more effectively learn from and adapt to their social contexts.

Importantly, our model extends this relational psychotherapy perspective on personality psychopathology by positing at least three

Figure 7

Integration of Internal Working Models During Psychotherapy



Note. The left panel illustrates agents' starting (beginning of psychotherapy) working models: (A) neutral, (B) negative, (C) positive, (D) split, and (E) empty, which are associated with (A) integrative (normally distributed), (B) negative (left-skewed), (C) positive (right-skewed), (D) dichotomous (both negative and positive), and (E) disorganized (equal probability of being positive, neutral, or negative) behaviors. The middle panel illustrates mental inferences (colored circles) which are based on these behaviors from the self (not visualized for clarity) and therapist-behaviors that are integrative: that is, normally distributed with high precision (gray circles). The right panel demonstrates that the starting working models became more integrated at the end of treatment by shifting toward more neutral mental states and changing their certainty (precision). See the online article for the color version of this figure.

formal predictions regarding its inner workings. First and foremost, our model suggests that the psychotherapy of personality pathology is predicated on ways of relating that are not "reactive" (i.e., the therapist "reacting in kind" toward an unstable patient, as Figure 6 illustrates), but rather "integrative" (i.e., the therapist relating in a consistent and integrated manner that regulates their patient, as Figure 7 illustrates). Second, our model predicts that the treatment duration of personality disorder is directly correlated with the severity of the disorder: more severe cases of personality disorder would require a lot more counter-evidence to subside, an assumption that is consistent with existing evidence on the long-term treatment of severe personality disorders (Lindfors et al., 2015). Finally, our model implies that the treatment of personality disorder operates by opening up channels of social learning outside the confines of psychotherapy, in keeping with recent perspectives on the social mechanisms of psychotherapy (Fonagy et al., 2015, 2019). This restoration of broader social learning (and "epistemic trust") is shown in Figure 8 which illustrates how patients who have revitalized their working models can more flexibly mentalize and learn from social settings outside psychotherapy, a capacity that was not active prior to psychotherapy (e.g., Figures 3 and 4).

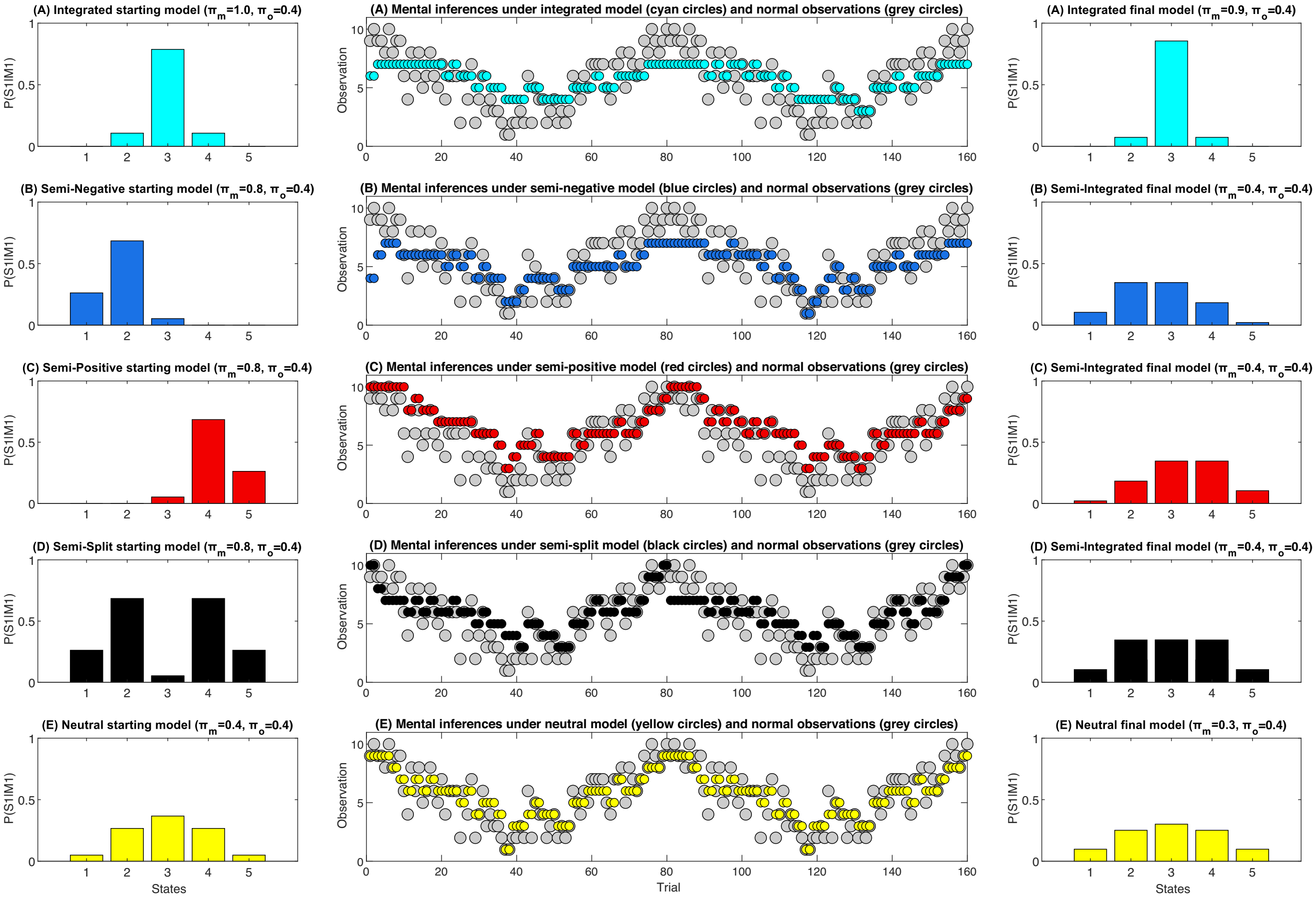
To summarize, we have outlined two mechanisms for treating personality pathology: a specific treatment mechanism (that tunes the precision of inner working models) and a general treatment mechanism (that opens up the capacity for social learning, both within and outside therapeutic settings). Together, these mechanisms highlight that what is central to the treatment of personality disorder might be meaningful relationships, which restore the patient's capacity for social openness and, in doing so, revitalize how patients re-present and relate to themselves and others.

Discussion

In this article, we have formalized personality disorder as a relational disorder: a maladaptive way of experiencing and relating to oneself and others. Our formal perspective was able to demonstrate

Figure 8

Adaptation of Working Models to Their Social Contexts



Note. The left panel illustrates the agents' starting models which are those after psychotherapy in Figure 7: specifically, (A) integrated, (B) seminegative, (C) semipositive, (D) semisplit, and (E) neutral. The middle panel illustrates flexible mental inferences (colored circles) which are based on behaviors from the self (that follow from working models and are not visualized for clarity) as well as normally distributed behaviors from others (gray circles). The right panel demonstrates that the starting working models adapt to their social context because they are more open to change (because of prior effects from psychotherapy). See the online article for the color version of this figure.

how various personality problems get established, maintained, and ameliorated over time. In the discussion that follows we expand on these matters, commenting on how they can assist in revitalizing the academic study and clinical practice of personality psychopathology.

Personality Functioning as Relational Functioning

To begin with, we note that contemporary research on personality disorders has been fragmented by trying to place either personality traits or personality functioning at the heart of personality pathology (see Hopwood, 2025; Zimmermann, 2022 for impartial commentaries). Our view, however, is that such either–or splits are false dichotomies because traits and functions are two sides of the same personality coin: the former describes the structure of personality, while the latter attempts to explain how that structure functions in everyday life (Wright & Kaurin, 2020). Arguably, while the former has a strong research base (given decades of inquiry), the latter has somewhat lagged behind (given not only its recency but also the predominance of descriptive models which do not, in our view, help separate it from traits; Zavlis & Fonagy, 2025).

On that basis, we suggest that our model could help revitalize the study of personality functioning by formalizing it as a dynamic process of mentalizing-relational functioning: that is, the ability to conceive oneself and others in intentional terms and in so doing adaptively relate to them. Indeed, consistent with our definition, personality dysfunction has been defined in largely relational terms: namely, as maladaptive self-relatedness (which impedes the capacity to construct a coherent identity and direction in life) and other-relatedness (which impedes the capacity to establish and maintain meaningful relationships; Sharp & Wall, 2021). Accordingly, our model suggests that the most fundamental personality pathologies might be maladaptive ways of relating to oneself and others, which arise, almost by definition, from maladaptive ways of experiencing oneself and others (which are in turn rooted in "distorted working models of the self and others"; Krueger, 2013, p. 355). In that sense, our model suggests that the core malfunctioning of personality is maladaptive relating highlighting that personality disorders may be more functionally understood as relational disorders: that is, disorders of maladaptive "self–other relatedness" (Sharp & Wall, 2018, p. 113).

Dynamic Modeling of Personality Functioning

Several implications follow from our functional definition of personality disorder. The first and most obvious is that this definition highlights the distinction between personality description and explanation. This distinction was first set forth by Allport (1937), who noted that "personality *is* and personality *does*," a sentiment currently echoed by dynamic theorists who distinguish between personality structures (five-factor traits) and personality functions (dynamic mechanisms; Cervone, 2005; Mischel, 2004; Revelle, 1995; Wright & Hopwood, 2016). Our model solidifies this dichotomy and validates a growing chorus of voices suggesting that there is value in conceptualizing personality (pathology) not only by what "it is" but also by what "it is supposed to do" (Wright & Kaurin, 2020, p. 193). At the same time, however, our model adds to these voices by providing a comprehensive computational framework on how personality (mal)functions.

To briefly illustrate this value of our framework, we consider one of its possible empirical applications: namely, the use of hidden Markov models (HMMs) in longitudinal data sets. To elaborate, data-driven HMMs can operationalize our framework by revealing latent personality states: that is, attractor states toward which patients gravitate (Haslbeck & Ryan, 2022). For example, applying HMMs to ambulatory data might reveal that some patients gravitate toward grandiose states (typified by inflated self-perceptions and other-harming behavior), other patients gravitate toward neurotic states (typified by deflated self-perceptions and self-harming behavior), and yet other patients cycle dynamically between these neurotic and grandiose states. In the online supplemental materials, we have outlined such empirical examples in greater detail, signposting interested readers to existing tutorials for these analyses (e.g., Visser & Speekenbrink, 2022). Our hope with this special issue article is to inspire researchers to move beyond the view of personality functioning as a static and unitary entity and start investigating it as a dynamic relational entity that comprises a multiplicity of self–other states that can shift over time and space.

Precise Tests of Personality Processes

A second and related implication of our model is that it provides a framework for generating such precise quantitative predictions on how particular personality states function. For example, as we have shown earlier, our model predicts that those with narcissistic traits will tend to alternate between self-states of worthlessness and grandiosity while those with borderline traits will tend to be a function of their environments. Interestingly, although some of those predictions were explicitly installed in our model (based on available evidence), others emerged without a priori expectations, highlighting the utility of formal modeling over traditional verbal theorizing in generating precise, undogmatic, and sometimes counterintuitive scientific hypotheses (Lewandowsky & Farrell, 2011).

Importantly, this hypothesis-generating utility could be leveraged to address recent concerns regarding the lack of theory-driven research in personality science (Benning & Smith, 2023; Hopwood, 2025; Kaurin et al., 2023). For example, the lack of transparency in preregistrations can be addressed with simulation studies such as this that demonstrate how specific personalities can be formalized so that they generate data that are consistent with theoretical hypotheses. In turn, these explicit hypotheses can be preregistered and then precisely tested in at least two ways: first, by examining the qualitative convergence of simulated data with empirical data; second, by examining the quantitative fit of formal models on empirical data (see Palminteri et al., 2017). Notably, such tests can be conducted not only in experimental data sets (which could provide stronger evidence for causal processes; Bailey et al., 2024) but also in longitudinal data sets (which might help us shed light on the timing of these processes; Hopwood, Bleidorn, & Wright, 2022). Historically, correlational and experimental research have been artificially segregated (Borsboom et al., 2009). It is our view, however, that together these lines of inquiry can enhance our knowledge of personality processes and culminate to an outline of personality (psychopathology) that is not merely descriptive but also explanatory.

Computational Processes in Psychotherapy

This leads us to the final implication of our model: the exploration of computational processes that ameliorate personality problems. Although traditional psychotherapy constructs (such as mentalizing, working models, and object-relational patterns) align remarkably with recent computational theorizing, little computational research has been conducted to formalize them as generative models and examine their theoretical nuances (see Barnby et al., 2023; Moutoussis et al., 2018). Our model takes a first step in this direction by formally outlining disparate clinical concepts and in so doing enabling their integration and formal examination in real-life therapeutic contexts.

For example, as we have outlined earlier, the idea that more uncertainty (certainty) in mentalizing ameliorates rigid (plastic) working models could be examined by tracking the variance of higher-order (prior) distributions during the course of metacognitive interventions, such as schema (Young et al., 2006), transference-focused (Kernberg et al., 2008), mentalization-based (Bateman & Fonagy, 2016), and cognitive-interpersonal (Dimaggio et al., 2015) psychotherapies. Likewise, these top-down mechanisms could be fruitfully linked to concrete bottom-up social difficulties, examining how they ameliorate during interpersonal therapies (e.g., Anchin & Pincus, 2010; Cain et al., 2024). Although we acknowledge that these hypotheses are somewhat speculative at the time of writing, we also note that they are remarkably consistent with both traditional and modern perspectives on the treatment of personality disorder, which stress the importance of "relational practice" (Bateman & Fonagy, 2016; Cain et al., 2024; Kernberg et al., 2008; Trevillion et al., 2022; Zavlis, 2023). On that basis, we are hopeful that a more formal and relational way of investigating them could lead to a more unified way of treating personality/relational problems.

Limitations and Future Directions

That being said, some limitations of our model must be acknowledged. The first and most obvious is that many of its computational predictions have not yet been empirically tested. To be sure, some of these predictions (such as the tendency of people with borderline personality disorder to overweigh their social experiences) are supported by emerging computational research on personality disorder (see Zavlis, Story, et al., 2024 for a systematic review). Nevertheless, more specific predictions (e.g., those on working model distributions) will require more direct empirical examinations. Although we have not conducted such examinations, we have provided an online tutorial of our generative framework to facilitate its empirical examination (https://github.com/OrestisZavlis/RelationalDisorder).

A second and related limitation concerns our simulations, which were constrained in several ways. First, our simulations were constrained as they did not explain the key mentalizing problems of all personality disorders (partly because of space constraints; partly because of our focus on dimensional personality pathology; and partly because of the lack of substantive theory to guide more specific modeling). Accordingly, future work may wish to focus on individual personality conditions and model their intricacies using our open framework. Second, our simulations focused on intentional mental states that reflect worthiness in oneself and trustworthiness in others. Although we note that this choice was motivated by the theory of mentalizing, we also acknowledge that our model could be applied to other kinds of personality states, including schematic belief states (in accordance with schema theory; Baldwin, 1992; Bartlett, 1995; Young et al., 2006) and dysregulated emotional states (in accordance with biosocial theory; Crowell et al., 2009; Linehan, 1993). For a tutorial on how to examine such latent personality states, please refer to the online supplemental materials. Finally, we note that our simulations focused only on perceptual inferences (how humans construe themselves and others), neglecting action components (how humans emit specific behaviors toward themselves and others). Future research could thus extend our framework by incorporating Markov decision processes: that is, stochastic choices of actions that move agents closer to their idiosyncratic goals (see Smith et al., 2022 for a tutorial). Such a process model would be in accordance with contemporary cybernetic perspectives on personality (DeYoung, 2015; Safron & DeYoung, 2021) and formalize them by outlining how specific personalities propel specific ways of relating to oneself and others as a means of achieving certain ends (e.g., a dependent personality exhibiting clinging behavior to alleviate separation anxiety). Notably, such a model would also enable us to test the novel view of "personality disorders" as "interpersonal disorders" by examining whether these disorders are best defined by both how a self relates to others (e.g., a person with dependent personality excusing the abusive behavior of their partner by idealizing them and devaluing themselves) and how others relate to the self (e.g., their partner with narcissistic proclivities perpetuating this pattern by denigrating their partner while idealizing themselves).

Finally, we note that we have not included people with lived experience of personality disorder in the process of developing our model. Although this omission may not be viewed as a limitation per se, we believe that it can be a problem because historically it has led to subjective and stigmatizing views on personality problems (e.g., "hysterical" personality); and, contemporarily, it has skewed research toward ideological debates rather than ways of further understanding and helping those who suffer from personality pathologies (Hopwood, 2025). On that basis, we believe that future researchers may wish to include patients in their research, particularly in theory-driven computational research, which we believe has the potential to de-stigmatize these problems by providing a more objective and precise way of casting them as problems with what a patient "does" (i.e., maladaptively relate), not as problems with who they "are" (i.e., maladaptive personality).

Conclusion

To conclude, we have provided a generative model of personality disorder. Our formal model extends current structural research on personality traits by providing a more functional perspective on how those traits manifest in everyday life. Through a series of simulations, we have argued that the fundamental function of personality is the ability to adaptively experience and relate to oneself and others. On that basis, we conclude that our model has formalized static personality disorders as dynamic relational disorders, highlighting that what might be central to them may not even be personality per se, but may rather be what S. Freud (1921/1955, p. 69) had intuited over a century earlier: human relationships.

References

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53–63. https://doi.org/10.1136/jnnp-2015-310737

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, Article 47. https://doi.org/10.3389/fpsyt.2013.00047

Ainsworth, M. D. S., Blehar, M. C., Wall, S., & Waters, E. (1978). *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum Associates.

Allport, G. W. (1937). *Personality: A psychological interpretation*. Henry Holt and Company.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Anchin, J. C., & Pincus, A. L. (2010). Evidence-based interpersonal psychotherapy with personality disorders: Theory, components, and strategies. In J. J. Magnavita (Ed.), *Evidence-based treatment of personality dysfunction: Principles, methods, and processes* (pp. 113–166). American Psychological Association. https://doi.org/10.1037/12130-005

Angyal, A. (1951). *Neurosis and treatment: A holistic theory*. Wiley.

Arnold, M. B. (1960). *Emotion and personality*. Psychological aspects (Vol. 1). Columbia University Press.

Asan, A. E., & Pincus, A. L. (2023). Examining schizotypal personality scales within and across interpersonal circumplex surfaces. *Assessment*, *30*(7), 2296–2317. https://doi.org/10.1177/10731911221143354

Atzil, S., Gao, W., Fradkin, I., & Barrett, L. F. (2018). Growing a social brain. *Nature Human Behaviour*, *2*(9), 624–636. https://doi.org/10.1038/s41562-018-0384-6

Back, M. D., Baumert, A., Denissen, J. J. A., Hartung, F.-M., Penke, L., Schmukle, S. C., Schönbrodt, F. D., Schröder-Abé, M., Vollmann, M., Wagner, J., & Wrzus, C. (2011). PERSOC: A unified framework for understanding the dynamic interplay of personality and social relationships. *European Journal of Personality*, *25*(2), 90–107. https://doi.org/10.1002/per.811

Back, M. D., & Vazire, S. (2015). The social consequences of personality: Six suggestions for future research. *European Journal of Personality*, *29*(2), 296–307. https://doi.org/10.1002/per.1998

Bailey, D. H., Jung, A. J., Beltz, A. M., Eronen, M. I., Gische, C., Hamaker, E. L., Kording, K. P., Lebel, C., Lindquist, M. A., Moeller, J., Razi, A., Rohrer, J. M., Zhang, B., & Murayama, K. (2024). Causal inference on human behaviour. *Nature Human Behaviour*, *8*(8), 1448–1459. https://doi.org/10.1038/s41562-024-01939-z

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion*. Rand Mcnally.

Baldwin, M. W. (1992). Relational schemas and the processing of social information. *Psychological Bulletin*, *112*(3), 461–484. https://doi.org/10.1037/0033-2909.112.3.461

Ball, J. S., & Links, P. S. (2009). Borderline personality disorder and childhood trauma: Evidence for a causal relationship. *Current Psychiatry Reports*, *11*(1), 63–68. https://doi.org/10.1007/s11920-009-0010-4

Barlow, D. H., Curreri, A. J., & Woodard, L. S. (2021). Neuroticism and disorders of emotion: A new synthesis. *Current Directions in Psychological Science*, *30*(5), 410–417. https://doi.org/10.1177/09637214211030253

Barnby, J. M., Dayan, P., & Bell, V. (2023). Formalising social representation to explain psychiatric symptoms. *Trends in Cognitive Sciences*, *27*(3), 317–332. https://doi.org/10.1016/j.tics.2022.12.004

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology*, *23*(3), 361–372. https://doi.org/10.1016/j.conb.2012.12.012

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bateman, A. W., & Fonagy, P. (2000). Effectiveness of psychotherapeutic treatment of personality disorder. *British Journal of Psychiatry*, *177*(2), 138–143. https://doi.org/10.1192/bjp.177.2.138

Bateman, A. W., & Fonagy, P. (2016). *Mentalization-based treatment for personality disorders: A practical guide*. Oxford University Press. https://doi.org/10.1093/med:psych/9780199680375.001.0001

Beck, A. T. (1961). A systematic investigation of depression. *Comprehensive Psychiatry*, *2*(3), 163–170. https://doi.org/10.1016/S0010-440X(61)80020-5

Beck, A. T., Davis, D. D., & Freeman, A. (Eds.). (2016). *Cognitive therapy of personality disorders* (3rd ed.). The Guilford Press.

Bender, D. S., Morey, L. C., & Skodol, A. E. (2013). Toward a model for assessing level of personality functioning in *DSM*-5, Part I: A review of theory and methods. In S. Huprich & C. Hopwood (Eds.), *Personality assessment in the DSM-5* (pp. 1–152). Routledge. https://doi.org/10.4324/9781315873091

Benning, S. D., & Smith, E. A. (2023). The registration continuum in personality disorder studies: Theory, rationale, and template. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 5–18. https://doi.org/10.1037/per0000602

Berrios, G. E. (1993). European views on personality disorders: A conceptual history. *Comprehensive Psychiatry*, *34*(1), 14–30. https://doi.org/10.1016/0010-440X(93)90031-X

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4). Springer.

Blatt, S. J. (2008). *Polarities of experience: Relatedness and self-definition in personality development, psychopathology, and the therapeutic process*. American Psychological Association. https://doi.org/10.1037/11749-000

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. Molenaar, M. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 67–97). Springer.

Bowlby, J. (1969). *Attachment and loss*. Random House.

Brummelman, E., Thomaes, S., Nelemans, S. A., Orobio De Castro, B., Overbeek, G., & Bushman, B. J. (2015). Origins of narcissism in children. *Proceedings of the National Academy of Sciences*, *112*(12), 3659–3662. https://doi.org/10.1073/pnas.1420870112

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*(1), 1–38. https://doi.org/10.1196/annals.1440.011

Cain, N. M., De Panfilis, C., Meehan, K. B., & Clarkin, J. F. (2017). A multisurface interpersonal circumplex assessment of rejection sensitivity. *Journal of Personality Assessment*, *99*(1), 35–45. https://doi.org/10.1080/00223891.2016.1186032

Cain, N. M., Hopwood, C. J., & Pincus, A. L. (2024). Psychotherapy through the lens of contemporary integrative interpersonal theory. In F. T. L. Leong, J. L. Callahan, J. Zimmerman, M. J. Constantino, & C. F. Eubanks (Eds.), *APA handbook of psychotherapy: Theory-driven practice and disorder-driven practice* (pp. 141–156). American Psychological Association. https://doi.org/10.1037/0000353-009

Cervone, D. (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology*, *56*(1), 423–452. https://doi.org/10.1146/annurev.psych.56.091103.070133

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Coker, L. A., Samuel, D. B., & Widiger, T. A. (2002). Maladaptive personality functioning within the big five and the five-factor model. *Journal of Personality Disorders*, *16*(5), 385–401. https://doi.org/10.1521/pedi.16.5.385.22125

Cozolino, L. J. (2014). *The neuroscience of human relationships 2e: Attachment and the developing social brain*. WW Norton.

Cramer, A. O. J., Van Der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414–431. https://doi.org/10.1002/per.1866

Crowell, S. E., Beauchaine, T. P., & Linehan, M. M. (2009). A biosocial developmental model of borderline personality: Elaborating and extending Linehan's theory. *Psychological Bulletin*, *135*(3), 495–510. https://doi.org/10.1037/a0015616

Deutsch, H. (1942). Some forms of emotional disturbance and their relationship to schizophrenia. *The Psychoanalytic Quarterly*, *11*(3), 301–321. https://doi.org/10.1080/21674086.1942.11925501

DeYoung, C. G. (2015). Cybernetic Big Five theory. *Journal of Research in Personality*, *56*(2), 33–58. https://doi.org/10.1016/j.jrp.2014.07.004

Dimaggio, G., Montano, A., Popolo, R., & Salvatore, G. (2015). *Metacognitive interpersonal therapy for personality disorders: A treatment manual*. Routledge. https://doi.org/10.4324/9781315744124

Dowgwillo, E. A., & Pincus, A. L. (2017). Differentiating dark triad traits within and across interpersonal circumplex surfaces. *Assessment*, *24*(1), 24–44. https://doi.org/10.1177/1073191116643161

Dowgwillo, E. A., Roche, M. J., & Pincus, A. L. (2018). Examining the interpersonal nature of criterion A of the *DSM-5* section III alternative model for personality disorders using bootstrapped confidence intervals for the interpersonal circumplex. *Journal of Personality Assessment*, *100*(6), 581–592. https://doi.org/10.1080/00223891.2018.1464016

Emanuel, A., & Eldar, E. (2022). Emotions as computations. *Neuroscience & Biobehavioral Reviews*, *144*(3), Article 104977. https://doi.org/10.1016/j.neubiorev.2022.104977

Erikson, E. H. (1950). *Childhood and society*. W. W. Norton.

Fairbairn, W. R. D. (1954). *An object-relations theory of the personality*. Basic Books.

Fogel, A. (1993). *Developing through relationships*. University of Chicago Press.

Fonagy, P. (1991). Thinking about thinking: Some clinical and theoretical considerations in the treatment of a borderline patient. *International Journal of Psychoanalysis*, *72*(4), 639–656. https://pubmed.ncbi.nlm.nih.gov/1797718/

Fonagy, P., Gergely, G., & Jurist, E. L. (2018). *Affect regulation, mentalization and the development of the self*. Routledge.

Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Development and Psychopathology*, *21*(4), 1355–1381. https://doi.org/10.1017/S0954579409990198

Fonagy, P., & Luyten, P. (2018). Attachment, mentalizing, and the self. In W. John Livesley & R. Larstone (Eds.), *Handbook of personality disorders: Theory, research, and treatment* (Vol. 2, pp. 123–140). The Guilford Press.

Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: A new conceptualization of borderline personality disorder and its psychosocial treatment. *Journal of Personality Disorders*, *29*(5), 575–609. https://doi.org/10.1521/pedi.2015.29.5.575

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2019). Mentalizing, epistemic trust and the phenomenology of psychotherapy. *Psychopathology*, *52*(2), 94–103. https://doi.org/10.1159/000501526

Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis*, *19*(1), 3–28. https://doi.org/10.1080/15294145.2017.1294031

Fradkin, I., Adams, R. A., Parr, T., Roiser, J. P., & Huppert, J. D. (2020). Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychological Review*, *127*(5), 672–699. https://doi.org/10.1037/rev0000188

Fraley, R. C., & Roisman, G. I. (2019). The development of adult attachment styles: Four lessons. *Current Opinion in Psychology*, *25*(4), 26–30. https://doi.org/10.1016/j.copsyc.2018.02.008

Freud, A. (1936). *The ego and the mechanisms of defence*. Imago Publishing.

Freud, S. (1921). *Group psychology and the analysis of the ego*. The International Psycho-Analytical Press.

Freud, S. (1930). *Civilization and its discontents*. The Hogarth Press.

Freud, S. (1938). Splitting of the ego in the defensive process. *The International Journal of Psychoanalysis*, *22*, 65–68.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*(4), Article 151. https://doi.org/10.3389/fpsyg.2012.00151

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. https://doi.org/10.1016/S2215-0366(14)70275-5

Fuchs, T. (2007). Fragmented selves: Temporality and identity in borderline personality disorder. *Psychopathology*, *40*(6), 379–387. https://doi.org/10.1159/000106468

Goldberg, A. (1990). *The prisonhouse of psychoanalysis*. Analytic Press.

Grapsas, S., Brummelman, E., Back, M. D., & Denissen, J. J. A. (2020). The "why" and "how" of narcissism: A process model of narcissistic status pursuit. *Perspectives on Psychological Science*, *15*(1), 150–172. https://doi.org/10.1177/1745691619873350

Greenburgh, A., & Raihani, N. J. (2022). Paranoia and conspiracy thinking. *Current Opinion in Psychology*, *47*(3), Article 101362. https://doi.org/10.1016/j.copsyc.2022.101362

Gunderson, J. G. (2007). Disturbed relationships as a phenotype for borderline personality disorder. *American Journal of Psychiatry*, *164*(11), 1637–1640. https://doi.org/10.1176/appi.ajp.2007.07071125

Gunderson, J. G., & 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

Haehner, P., Sleep, C. E., Miller, J. D., Lynam, D. R., & Hopwood, C. J. (2024). The longitudinal (co)development of personality traits and the level of personality functioning after negative life events. *Clinical Psychological Science*, *12*(4), 782–801. https://doi.org/10.1177/21677026231197607

Hartmann, H. (1958). *Ego psychology and the problem of adaptation* (D. Rapaport, Trans.). International Universities Press. https://doi.org/10.1037/13180-000

Haslbeck, J. M. B., & Ryan, O. (2022). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, *57*(5), 735–766. https://doi.org/10.1080/00273171.2021.1896353

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, *27*(6), 930–957. https://doi.org/10.1037/met0000303

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2011). *Acceptance and commitment therapy: The process and practice of mindful change*. Guilford Press.

Heider, F. (1983). *The psychology of interpersonal relations*. Lawrence Erlbaum.

Hopwood, C. J. (2018). Interpersonal dynamics in personality and personality disorders. *European Journal of Personality*, *32*(5), 499–524. https://doi.org/10.1002/per.2155

Hopwood, C. J. (2024). If personality disorder is just maladaptive traits, there is no such thing as personality disorder. *Journal of Psychopathology and Clinical Science*, *133*(6), 427–428. https://doi.org/10.1037/abn0000922

Hopwood, C. J. (2025). Personality functioning, problems in living, and personality traits. *Journal of Personality Assessment*, *107*(2), 143–158. https://doi.org/10.1080/00223891.2024.2345880

Hopwood, C. J., Ansell, E. B., Pincus, A. L., Wright, A. G., Lukowitsky, M. R., & Roche, M. J. (2011). The circumplex structure of interpersonal sensitivities. *Journal of Personality*, *79*(4), 707–740. https://doi.org/10.1111/j.1467-6494.2011.00696.x

Hopwood, C. J., Bleidorn, W., & Wright, A. G. (2022). Connecting theory to methods in longitudinal research. *Perspectives on Psychological Science*, *17*(3), 884–894. https://doi.org/10.1177/17456916211008407

Hopwood, C. J., & Good, E. W. (2019). Structure and correlates of interpersonal problems and sensitivities. *Journal of Personality*, *87*(4), 843–855. https://doi.org/10.1111/jopy.12437

Hopwood, C. J., Kotov, R., Krueger, R. F., Watson, D., Widiger, T. A., Althoff, R. R., Ansell, E. B., Bach, B., Michael Bagby, R., Blais, M. A., Bornovalova, M. A., Chmielewski, M., Cicero, D. C., Conway, C., De Clercq, B., De Fruyt, F., Docherty, A. R., Eaton, N. R., Edens, J. F., … Zimmermann, J. (2018). The time has come for dimensional personality disorder diagnosis. *Personality and Mental Health*, *12*(1), 82–86. https://doi.org/10.1002/pmh.1408

Hopwood, C. J., Wright, A. G., Ansell, E. B., & Pincus, A. L. (2013). The interpersonal core of personality pathology. *Journal of Personality Disorders*, *27*(3), 270–295. https://doi.org/10.1521/pedi.2013.27.3.270

Hopwood, C. J., Wright, A. G., & Bleidorn, W. (2022). Person–environment transactions differentiate personality and psychopathology. *Nature Reviews Psychology*, *1*(1), 55–63. https://doi.org/10.1038/s44159-021-00004-0

Hutsebaut, J., & Bender, D. S. (2024). The clinical utility of the level of personality functioning scale: A treatment perspective. *Journal of Psychiatric Practice*, *30*(6), 411–420. https://doi.org/10.1097/PRA.0000000000000822

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guilford Press.

Jung, C. G. (1919). Instinct and the unconscious. *British Journal of Psychology, 1904–1920*, *10*(1), 15–23. https://doi.org/10.1111/j.2044-8295.1919.tb00003.x

Kaufman, E. A., & Meddaoui, B. (2021). Identity pathology and borderline personality disorder: An empirical overview. *Current Opinion in Psychology*, *37*(1), 82–88. https://doi.org/10.1016/j.copsyc.2020.08.015

Kaurin, A., King, K. M., & Wright, A. G. (2023). Studying personality pathology with ecological momentary assessment: Harmonizing theory and method. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 62–72. https://doi.org/10.1037/per0000596

Kerber, A., Ehrenthal, J. C., Zimmermann, J., Remmers, C., Nolte, T., Wendt, L. P., Heim, P., Müller, S., Beintner, I., & Knaevelsrud, C. (2024). Examining the role of personality functioning in a hierarchical taxonomy of psychopathology using two years of ambulatory assessed data. *Translational Psychiatry*, *14*(1), Article 340. https://doi.org/10.1038/s41398-024-03046-z

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., Yeomans, F. E., Clarkin, J. F., & Levy, K. N. (2008). Transference focused psychotherapy: Overview and update. *The International Journal of Psychoanalysis*, *89*(3), 601–620. https://doi.org/10.1111/j.1745-8315.2008.00046.x

Klein, M. (1923). The development of a child. *The International Journal of Psycho-Analysis*, *4*(1), 419–474.

Klein, M. (1930). The importance of symbol-formation in the development of the ego. *International Journal of Psychoanalysis*, *11*(1), 24–39.

Klein, M. (1946). Notes on some schizoid mechanisms. *The International Journal of Psychoanalysis*, *27*(3), 99–110.

Knox, J. (2003). *Archetype, attachment, analysis: Jungian psychology and the emergent mind*. Routledge.

Krueger, R. F. (2013). Personality disorders are the vanguard of the post-*DSM*-5.0 era. *Personality Disorders: Theory, Research, and Treatment*, *4*(4), 355–362. https://doi.org/10.1037/per0000028

Lazarus, R. S. (1966). *Psychological stress and the coping process*. McGraw-Hill.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Sage Publications.

Lilienfeld, S. O., Watts, A. L., Murphy, B., Costello, T. H., Bowes, S. M., Smith, S. F., Latzman, R. D., Haslam, N., & Tabb, K. (2019). Personality disorders as emergent interpersonal syndromes: Psychopathic personality as a case example. *Journal of Personality Disorders*, *33*(5), 577–622. https://doi.org/10.1521/pedi.2019.33.5.577

Lindfors, O., Knekt, P., Heinonen, E., Härkänen, T., Virtala, E., & Helsinki Psychotherapy Study Group. (2015). The effectiveness of short-and long-term psychotherapy on personality functioning during a 5-year follow-up. *Journal of Affective Disorders*, *173*(4), 31–38. https://doi.org/10.1016/j.jad.2014.10.039

Linehan, M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Press.

Luyten, P., & Blatt, S. J. (2011). Integrating theory-driven and empirically-derived models of personality development and psychopathology: A proposal for *DSM* V. *Clinical Psychology Review*, *31*(1), 52–68. https://doi.org/10.1016/j.cpr.2010.09.003

Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The mentalizing approach to psychopathology: State of the art and future directions. *Annual Review of Clinical Psychology*, *16*(1), 297–325. https://doi.org/10.1146/annurev-clinpsy-071919-015355

Luyten, P., Campbell, C., & Fonagy, P. (2020). Borderline personality disorder, complex trauma, and problems with self and identity: A social-communicative approach. *Journal of Personality*, *88*(1), 88–105. https://doi.org/10.1111/jopy.12483

Luyten, P., & Fonagy, P. (2015). The neurobiology of mentalizing. *Personality Disorders: Theory, Research, and Treatment*, *6*(4), 366–379. https://doi.org/10.1037/per0000117

Maisto, D., Barca, L., Van den Bergh, O., & Pezzulo, G. (2021). Perception and misperception of bodily symptoms from an active inference perspective: Modelling the case of panic disorder. *Psychological Review*, *128*(4), 690–710. https://doi.org/10.1037/rev0000290

McWilliams, N. (2011). *Psychoanalytic diagnosis: Understanding personality structure in the clinical process*. Guilford Press.

Mikulincer, M., & Shaver, P. R. (2010). *Attachment in adulthood: Structure, dynamics, and change*. Guilford Publications.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, *55*(1), 1–22. https://doi.org/10.1146/annurev.psych.55.042902.130709

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/10.1037/0033-295X.102.2.246

Mitchell, S. A. (1991). Contemporary perspectives on self: Toward an integration. *Psychoanalytic Dialogues*, *1*(2), 121–147. https://doi.org/10.1080/10481889109538889

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Morey, L. C. (2019). Thoughts on the assessment of the *DSM*-5 alternative model for personality disorders: Comment on Sleep et al. (2019). *Psychological Assessment*, *31*(10), 1192–1199. https://doi.org/10.1037/pas0000710

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian Inferences about the self (and others): A review. *Consciousness and Cognition*, *25*(2), 67–76. https://doi.org/10.1016/j.concog.2014.01.009

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2018). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry*, *2*(1), 50–73. https://doi.org/10.1162/CPSY_a_00014

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). A formal model of interpersonal inference. *Frontiers in Human Neuroscience*, *8*(2), Article 160. https://doi.org/10.3389/fnhum.2014.00160

Murphy, P., Bentall, R. P., Freeman, D., O'Rourke, S., & Hutton, P. (2018). The paranoia as defence model of persecutory delusions: A systematic review and meta-analysis. *The Lancet Psychiatry*, *5*(11), 913–929. https://doi.org/10.1016/S2215-0366(18)30339-0

Nagel, M. G., Marcus, D. K., & Zeigler-Hill, V. (2023). Bipolar disorders and narcissism: Diagnostic concerns, conceptual commonalities and potential antecedents. *Clinical Psychology & Psychotherapy*, *30*(2), 235–249. https://doi.org/10.1002/cpp.2796

Nysaeter, T. E., Hummelen, B., Christensen, T. B., Eikenaes, I. U.-M., Selvik, S. G., Pedersen, G., Bender, D. S., Skodol, A. E., & Paap, M. C. S. (2023). The incremental utility of criteria A and B of the *DSM*-5 alternative model for personality disorders for predicting *DSM*-IV/*DSM*-5 section II personality disorders. *Journal of Personality Assessment*, *105*(1), 111–120. https://doi.org/10.1080/00223891.2022.2039166

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Pincus, A. L. (2005). A contemporary integrative interpersonal theory of personality disorders. In M. F. Lenzenweger & J. F. Clarkin (Eds.), *Major theories of personality disorder* (2nd ed., pp. 282–331). The Guilford Press.

Pincus, A. L., Cain, N. M., & Halberstadt, A. L. (2020). Importance of self and other in defining personality pathology. *Psychopathology*, *53*(3–4), 133–140. https://doi.org/10.1159/000506313

Pincus, A. L., & Lukowitsky, M. R. (2010). Pathological narcissism and narcissistic personality disorder. *Annual Review of Clinical Psychology*, *6*(1), 421–446. https://doi.org/10.1146/annurev.clinpsy.121208.131215

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. https://doi.org/10.1037/a0037250

Rauthmann, J. F., & Sherman, R. (2019). Toward a research agenda for the study of situation perceptions: A variance componential framework. *Personality and Social Psychology Review*, *23*(3), 238–266. https://doi.org/10.1177/1088868318765600

Read, S. J., & Miller, L. C. (2002). Virtual personalities: A neural network model of personality. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *The dynamic perspective in personality and social psychology* (pp. 357–369). Psychology Press.

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, *117*(1), 61–92. https://doi.org/10.1037/a0018131

Revelle, W. (1995). Personality processes. *Annual Review of Psychology*, *46*(1), 295–328. https://doi.org/10.1146/annurev.ps.46.020195.001455

Rigoli, F. (2022). Prisoner of the present: Borderline personality and a tendency to overweight cues during Bayesian inference. *Personality Disorders: Theory, Research, and Treatment*, *13*(6), 609–618. https://doi.org/10.1037/per0000549

Ringwald, W. R., Hopwood, C. J., Pilkonis, P. A., & Wright, A. G. C. (2021). Dynamic features of affect and interpersonal behavior in relation to general and specific personality pathology. *Personality Disorders: Theory, Research, and Treatment*, *12*(4), 365–376. https://doi.org/10.1037/per0000469

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Roche, M. J. (2018). Examining the alternative model for personality disorder in daily life: Evidence for incremental validity. *Personality Disorders: Theory, Research, and Treatment*, *9*(6), 574–583. https://doi.org/10.1037/per0000295

Rodriguez-Seijas, C., Rogers, B. G., & Asadi, S. (2023). Personality disorders research and social decontextualization: What it means to be a minoritized human. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 29–38. https://doi.org/10.1037/per0000600

Ryan, R. M., Deci, E. L., Grolnick, W. S., & La Guardia, J. G. (2015). The significance of autonomy and autonomy support in psychological development and psychopathology. In D. Cicchetti & D. J. Cohen (Eds.), *Developmental psychopathology: Theory and method* (2nd ed., pp. 795–849). John Wiley & Sons. https://doi.org/10.1002/9780470939383.ch20

Safron, A., & DeYoung, C. G. (2021). Integrating cybernetic Big Five theory with the free energy principle: A new strategy for modeling personalities as complex systems. In D. Wood, S. J. Read, P. D. Harms, & A. Slaughter (Eds.), *Measuring and modeling persons and situations* (pp. 617–649). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-819200-9.00010-7

Saxbe, D. E., Beckes, L., Stoycos, S. A., & Coan, J. A. (2020). Social allostasis and social allostatic load: A new model for research in social dynamics. *Perspectives on Psychological Science*, *15*(2), 469–482. https://doi.org/10.1177/1745691619876528

Sharp, C. (2022). Fulfilling the promise of the LPF: Comment on Morey et al. (2022). *Personality Disorders: Theory, Research, and Treatment*, *13*(4), 316–320. https://doi.org/10.1037/per0000567

Sharp, C., & Wall, K. (2018). Personality pathology grows up: Adolescence as a sensitive period. *Current Opinion in Psychology*, *21*(1), 111–116. https://doi.org/10.1016/j.copsyc.2017.11.010

Sharp, C., & Wall, K. (2021). *DSM*-5 level of personality functioning: Refocusing personality disorder on what it means to be human. *Annual Review of Clinical Psychology*, *17*(1), 313–337. https://doi.org/10.1146/annurev-clinpsy-081219-105402

Sharp, C., Wright, A. G., Fowler, J. C., Frueh, B. C., Allen, J. G., Oldham, J., & Clark, L. A. (2015). The structure of personality pathology: Both general ("g") and specific ("s") factors? *Journal of Abnormal Psychology*, *124*(2), 387–398. https://doi.org/10.1037/abn0000033

Shoda, Y. (2007). Computational modeling of personality as a dynamical system. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 633–651). The Guilford Press.

Shoda, Y., LeeTiernan, S., & Mischel, W. (2002). Personality as a dynamical system: Emergence of stability and distinctiveness from intra and interpersonal interactions. *Personality and Social Psychology Review*, *6*(4), 316–325. https://doi.org/10.1207/S15327957PSPR0604_06

Siegel, D. J. (2012). *The developing mind: How relationships and the brain interact to shape who we are*. Guilford Publications.

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, *107*(2), Article 102632. https://doi.org/10.1016/j.jmp.2021.102632

Speed, B. C., Goldstein, B. L., & Goldfried, M. R. (2018). Assertiveness training: A forgotten evidence-based treatment. *Clinical Psychology: Science and Practice*, *25*(1), Article e12216. https://doi.org/10.1111/cpsp.12216

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

Sullivan, H. S. (1953). *The interpersonal theory of psychiatry* (pp. xviii, 393). W. W. Norton.

Tomkins, S. S. (1978). Script theory: Differential magnification of affects. *Nebraska Symposium on Motivation*, *26*, 201–236. https://psycnet.apa.org/record/1982-11366-001

Trevillion, K., Stuart, R., Ocloo, J., Broeckelmann, E., Jeffreys, S., Jeynes, T., Allen, D., Russell, J., Billings, J., Crawford, M. J., Dale, O., Haigh, R., Moran, P., McNicholas, S., Nicholls, V., Foye, U., Simpson, A., Lloyd-Evans, B., Johnson, S., & Oram, S. (2022). Service user perspectives of community mental health services for people with complex emotional needs: A co-produced qualitative interview study. *BMC Psychiatry*, *22*(1), Article 55. https://doi.org/10.1186/s12888-021-03605-4

Trotta, A., Kang, J., Stahl, D., & Yiend, J. (2021). Interpretation bias in paranoia: A systematic review and meta-analysis. *Clinical Psychological Science*, *9*(1), 3–23. https://doi.org/10.1177/2167702620951552

Visser, I., & Speekenbrink, M. (2022). *Mixture and hidden Markov models with R*. Springer. https://doi.org/10.1007/978-3-031-01440-6

Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research Consortium, Nolte, T., Fonagy, P., Montague, P. R., Zimmermann, J., & Olaru, G. (2023). Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Personality and Mental Health*, *17*(2), 117–134. https://doi.org/10.1002/pmh.1566

Widiger, T. A., & Costa, P. T. (2012). Integrating normal and abnormal personality structure: The five-factor model. *Journal of Personality*, *80*(6), 1471–1506. https://doi.org/10.1111/j.1467-6494.2012.00776.x

Winnicott, D. W. (1953). Transitional objects and transitional phenomena: A study of the first not-me possession. *The International Journal of Psychoanalysis*, *34*(2), 89–97.

Wright, A. G. C., & Hopwood, C. J. (2016). Advancing the assessment of dynamic psychological processes. *Assessment*, *23*(4), 399–403. https://doi.org/10.1177/1073191116654760

Wright, A. G. C., & Kaurin, A. (2020). Integrating structure and function in conceptualizing and assessing pathological traits. *Psychopathology*, *53*(3–4), 189–197. https://doi.org/10.1159/000507590

Wright, A. G. C., Pincus, A. L., & Hopwood, C. J. (2023). Contemporary integrative interpersonal theory: Integrating structure, dynamics, temporal scale, and levels of analysis. *Journal of Psychopathology and Clinical Science*, *132*(3), 263–276. https://doi.org/10.1037/abn0000741

Wright, A. G. C., Ringwald, W. R., Hopwood, C. J., & Pincus, A. L. (2022). It's time to replace the personality disorders with the interpersonal disorders. *American Psychologist*, *77*(9), 1085–1099. https://doi.org/10.1037/amp0001087

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2006). *Schema therapy: A practitioner's guide*. Guilford Press.

Zavlis, O. (2023). Complex relational needs impede progress in NHS Talking Therapies (IAPT): Implications for public mental health. *Frontiers in Public Health*, *11*(4), Article 1270926. https://doi.org/10.3389/fpubh.2023.1270926

Zavlis, O. (2024a). Computational approaches to mental illnesses. *Nature Reviews Psychology*, *3*(1), Article 650. https://doi.org/10.1038/s44159-024-00360-7

Zavlis, O. (2024b). *The illusion of personality: Why personality disorders are actually relational disorders*. PsyArXiv. https://doi.org/10.31234/osf.io/b4d6v

Zavlis, O., Bentall, R., Fonagy, P., & Rigoli, F. (2024). *A formal theory of mood instability*. https://doi.org/10.31219/osf.io/rgv5m

Zavlis, O., & Fonagy, P. (2025). Beyond descriptive models of personality problems. *Journal of Personality Assessment*, *107*(2), 164–167. https://doi.org/10.1080/00223891.2024.2430322

Zavlis, O., Story, G. W., Fonagy, P., & Moutoussis, M. (2024). *Computational modeling of interpersonal dynamics in psychopathology: A systematic review and agenda for future work*. https://osf.io/5crp4/download

Zimmermann, J. (2022). Beyond defending or abolishing criterion A: Comment on Morey et al. (2022). *Personality Disorders: Theory, Research, and Treatment*, *13*(4), 321–324. https://doi.org/10.1037/per0000561

Received October 14, 2024
Revision received April 1, 2025
Accepted April 2, 2025 ■