



## City Research Online

### City, University of London Institutional Repository

---

**Citation:** Zavlis, O., Bentall, R. P., Fonagy, P. & Rigoli, F. (2025). A Formal Theory of Mood Instability. *Clinical Psychological Science*, doi: 10.1177/21677026251363862

This is the published version of the paper.

This version of the publication may differ from the final published version.

---

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/35772/>

**Link to published version:** <https://doi.org/10.1177/21677026251363862>

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

# A Formal Theory of Mood Instability



Orestis Zavlis<sup>1</sup>, Richard P. Bentall<sup>2</sup>, Peter Fonagy<sup>1</sup>,  
and Francesco Rigoli<sup>3</sup>

<sup>1</sup>Department of Psychology and Language Sciences, Unit of Psychoanalysis, University College London;

<sup>2</sup>Department of Psychology, Unit of Clinical Psychology, University of Sheffield; and <sup>3</sup>School of Health & Psychological Sciences, City, University of London

Clinical Psychological Science  
1–27

© The Author(s) 2025



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/21677026251363862

www.psychologicalscience.org/CPS



## Abstract

Despite empirical progress, theoretical understanding of mood instability remains stagnant. A major reason for this stagnation concerns the field's reliance on narrative theories that cannot integrate disparate quantitative perspectives on mood dynamics. Here, we address the limitations of narrative theorizing by developing a formal theory of mood instability. Our theory is predicated on the computational process of “evaluation”: the process of appraising the value of stimuli, which has long been theorized to be central to mood dynamics. Building on reinforcement-learning models of evaluation, we propose a dynamic framework, which we use to simulate various evaluative situations. Our simulations can generate and thereby formally integrate three well-known types of mood instability: emotional rigidity/inertia, transience/instability, and sensitivity/reactivity. We discuss how this formal perspective could enhance the theory, clinical utility, and measurement of mood instability.

## Keywords

affective disorders, computer simulation, cognition and emotion, open data, open materials

Received 12/2/23; Revision accepted 7/6/25

The core of affective experience, whatever its guise, is its evaluative nature. Affect introduces value in a world of factual perceptions and sensations.

—Nico Frijda

Burdening up to 14% of the general population, mood instability (MI; or “emotional instability”)<sup>1</sup> is an important yet often neglected endophenotype (Black et al., 2006; Marwaha, Parsons, Flanagan, & Broome, 2013). Although archetypally synonymous with two psychiatric conditions (i.e., bipolar disorder and borderline personality disorder [BPD]), MI is found, to some extent, in virtually all psychopathologies—with, indeed, as many as eight out of 10 psychiatric inpatients reporting some level of MI (Gilbert et al., 2005; Marwaha et al., 2014). Recently, some progress has been made in understanding MI. Neurobiologically, for instance, projections from the limbic system (amygdala, in particular) to the salience network have been consistently and transdiagnostically implicated in MI (for a review, see Broome, He, et al., 2015). Likewise, psychologically, MI has been associated with various mental-health

problems, such as poor functioning (Bowen et al., 2012), lower levels of well-being (Hills & Argyle, 2001), and increased suicidality (Marwaha, Parsons, & Broome, 2013).

Despite empirical progress, however, theoretical understanding of MI remains stagnant. Researchers in the field disagree on even basic questions, such as what the definition of MI is, what causes it, and whether different “types” of it exist (Broome, Saunders, et al., 2015). Consequently, the nomenclature surrounding MI is equivocal. Terms such as “emotional instability,” “mood swings,” “affective volatility” (or “lability”), “emotional dysregulation,” and “emotional impulsiveness” are used interchangeably by researchers and clinicians alike (for reviews, see Broome, Saunders, et al., 2015; Koenigsberg, 2010; Marwaha et al., 2014). Moreover, researchers from different disciplines (from psychiatry

## Corresponding Author:

Orestis Zavlis, Department of Psychology and Language Sciences, Unit of Psychoanalysis, University College London

Email: orestis.zavlis.23@ucl.ac.uk

to psychology and neuroscience) focus on putatively distinct “aspects” of MI, yielding varied definitions—from “excessive rise of emotion and delayed return to baseline” to “frequent oscillations between affective categories” and “intense emotional reactivity” (Broome, He, et al., 2015; Koenigsberg, 2010). These definitional ambiguities impede the measurement and pragmatic utility of MI in both research and clinical settings (Broome, Saunders, et al., 2015; Zimmerman et al., 2010).

The ambiguities around MI can be traced back to the field’s reliance on natural language as a means of specifying its theoretical aspects. Because of the vagaries of language (Bolton & Hill, 2004; Wittgenstein, 2010), theories specified at the narrative level have been argued to be severely limited (see Fried, 2020; Haslbeck, Bringmann et al., 2021; Robinaugh et al., 2021; Smaldino, 2020; Vallacher et al., 2017). First and foremost, narrative theories are vulnerable to implicit assumptions and hidden contradictions—consider, as an example, the conflation of self-esteem instability and mood instability (Broome, Saunders, et al., 2015). Second, even without such internal contradictions, narrative theories cannot be rigorously examined against empirical reality—because of their generality, such theories cannot be subjected to significant risk of refutation and thus have the propensity of remaining neither fully corroborated nor refuted (Robinaugh et al., 2021). Finally, narrative theories cannot address the growing complexity of quantitative approaches to mood dynamics—for example, despite progress in creating quantitative definitions for several MI types (i.e., “affective inertia” or “stress reactivity”; Dejonckheere et al., 2019), theoretical research has not progressed in parallel to provide a unifying account of how these quantitative taxonomies might be understood theoretically.

Formal theories may address these shortcomings. First and foremost, formal theories can define constructs more precisely by formalizing them with the language of mathematics (see Adams et al., 2016; Huys et al., 2011; Lewandowsky & Farrell, 2011). Second, because of their mathematical specificity, formal theories can generate precise quantitative predictions that can be more explicitly examined against empirical observations (Huys et al., 2016; Zavlis, 2024a). Finally, because of their generative nature, formal theories enable researchers to unify disparate theoretical and quantitative perspectives by showing how they could all be generated from specific computational processes (Friston et al., 2014; Montague et al., 2012). Because of these properties, formal theories could present a fruitful avenue for clarifying the nature of MI.

In this article, we aim to develop a formal theory of MI. Our theory is grounded on Marr’s (1982) approach to computational modeling, which involves formalizing

the psychological processes that the human mind/brain employs to solve particular life tasks. This approach has a long history in computational neuroscience (and most recently, computational psychiatry) and can provide a notable adjunct to contemporary quantitative perspectives to MI. For example, although several longitudinal models have already been constructed with the aim of elucidating mood dynamics (by creating well-defined parameters such as “affective inertia,” “instability,” and “reactivity”; Vanhasbroeck et al., 2021), these models do not address how their parameters might themselves be generated by a common computational process. As a consequence, whether these models could be unified under a common generative framework remains, at the time of writing, uncertain.

Our framework, as we outline later, bridges these disparate quantitative approaches by positing a computational process that can generate the various types of MI. In particular, our framework suggests that at least three distinct types of MI (i.e., “emotional rigidity,” or inertia; “emotional transience,” or instability; and “emotional reactivity,” or sensitivity) could be generated and thereby integrated by a common computational process: the process of assigning value to stimuli (i.e., “evaluation”). Before we turn to our model and the explanations it furnishes, some background information on these mood parameters is first necessitated. We provide this background information in the next section, in which we briefly review existing longitudinal approaches to MI.

## Longitudinal Approaches to MI

Existing longitudinal approaches have demonstrated that MI can be parsed into at least three basic types: inertia, instability, and reactivity. These types have, in turn, been operationalized as statistical parameters (e.g., variance or regression slopes) in different quantitative models. We briefly review this line of research next.

First, emotional inertia can be interpreted as the tendency of emotions to resist change (Suls et al., 1998). This MI type is typically estimated as the correlation of a given (mood) variable with itself over time (autocorrelation); higher autocorrelations imply a greater tendency of a certain emotion to “carry over” to the next time point (Box et al., 2015). Research on inertia has typically relied on (discrete-time) (vector-)autoregressive (VAR) models (Bringmann et al., 2018; Ernst et al., 2024; Haslbeck, Bringmann et al., 2021), revealing that depression is typified by strong autocorrelations of negative emotions (Koval et al., 2013). More recent network approaches have extended these univariate models in a multivariate context by estimating networks of emotion autocorrelations, revealing that they are

inflated in individuals who score high on trait neuroticism (Bringmann et al., 2016). Finally, more sophisticated dynamical models that operate on a continuous timescale, such as the continuous-time VAR model (see Guthrie et al., 2020; Oud & Jansen, 2000; Uhlenbeck & Ornstein, 1930), the damped linear oscillator (Boker & Nesselroade, 2002; Chow et al., 2005; Hu et al., 2014), or the model of intraindividual variability in affect (Wirth et al., 2022), have operationalized inertia as a regulation parameter that enables individuals to return to their baseline mood after having deviated from it. Although the estimation of emotional inertia may differ across these models, its interpretation converges, implying that the only fundamental difference between the models may be in the types of data they accommodate (see Vanhasbroeck et al., 2021).

The second well-known MI type is emotional instability, which refers to the fluctuation of emotions over time. Traditionally, emotional instability was approximated with the variance of time series (Larsen & Diener, 1987), revealing inflated mood variability in people with borderline personality (Cowdry et al., 1991; Stein, 1996; Stiglmayr et al., 2001). Despite its straightforwardness, however, this operationalization was eventually argued to be limited because it did not consider the temporal dependency of mood variability: that is, the fact that although emotions may be highly variable, they may still be highly autocorrelated over time, implying mood rigidity (inertia) rather than instability (Larsen, 1987). Accordingly, measures that combine both time variability and time dependency have been introduced, including the mean of all squared differences of affect states (Ebner-Priemer et al., 2009) and the probability of acute change (Jahng et al., 2008). Research using these more sophisticated measures of “instability” has shown that they are more elevated in patients with BPD compared with individuals with depression (P. Santangelo et al., 2014; Trull et al., 2008) and that they are negatively correlated with inertia (e.g., Dejonckheere et al., 2019; Thompson et al., 2012). These patterns imply that the concept of emotional instability can be considered as the inverse of inertia because it reflects the tendency of emotions to rapidly change rather than persist over time.

The final type of MI refers to emotional reactivity: the tendency to experience strong emotions in response to particular life situations. Emotional reactivity is intimately dependent on life events, implying that its measurement differs from the other mood concepts in that it requires a variable denoting one’s “appraisal” (or evaluation) of life events (Silk, 2019; Thompson et al., 2012). In empirical research, affects are regressed on this appraisal of life events, yielding a coefficient that reflects one’s propensity to experience strong affects in

response to certain life events within naturalistic (Dejonckheere et al., 2019) or laboratory contexts (Lapate & Heller, 2020). Traditionally, reactivity has been conceptualized in this discrete manner and shown to be elevated in individuals with BPD (Bortolla et al., 2020) and deflated in individuals with major depressive disorder (Bylsma et al., 2008). More recent generative models have extended these approaches by formalizing the link between appraisals and emotions in a probabilistic way (although this link has yet to be examined empirically; see Ryan et al., 2025). Whether reactivity is conceptualized with traditional statistical models or more recent generative models, its interpretation remains the same, reflecting the potency of emotions in response to particular life conditions.

To summarize, existing longitudinal work parses out MI into at least three quantitative subtypes: rigidity (or inertia), instability (or transience), and reactivity (or sensitivity). These quantitative types are immensely helpful quantitatively because they can reveal interesting patterns of MI over time (e.g., that people with borderline personality will be extremely reactive to social stimuli; Sadikaj et al., 2013). However, such approaches are somewhat limited theoretically because they do not elucidate the cognitive processes that humans employ to generate these patterns of MI in the first place.

Cognitive perspectives can address these explanatory shortcomings. Rather than directly modeling longitudinal data, these approaches leverage principles from cognitive sciences to posit a computational process assumed to be used by the human mind (Adams et al., 2016; Huys et al., 2016; Montague et al., 2012). This cognitive process is then used to simulate healthy psychological effects before turning to their maladaptive variants. In that sense, this cognitive approach offers a top-down (rather than bottom-up) perspective to mood dynamics because it starts from underlying cognitive principles and then derives formal predictions that can be more precisely examined against empirical observations.

In this article, we adopt this top-down perspective to build a formal theory of MI. This formal theory is based on the notion of evaluation: the process by which humans assign value to stimuli. This notion of evaluation (aka “appraisal”) is featured in virtually all theories on emotion, including traditional theories on basic emotions (Ekman & Cordaro, 2011), contemporary perspectives on the social nature of emotions (Barrett, 2017), and cognitive approaches on the functional aspects of emotions (Keltner & Gross, 1999; for an integrative review, see Lange et al., 2020). Here, we build on contemporary computational models of evaluation to assess whether these can generate the above-mentioned MI types. Before we introduce our formal framework, though, we first review existing cognitive

approaches to evaluation and emotion in the next section.

## Cognitive Approaches to Evaluation and Emotion

Existing cognitive work has highlighted that the process of evaluation has at least three key properties: It is reference-dependent, predicated on precision (or certainty in evaluation), and encompasses a nonlinear value function. We briefly review these properties in turn before showing how they have been integrated by the model used in this paper.

The first property is based on a wealth of evidence illustrating how mood and emotions are fundamentally dependent on prediction errors, that is, on mismatches between rewards that a person receives ( $R$ ) versus rewards the person expects ( $V$ ). For example, research by Rutledge et al. (2014) revealed that self-reported happiness is predicated on the discrepancy between one's expectation of rewards and one's actual (received) rewards (for replications, see also Villano et al., 2020; Vinckier et al., 2018; Will et al., 2017). Likewise, seminal work by Eldar and colleagues showcased that the same reward-prediction errors are linked to momentary affective states (Emanuel & Eldar, 2022) and are robustly associated with phasic dopaminergic activity (Eldar et al., 2016; Eldar & Niv, 2015). Together, these patterns suggest that affective valence is based not on life outcomes per se but rather on their discrepancy with a person's reference (e.g., expectation). A model of evaluative processes, then, should account for such reference effects.

A second important aspect emerging from empirical research concerns the role of (un)certainty: that is, how certain or precise individuals are in their evaluative inferences (an aspect also at the core of some recent reinforcement-learning models; Bellemare et al., 2023). This certainty facet has been well documented by research showing that the more certain one is in one's evaluation, the stronger one's affective response, either for better (reality exceeds expectations) or for worse (reality falls short of expectation; Clark et al., 2018; Hesp et al., 2021). Moreover, parameters of (un)certainty are at the core of recent formal models of various psychopathologies which assert that people with psychosis and borderline personality tend to experience stronger emotions because which are overcertain in their evaluations (Henco et al., 2020; Rigoli, 2022a; Zavlis et al., 2025a; Zavlis et al., 2025b). These patterns imply that the concept of evaluative certainty could be leveraged to formalize emotional reactivity, a possibility that we turn to subsequently.

The final property relating to evaluation is the well-known nonlinearity characterizing the value function,

an aspect that is present in virtually all models of evaluation. To illustrate, consider well-replicated research that has indicated how income gains beyond a certain degree (i.e.,  $\approx \$75,000$ ) yield progressively smaller gains in well-being (Kahneman & Deaton, 2010). Such research highlights that objective rewards (i.e., income) might map on subjective values (i.e., affect) in a logistic-like manner, implying that after a certain point, objective wealth accumulates without proportional affective gains. A wealth of evidence has showcased several such nonlinearities in evaluative inferences (see Barberis, 2013; Bruch & Feinberg, 2017), implying that formal models of evaluation need to account for the nonlinear relation between "objective" rewards and the "subjective" (aka "affective") values experienced in response to those rewards.

To summarize, cognitive research on evaluative inference has highlighted at least three central properties that should be captured by a model of this process: the reference point, (un)certainty, and nonlinear nature of the value function. A recent computational model has attempted to synthesize these various properties in a unifying account (see Rigoli, 2019). This "logistic" account of evaluation has also been effectively applied to the field of psychopathology, illustrating how maladaptive evaluative processes could lead to various mental-health problems, such as depression, addiction, and personality problems (see Rigoli et al., 2021). Here, we build on this model to develop a formal framework on MI. Before we do so, however, we first outline the logistic model in its current form.

## Static Evaluation

The logistic model of evaluation (Rigoli, 2019) proposes that the affective value  $V(R)$  of a stimulus is derived from a prediction error between its raw value ( $R$ ) and reference value ( $\mu$ ), moderated by a weight parameter ( $\pi$ ) and filtered through a logistic function:

$$V(R) = \frac{1}{1 + e^{-\pi(R - \mu)}}. \quad (1)$$

Given the logistic nature of Equation 1, an alternative expression is

$$V(R) = \text{logistic}(\pi(R - \mu)), \quad (2)$$

where  $V(R)$  is the subjective/affective value of a given stimulus (which, given the logistic function, is bounded between 0 and 1), ( $R$ ) is the "raw" value of a stimulus (e.g., \$10),  $\mu$  is the reference point (e.g., a desired value \$20), and  $\pi$  is the "weight" attached to a given prediction error (for a detailed outline, see Table 1).



**Table 1.** Outline of Computational Parameters

Parameter name	Parameter symbol	Parameter interpretation	Parameter example
Raw value	$(R)$	The objective (or raw) value of a stimulus	A grade of 95/100
Affective value	$V(R)$	The subjective-affective value of a stimulus	How happy/sad you feel when receiving that grade of 95/100
Reference point	$\mu$	The “standard” for a particular stimulus	A standard (desired) grade of 80/100
Weight	$\pi$	The certainty (weight) of the prediction error	Greater certainty (0.1 vs. 0.001) results in stronger emotional response
Learning rate	$\alpha$	The rate at which the reference changes to align with raw values	Maximal learning ( $\alpha=1$ ) yields a new reference of exactly 95/100

To briefly expand each of these terms, consider first the prediction error:  $R - \mu$ . In a school context, as illustrated by Villano et al. (2020), the raw value of a stimulus ( $R$ ) is a student’s actual grade, and the referenced value of a stimulus ( $\mu$ ) is a student’s desired grade. Villano et al. demonstrated that students’ emotional responses are most robustly predicted by the difference between the students’ desired (referenced) grade and their actual (raw) grade ( $R - \mu$ ), not by the raw grade itself. This naturalistic empirical pattern is consistent with a wealth of laboratory evidence showing that indeed, emotions are the outcomes of expectation-outcome prediction errors (Emanuel & Eldar, 2022). The logistic model leverages these error terms to suggest that the affective valence of an event is dependent on a frame of reference: When the raw value of an experience is higher than one’s reference, reality exceeds expectations ( $R > \mu$ ), and affective valence is positive; when the raw value is lower, reality falls short of expectations ( $R < \mu$ ), and affective valence is negative; finally, when the two are equal, reality meets expectations ( $R = \mu$ ), and affective valence is neutral.

Note that the reference point can take various interpretations depending on a researcher’s theoretical tradition or empirical goals. In economics, for example, the reference point has traditionally been viewed as a rational expectation of a future reward (i.e., “Based on all available information, I rationally predict to receive a certain reward”; Muth, 1961). In reinforcement learning, however, the reference reflects a person’s subjective expectation of a future reward based on previous rewards (i.e., “Based on what I have encountered before, I subjectively expect to get a particular reward”; Sutton & Barto, 2020). Our model builds on this latter perspective by viewing the reference point as a person’s standard for rewards (i.e., “I want this” rather than “I rationally expect this”). We believe that this interpretation is more psychologically accurate because it naturally accounts for everyday scenarios whereby a person may still feel sadness even in response to rationally predictable negative events—for instance, losing a

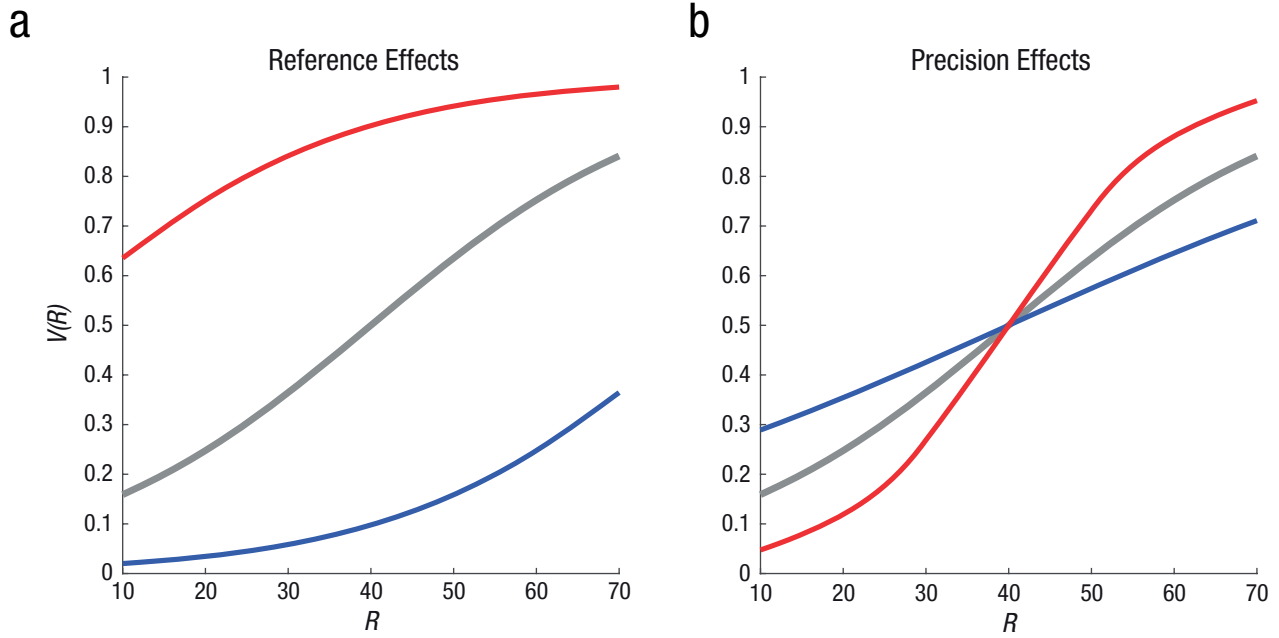
loved one from a terminal illness, which is something that one might have (rationally) predicted but still not wish to have happened.

Beyond such reference effects, the second aspect of the logistic model is the weight parameter,  $\pi$ , which either inflates or deflates the prediction error. To illustrate, consider two students who have the same standard for a grade ( $\mu = 70/100$ ) but who both receive a much lower grade ( $R = 30/100$ ), resulting in a negative emotion of “disappointment” ( $R - \mu = 30 - 70 = -40$ ). The weight parameter controls how much each student will “weigh” (i.e., ascribe meaning to) this feeling of “disappointment”: Although a student who weakly weighs this event ( $\pi = 0.001$ ) will not be much upset ( $0.001 \times (-40) = -0.04$ ), the student who strongly weighs it ( $\pi = 0.1$ ) will be devastated ( $0.1 \times (-40) = -4$ ). In that sense, the weight parameter shapes the intensity of affect by determining how strongly its prediction errors will be weighted. As argued later, this weight parameter can be interpreted as one’s certainty over the meaning of a prediction error: The more certain a person is regarding what the prediction error means, the stronger the person’s affective response will be.

The final key feature of the logistic model is its logistic function, which bounds the affective value of a stimulus between 0 and 1. This function implies that affective values below 0.5 are negatively valenced, those above 0.5 are positively valenced, and those equal to 0.5 are neutral. These outcomes emerge because negative values in a logistic function (here, negative prediction errors) yield  $V(R) < 0.5$ , positive values yield  $V(R) > 0.5$ , and values of 0 yield  $V(R) = 0.5$ . This function is a key aspect of this model because it accounts for the nonlinear relationship between objective rewards and subjective values, as evinced by a wealth of cognitive research on this topic (Rigoli, 2019).

## Psychopathology

Having introduced the logistic model of evaluation, we now turn to its application in psychopathology, as



**Fig. 1.** (a) Maladaptive evaluation based on reference points. Blue line illustrates depressive evaluation (high reference), which yields consistently low affective values,  $V(R)$ , even at high raw values,  $R$ ; red line illustrates manic evaluation (low reference), which yields consistently high affective values,  $V(R)$ , even at low raw values,  $R$ . (b) Maladaptive evaluation based on the weight parameter. Blue line illustrates apathetic evaluation (low weight), which yields similar affective values,  $V(R)$ , across various raw values,  $R$ ; red line illustrates reactive evaluation (high weight), which yields a wide range of affective values,  $V(R)$ , across various raw values,  $R$ .

exemplified by the work of Rigoli et al. (2021). The general idea around “adaptive evaluation” is that mental health ensues when an agent’s evaluation is based on parameters that reflect the true statistics of the agent’s environment: specifically, when the agent’s reference ( $\mu$ ) reflects the average of the agent’s environment’s rewards and the agent’s weight ( $\pi$ ) reflects the inverse of standard deviation (i.e., the consistency) of these rewards.<sup>2</sup> By contrast, psychopathology ensues when either parameter deviates markedly from the true environment statistics. Thus, the logistic model suggests that various psychopathologies can be explained by distinct alterations in these two evaluative parameters.

To illustrate, consider an agent who dwells in an environment offering the following life outcomes with equal probability:  $R = \{10, 20, 30, 40, 50, 60, 70\}$ . (Note that the average and precision [defined as  $1/SD$ ]<sup>2</sup> of these outcomes are  $\mu = 40$  and  $\pi = 0.05$ , respectively.) If the agent’s evaluation relies on these true context statistics ( $\mu = 40$ ,  $\pi = 0.05$ ), then the agent’s evaluation is deemed “adaptive,” resulting in the following affective values:  $V(R) = \{0.1824, 0.2689, 0.3775, 0.5, 0.6225, 0.7311, 0.8176\}$  (see Fig. 1). The reason these values are adaptive (or “healthy”) is because in accordance with the agent’s environment,  $R$  (raw) values below  $\mu = 40$  are perceived negatively ( $V(R) < 0.5$ ),  $R$  values above  $\mu = 40$  are perceived positively ( $V(R) > 0.5$ ), and  $R$  values equal to  $\mu = 40$  are perceived as “neutral” ( $V(R) = 0.5$ ).

Consider now the case of another agent who, while dwelling in the same environment, embodies a reference point radically higher than the average of the agent’s surrounding rewards (e.g.,  $\mu = 80$  rather than  $\mu = 40$ ). In this case, the reference point is higher than all life outcomes ( $\mu > R$ ), casting all such outcomes in a negative light. This scenario has been argued to explain depression and addiction, whereby a high reference point—that could be expressed either psychologically (in terms of an unrealistic standard for rewards) or neurobiologically (in terms of an altered neuromodulatory set point)—could result in persistently negative affect: Most stimuli pale in comparison with the high reference and are thereby experienced in a negative manner (see Rigoli, 2022b; Rigoli & Martinelli, 2021; Rigoli et al., 2021). The exact opposite scenario has been proposed to reflect (hypo)mania: If an agent’s reference point is smaller than the average of the agent’s contextual distribution (e.g.,  $\mu = 0$  rather than  $\mu = 40$ ), then most (and in this case, all) stimuli will be experienced positively, yielding persistently inflated affect (Rigoli et al., 2021; see Fig. 1). In this sense, the reference point determines affective valence: Low and high reference points typically yield positive and negative emotions, respectively (e.g., Clark et al., 2018).

In turn, the weight parameter ( $\pi$ ) shapes the intensity of these experiences, giving rise to other psychopathologies. In particular, when the weight is low (e.g.,

$\pi = 0.025$  rather than  $\pi = 0.05$ ), emotional experience is less intense: less rewarding if positive and less punishing if negative. This situation has been argued to parallel the clinical condition of apathy, a negative symptom of psychosis and related pathologies, in which all stimuli are experienced as equally rewarding and punishing, resulting in lower motivation to pursue any one of them (see Rigoli & Martinelli, 2023). Conversely, when the weight is high (e.g.,  $\pi = 0.1$  rather than  $\pi = 0.05$ ), intense emotionality (aka, emotional reactivity) emerges. An example here could be the strong emotional reactions that individuals with borderline personality typically display even in response to “objectively” mild events (Rigoli et al., 2021; see Fig. 1).

Thus, through aberrant alterations of the reference point, on the one hand, and the weight parameter, on the other, the logistic model can reproduce archetypal patterns of several psychopathologies. But can the same model also shed light on the nature of MI? We argue that as it currently stands, the model is unable to do so. This is because the logistic model assumes that its parameters (i.e., reference and weight parameters) remain fixed over time, thereby neglecting the mechanisms that underpin their evolution. This assumption is problematic when investigating phenomena such as MI, for which, as its very name suggests, dynamical processes are of central interest (e.g., Durstewitz et al., 2021). On this basis, we propose an extension of the logistic model that addresses the question of how evaluative parameters are formed and shaped over time. We then employ this dynamic version to investigate whether any insights on MI could be gained.

## Dynamic Evaluation

Adding a temporal dimension to the logistic model, we propose a dynamic framework of (reference-based) evaluation that is predicated on the following discrete equations:

$$V(R_t) = \text{logistic}(\pi(R_t - \mu_t)) \quad (3)$$

$$\mu_{t+1} = \mu_t + \alpha(R_t - \mu_t), \quad (4)$$

where  $V(R_t)$  refers to the affective value at time  $t$ ,  $R_t$  refers to its raw counterpart at time  $t$ ,  $\mu_t$  and  $\mu_{t+1}$  are the reference points at times  $t$  and  $t + 1$ ,  $\pi$  is the weight parameter, and  $\alpha$  [0,1] is the learning rate (at which the reference point gets updated over time; Table 1).

Equations 3 and 4 describe dynamic evaluation, that is, evaluation over time. Equation 3 is similar to Equation 2; the only difference is its extension over time—whereby, at each time point ( $t$ ), the raw value ( $R_t$ ) of a given stimulus is converted into its subjective-affective

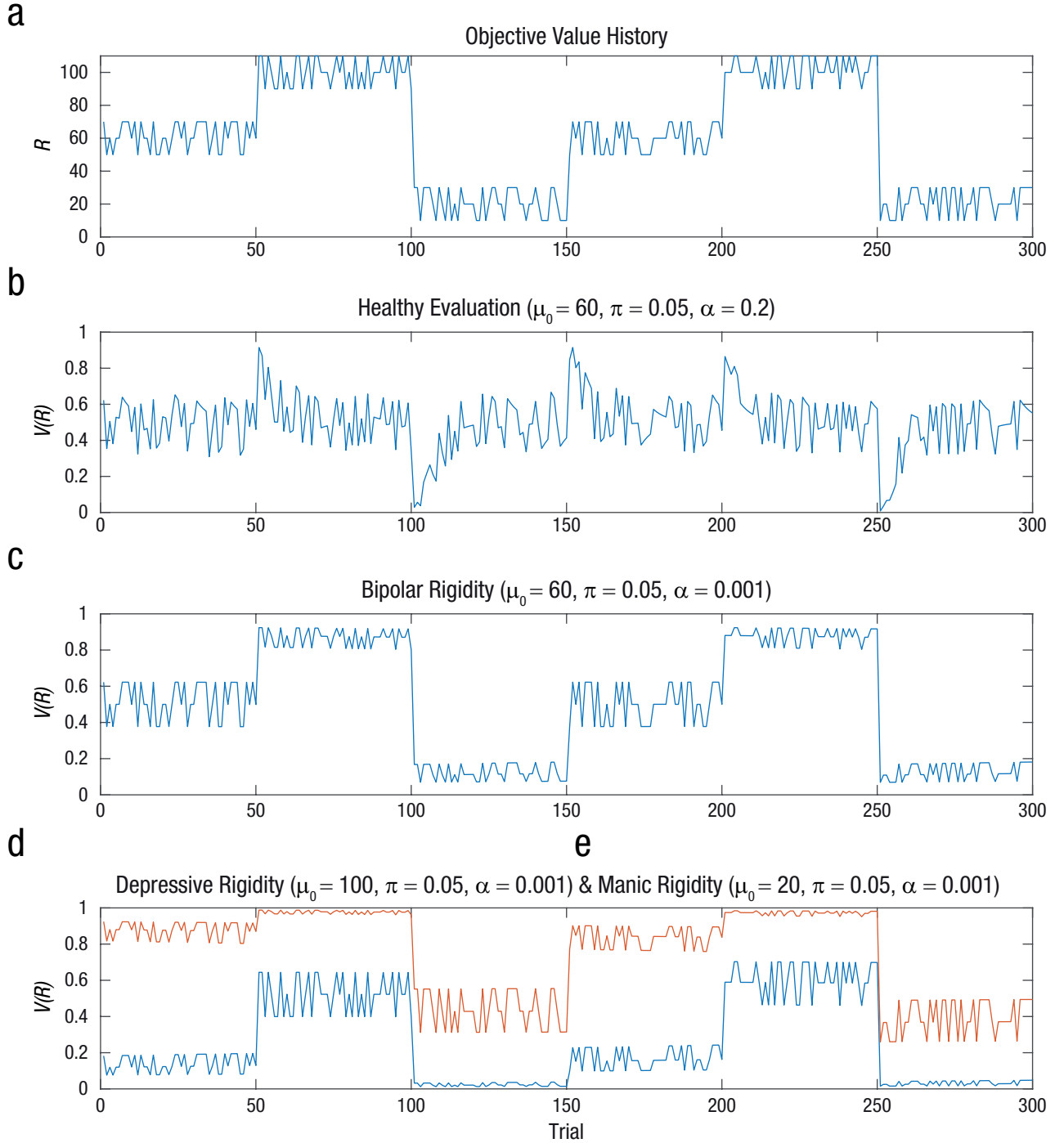
counterpart,  $V(R_t)$ , via the logistic function. Equation 4 implies that the reference point ( $\mu$ ) changes over time based on the Rescorla-Wagner rule, according to which “errors” (i.e., mismatches between one’s reference [ $\mu_t$ ] and reality [ $R_t$ ], at each time point) drive learning about what the reference point should be in the future ( $\mu_{t+1}$ ).

Three parameters define our dynamic model (Table 1). The first is the weight parameter ( $\pi$ ), which, as with the static model, is a constant that determines how strongly agents weigh prediction errors during evaluation. The second parameter, also a constant, is the learning rate ( $\alpha$ ), which determines the extent to which the reference point ( $\mu_t$ ) changes in response to new events ( $R_t$ ); larger values reflect faster learning, that is, faster convergence of a reference with each new raw event:  $\mu_t \rightarrow \mathbb{E}(R_t)$ . The third parameter is the starting reference point ( $\mu_0$ ), which indicates the reference point that an agent entertains at the beginning of a period under scrutiny (e.g., childhood vs. adulthood).

Focusing on these parameters, we sought to examine whether dynamic evaluation could provide any insights on the nature of MI. To do so, we employed an approach akin to the one adopted previously with static evaluation. That is, we assumed mental health to be characterized by moderate parameter values and mental illness to be linked with excessively high or low parameter values. Looking at the consequences of altering these parameters (either in isolation or in combination), we assessed whether the ensuing temporal profiles resemble empirical manifestations of MI. As we detail below, this approach allowed us to generate the three basic forms of MI (inertia/rigidity, instability, and reactivity) from alterations in our model’s parameters (i.e., reference point, learning rate, and weight parameters, respectively).

For our analyses, we simulated a scenario in which on each trial, an agent experiences an objective (or raw) value ( $R_t$ ) that is sampled from one of three distributions, or contexts, which alternate randomly every 50 trials. These include a high-value context (in which the objective value can be 90, 100, or 110 with equal probability), a moderate-value context (in which the objective value can be 50, 60, or 70 with equal probability), and a low-value context (in which the objective value can be 10, 20, or 30 with equal probability). These contexts were built to mimic life periods of prosperity (e.g., a particularly successful period characterized by an enjoyable and well-remunerated job), normality (e.g., a period characterized by somewhat ordinary life experiences), and adversity (e.g., a period characterized by unemployment and anxiety), respectively. The sequence of outcomes ( $R$ ) sampled over time is illustrated in Figures 2a to 4a. The same sequence was used for all simulated agents.

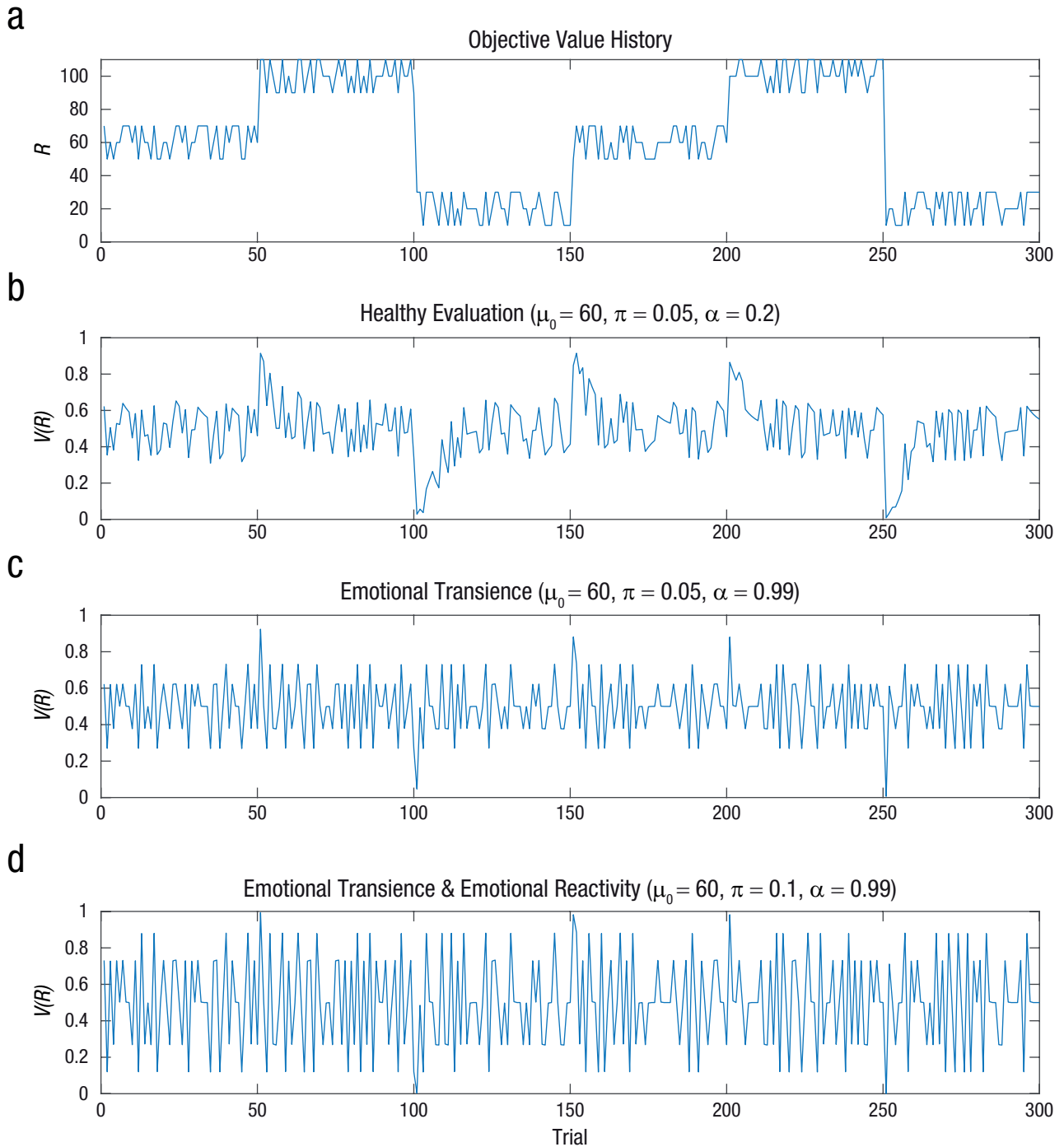




**Fig. 2.** (a) The raw value of stimuli,  $R$ . (b) Healthy evaluation is characterized by strong affects in response to contextual changes but gradual adaptation to each new context. (c) Bipolar rigidity is characterized by periods of inflated (mania) and deflated (depression) affective values, interspersed by moderate affective values (euthymia). (d) Depressive rigidity is characterized by persistently deflated affective values, interspersed by moderate affective values (euthymia) in overly positive life contexts. (e) Manic rigidity is typified by the opposite pattern of depressive rigidity, showing persistently inflated affective values that are interspersed by moderate affective values (euthymia) in overly negative life contexts.

To begin with, we simulate the affective values experienced over time by a “healthy” agent who embodies the following moderate parameter values:  $\mu_0 = 40$ ,  $\pi =$

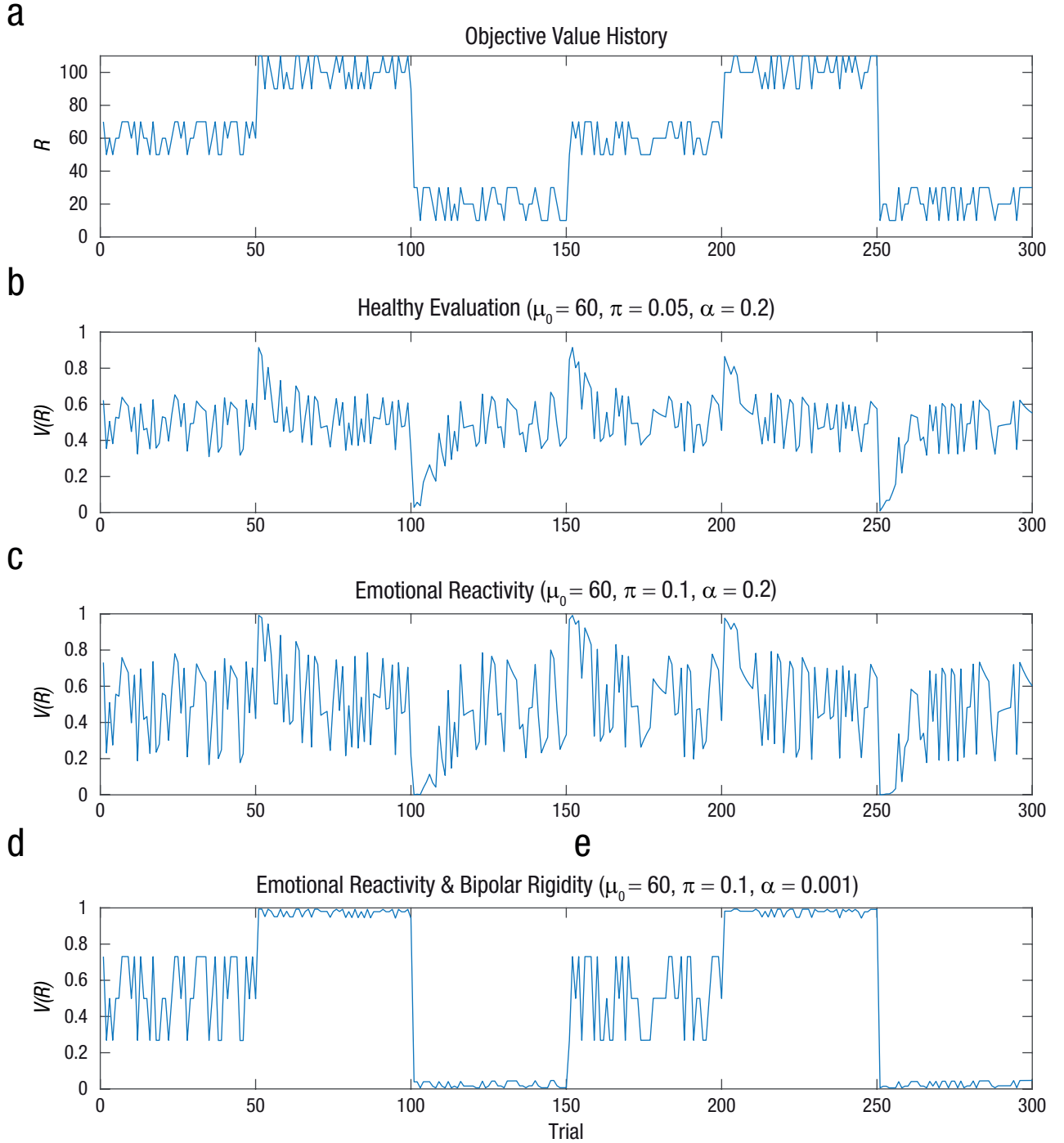
0.05, and  $\alpha = 0.2$ . These parameter values produce what can be referred to as “healthy evaluation” (see Figs. 2b–4b). In Figures 2b through 4b, we show that the



**Fig. 3.** (a) The raw value of stimuli,  $R$ . (b) Healthy evaluation is characterized by strong affects in response to contextual changes but gradual adaptation to each new context. (c) Emotional transience is characterized by fleeting affective values that are inconsistent across similar life events,  $R$ . (d) Combined emotional transience and rigidity are characterized by fleeting, inconsistent, and inflated affective values across all life events,  $R$ .

healthy agent can gradually adapt to new contexts. Specifically, although the agent experiences extremely negative (positive) values when the agent shifts to a worse (better) context, the agent's affective values progressively adapt to the new situation. This pattern of

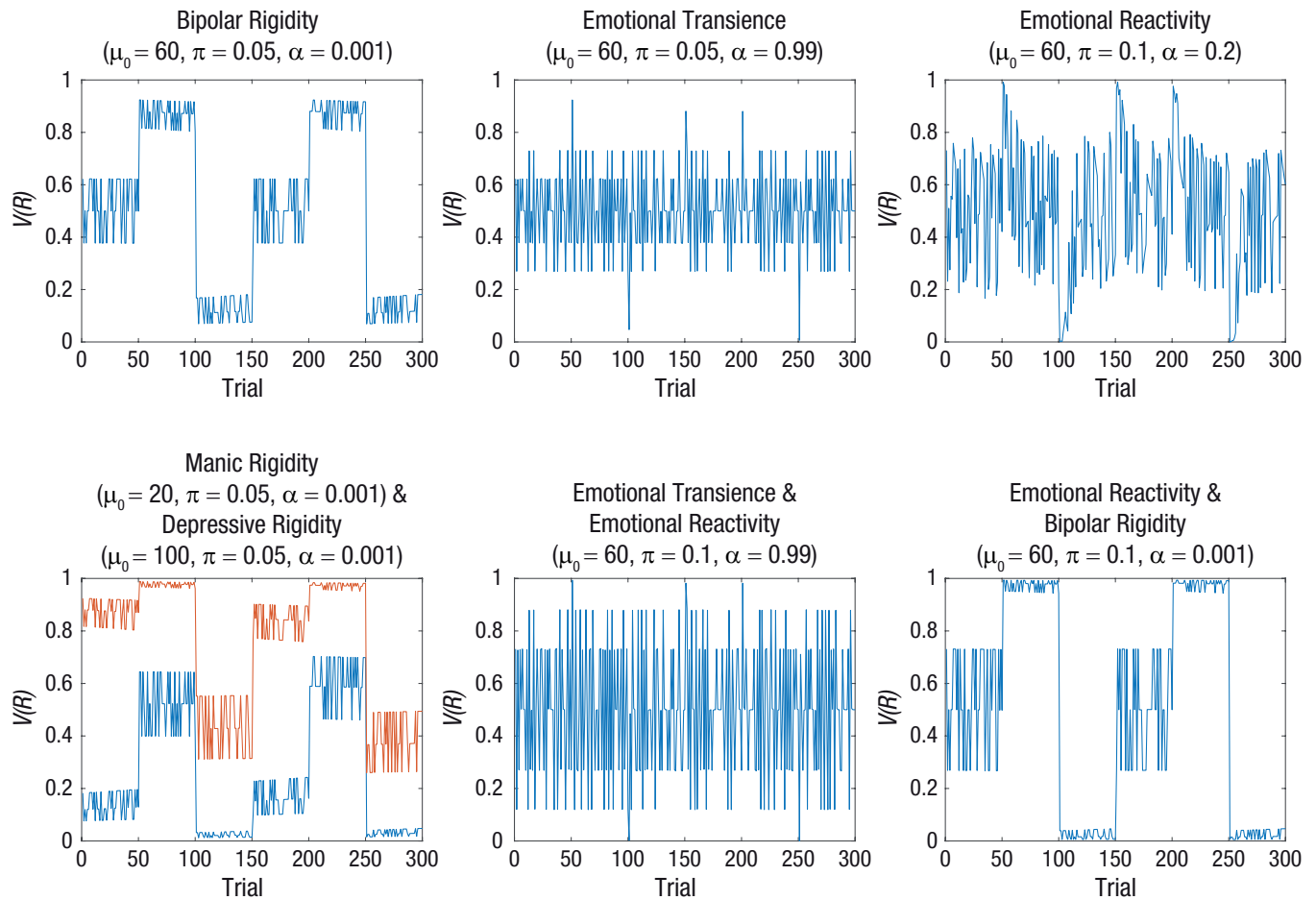
strong response followed by gradual adaptation resembles the affective dynamics that are typically experienced by nonclinical populations. For example, in response to typical adversities (e.g., bereavement) or successes (e.g., job promotion), most people tend to



**Fig. 4.** (a) The raw value of stimuli,  $R$ . (b) Healthy evaluation is characterized by strong affects in response to contextual changes but gradual adaptation to each new context. (c) Emotional reactivity is characterized by inflated affective values across all life events,  $R$ . (d) Combined emotional reactivity and bipolar rigidity are characterized by periods of extremely inflated (mania) and extremely deflated (depression) affective values, interspersed by extremely strong and fluctuating affective values (mixed states).

have strong initial reactions followed by gradual adaptation to the new status quo (Kalisch et al., 2019; Thompson et al., 2010). Our healthy agent reproduces this pattern because of the agent’s moderate parameter

values—particularly in the learning rate ( $\alpha$ ), which allows the agent to gradually adapt to the new context (by flexibly shifting the reference point to the new state of affairs:  $\mu_t \rightarrow \mathbb{E}(R_t)$ ).



**Fig. 5.** All maladaptive affective cases explored in this article: rigidity (including bipolar, manic, and depressive rigidities), transience (including pure transience but also transience that is combined with rigidity), and reactivity (including pure reactivity but also reactivity that is combined with bipolar rigidity).

Using the scenario of healthy evaluation as a baseline, we next explore whether altering any parameters ( $\mu_0$ ,  $\pi$ , or  $\alpha$ ) yields maladaptive consequences that resemble empirical manifestations of MI. As we illustrate below, this procedure allowed us to generate, and thus mathematically integrate, three well-known flavors of MI: emotional rigidity (or inertia), emotional transience (or instability), and emotional reactivity (or sensitivity; see Fig. 5).

### **Emotional rigidity**

First, we consider the case of emotional rigidity (inertia): the tendency of experiencing prolonged emotions, which can be formalized using an extremely low learning rate (here,  $\alpha = 0.001$ ). Below, we detail three scenarios of emotional rigidity: one characterized by a moderate starting reference point ( $\mu_0 = 60$ ), another characterized by an extremely high starting reference point ( $\mu_0 = 100$ ), and a final one characterized by an extremely low starting reference point ( $\mu_0 = 10$ ).

Figure 2c depicts the first of these scenarios, that is, “bipolar rigidity,” which is characterized by a low learning rate ( $\alpha = 0.001$ ) and a moderate starting reference point ( $\mu_0 = 60$ ). Here, the bipolar agent’s behavior is similar to the healthy agent’s behavior during the moderate-value context but radically different during the low- and high-value contexts. Specifically, during the high-value (low-value) context, the bipolar agent experiences persistently positive (negative) affects. Furthermore, during these contexts, affective values get stuck at their extreme ends and oscillate less (e.g., in the low-value context, responses to outcomes 10, 20, and 30 are virtually identical), reflecting a reduced ability to discriminate life experiences.

This scenario bears striking resemblance to manic-depressive illness (or bipolarity). For example, much like our simulation illustrates (Fig. 2c), bipolarity is characterized by severe episodes of mania (or hypomania) and depression that persist temporally (more than a week for depression; more than 4 days for mania) and recur cyclically (Goodwin et al., 2007).

Furthermore, such affective episodes are typically triggered by significant life experiences, such as severe adversities (in the case of depression) and goal-attainment events (in the case of mania; Johnson et al., 2011; Zavlis et al., 2023). Our model captures this pattern as well because it suggests that the mechanism by which manic and depressive episodes emerge, persist, and alternate is through emotionally intense events that signal significant life changes. Finally, consistent with our proposal (that places abnormal evaluative processes at the heart of bipolar disorder), empirical observations suggest that negative and positive appraisals (“I am a failure” and “I am the best,” respectively) can both generate and maintain bipolar episodes (for an integrative review, see Mansell et al., 2007).

Two further notable points can be made from our simulation on bipolarity. First, an implication from our proposal is that during depressive and manic episodes, individuals with bipolar disorder manifest a diminished ability to discriminate among stimuli. In other words, their emotional response to different experiences is predicted to be less differentiated. To our knowledge, this phenomenon has not yet been tested empirically and thus presents a novel prediction of our theory. Second, we note that a puzzling feature of some cases of bipolarity is mixed affective states, that is, states that combine depressive and manic features (Vieta & Valentí, 2013). Although our present simulation does not account for such states, in a subsequent section (Combined Mood States), we extend it so that it does.

A second form of emotional rigidity or inertia (i.e., “depressive rigidity,” as depicted in Fig. 2d) ensues when an extremely low learning rate ( $\alpha = 0.001$ ) is coupled with an extremely high starting reference point ( $\mu_0 = 100$ ). Compared with the healthy agent, this agent exhibits affective values that are persistently depressed: The values are moderate for the high-value context, low for the moderate-value context, and extremely low for the low-value context. Moreover, similar to the previous scenario of bipolarity, certain contexts (i.e., the medium- and low-value ones) are characterized by diminished oscillations in affective values, reflecting a reduced ability to discriminate life experiences.

This scenario is reminiscent of major depression. For instance, similar to what Figure 2d depicts, people with depression tend to persistently experience severely depressed mood (see Rottenberg, 2005; Rottenberg et al., 2005). Crucially, their depressed mood persists even within moderate and positive contexts, reflecting the clinical symptom of “anhedonia” (i.e., the inability to experience pleasure; Loas, 1996). On this, much like our model predicts unvarying responses across varying outcomes, empirical research has shown that depressed patients manifest the same negative affect in response

to stimuli of varying emotional intensities (Rottenberg, 2005; Rottenberg et al., 2005). According to our proposal, this tendency of the depressed agent to gravitate toward low mood can be attributed to the agent’s excessively high reference point, which is resistant to change in light of new experiences (because of the low learning rate; Clark et al., 2018). This high reference point could be interpreted in at least two ways. Neurobiologically, it could be interpreted as an aberrant neuromodulatory set point, explaining the more “endogenous” cases of depression (e.g., high trait neuroticism; Barlow et al., 2021; Clark et al., 2018; Ormel et al., 2013). Psychologically, it might be interpreted as an unattainably high standard (desire) for rewards (see M. M. Smith et al., 2018), explaining the more psychological and “perfectionistic” cases of depression (e.g., those found in narcissism; M. M. Smith et al., 2016).

Our final scenario of emotional rigidity (see Fig. 2d, red line) occurs when both the learning rate ( $\alpha = 0.001$ ) and the reference point ( $\mu_0 = 10$ ) are extremely low. Affective values in this case are persistently inflated: They are moderate for the low-value context, high for the moderate-value context, and extremely high for the high-value context. Moreover, the low- and moderate-value contexts are once again characterized by diminished oscillations in affective values, reflecting a reduced ability to discriminate varying life experiences.

In principle, this scenario depicts a profile of “pure” (aka, unipolar) mania whereby euthymic states alternate with inflated affective states. Yet at the empirical level, there is ongoing debate over the existence of cases of “unipolar mania”; some clinicians have argued that such cases are, in fact, nonexistent (e.g., Solomon et al., 2003; Stokes et al., 2020). On that basis, notwithstanding its theoretical significance, this last scenario of mood rigidity might not be relevant empirically. Still, it raises an interesting question: If the general logic proposed by the model is sound, why does the case of pure mania never occur in practice?

A possibility is that within the population, the learning rate and the reference point are not independent but are negatively correlated. That is, people characterized by lower learning rate might also tend to have higher reference points. This may occur because, for instance, the neural substrate of the learning rate and of the reference point partially overlap, implying some degree of correlation between the two factors. This possibility is compatible with the empirical observation that cases of unipolar mania (defined by our model as having both excessively low reference point and low learning rate) are virtually nonexistent because they are almost always followed by at least one major depressive episode (Solomon et al., 2003). Moreover,



this possibility implies that depression (characterized by an excessively high reference point and an excessively low learning rate) is more prevalent than bipolar disorders (characterized by low reference points and low learning rates), which is indeed the case at the empirical level (Mitchell & Malhi, 2004).

To summarize, the abovementioned mood disorders exhibit a flavor of MI that is characterized by emotional rigidity. In the psychological literature, this rigidity is typically operationalized by autocorrelations that denote the persistence (i.e., inertia) of particular emotions (e.g., Box et al., 2015). Here, we have formalized this phenomenon with our learning parameter, showing how a low learning rate leads to overly rigid evaluations that yield persistently inflated and/or deflated emotions. In that sense, our learning parameter has formalized the concept of inertia by suggesting that it is generated from an inability to adapt to changing life contexts.

### ***Emotional transience***

If emotional rigidity (or inertia) emerges when learning is too slow (rendering agents too rigid), then emotional transience (or instability) emerges when learning is too fast (rendering agents too plastic). Thus, emotional transience can be considered to be the polar opposite of emotional rigidity (because it is predicated on an excessively high, not low, learning rate).

Figure 3c illustrates the case of emotional transience (or instability), which ensues when learning is excessively high ( $\alpha = 0.99$ ). Comparing this scenario of transience against the one of health, we find that two key differences emerge. First, although the healthy agent requires some trials to adapt to a new context (i.e., the agent's responses are initially extreme and gradually adapt), the transient agent adapts instantly (after a single trial). Second, once adaptation has ensued, the healthy agent manifests a consistent response to the same outcome; for instance, the healthy agent always elicit the same (neutral) response of  $V(R) \approx 0.5$  to the objective outcome  $R_t = 60$ . By contrast, the responses of the transient agent appear to be inconsistent; for example, the same outcome  $R_t = 60$  is sometimes experienced positively ( $V(R) > 0.5$ ; Trial 3) yet other times experienced negatively ( $V(R) < 0.5$ ; Trial 50). These evaluative inconsistencies are due to the excessive learning rate ( $\alpha = 0.99$ ), which renders the agent's reference point almost equivalent to the agent's most recent experience ( $\mu_{t+1} \approx R_t$ ). Thus, the transient agent's reference is in constant flux, producing inconsistent responses to the same outcomes.

This pattern of evaluative inconsistency and emotional transience is reminiscent of the pervasive instability (in emotions, relations, and self-perceptions) that

typifies BPD. For instance, in line with our simulation, people with BPD are known to be “a function of their present situations” (Hochschild Tolpin, et al., 2004, p. 112), frequently changing their personal tastes and preferences (e.g., their core values, goals, and friends) in accordance to their most recent experiences. Moreover, experimental evidence suggests that during mind-wandering tasks, people with BPD are more likely than healthy control subjects to generate inconsistent and extreme evaluations about themselves and others (Kanske et al., 2016). Note that such evaluations appear to vary quickly, in line with our notion of “transiency” (Kanske et al., 2016). Finally, evaluative inconsistencies have long been theorized to be central to “self-instability” (Kaufman & Meddaoui, 2021), an aspect of BPD that was recently demonstrated to account for its remarkable affective instability (Kockler et al., 2022; P. S. Santangelo et al., 2017, 2020).

Our model formalizes this self-instability by casting it as a form of evaluative inconsistency: the rapid changing of reference points by which one evaluates life events. To illustrate, consider the situation in which a seemingly “perfect” romantic date ( $R_1 = 100$ ) sets up a high “standard” for another one ( $\mu_2 \approx R_1$ ), only to lead to a strong disappointment (or even a “fear of abandonment”) in a BPD agent when it is merely rescheduled by the other party ( $R_2 = 40 \ll \mu_2 \approx 100$ ). Indeed, in line with this formulation, prominent treatment protocols for BPD, such as dialectical-behavior therapy, focus on diverting patients away from such “ruminative” states (regarding less-than-ideal outcomes) and toward states of “acceptance” (that could be interpreted as using more consistent reference points when evaluating sub-optimal situations; Dimeff & Linehan, 2001). Our model is consistent with these psychotherapeutic approaches and provides a mathematical intuition of why they might be successful in treating conditions characterized by transient emotions and self-perceptions (including not only BPD but also its frequent comorbidities, such as eating pathologies; Linehan & Chen, 2005).

To summarize, emotional transience reflects fleeting, rather than persisting, emotionality. Also known as “emotional instability,” this type is usually operationalized in terms of high variability and low autocorrelation of mood time series (Dejonckheere et al., 2019). Note that these fleeting dynamics were shown to be the polar opposite of those that typify emotional rigidity (Thompson et al., 2012), in keeping with our model that places these types on opposite ends of a learning-rate spectrum. Our model clarifies the nature of these concepts by suggesting that rigidity can be generated from overly consistent evaluations (rooted in an inability to adapt to changing life conditions), whereas transience can be generated from overly inconsistent

evaluations (rooted in the tendency to overidentify with changing life conditions; Deutsch, 1942).

### **Emotional reactivity**

Our final type of MI is generated from an inflated weight parameter ( $\pi$ ). When this occurs, agents become emotionally reactive: They experience inflated affects because they overweigh the “meaning” of their life experiences.

These patterns of reactivity are illustrated in Figure 4c, in which all parameters are set to a moderate and adaptive range ( $\mu_0 = 60$ ,  $\alpha = 0.2$ ) except the weight parameter, which is set to be high and maladaptive ( $\pi = 0.1$ ). Comparing this scenario of reactivity to the one of health (Fig. 4b), we find that one key similarity emerges (because of the adaptive learning rate): Affective values are more extreme when the context shifts and progressively adapt to the new status quo. Nevertheless, one key difference is also evident: The reactive agent exhibits greater oscillations in affective values, reflecting an enhanced reactivity to life outcomes. Formally, these inflated life outcomes occur because the high weight inflates the discrepancy between one’s reference and reality ( $R_t - \mu_t$ ), magnifying their affective values  $V(R_t)$ .

We argue that this magnification of affective values reflects several “reactive” temperaments that are typified by an overinterpretation of life events. For instance, the so-called “narcissistic rage,” which entails strong and inappropriate anger in the face of minimal provocation, can be considered a form of emotional reactivity (Krizan & Johar, 2015). Likewise, “defensive aggression,” which is typical of antisocial personalities, is a reactive form of aggression that occurs in response to perceived provocation (Vaidyanathan et al., 2011). Finally, contemporary classification systems describe a “marked reactivity in mood” as a key symptom of BPD (see American Psychiatric Association, 2013, p. 663), suggesting that individuals with borderline personality can also exhibit emotional reactivity (beyond the previously mentioned emotional transiency).

Note that our model pinpoints these reactive temperaments to a computational parameter ( $\pi$ ) that has a clear and formal interpretation: that is, overweighing the meaning of life experiences, which results in overconfident inferences (e.g., “I am sure you hate me”). Our definition for this parameter mirrors that of “hypermentalizing,” which has been defined as the overcertainty regarding the meaning of social experiences when there is minimal or no objective data to support such inferences (Sharp, 2014). Accordingly, both emotional reactivity in general and hypermentalizing in particular have been reported in a number of psychiatric disorders, including psychosis (Myin-Germeys,

Krabbendam, et al., 2003; Myin-Germeys, Peeters, et al., 2003), bipolar disorder (Henry et al., 2007), and attention-deficit/hyperactivity disorder (Isaksson et al., 2019; for a meta-analysis, see McLaren et al., 2022). Our model is in line with this evidence and provides a formal interpretation of why hypermentalizing (or overcertain) states may lead to emotional reactivity.

The clinical conditions just mentioned manifest excessive reactivity particularly in the domain of negative affect. Regarding the domain of positive affect, evidence suggests that various facets of extraversion (e.g., enthusiasm or sensation-seeking) can be interpreted in terms of an enhanced responsivity to reward (Lucas et al., 2000; Smillie, 2013). Note that when extreme, these personality facets can promote “impulsivity” (Revelle, 1997), a symptom found in a plethora of psychiatric conditions, including bipolar disorder, personality disorder, and substance abuse, to name a few (Moeller et al., 2001). Although impulsivity is a multidimensional trait (comprising emotional, motivational, and cognitive components), one of its central facets includes the “rapid and unplanned reaction to positive internal or external stimuli” (see Moeller et al., 2001, p. 1784). Taken from this perspective, our “positively valenced” reactivity could be regarded as one among many mechanisms that promote impulsivity (see Strickland & Johnson, 2021).

In summary, these patterns illustrate a type of MI that is characterized by strong emotional reactivity, sometimes also referred as “stress reactivity” or “emotional sensitivity” (Marwaha et al., 2014). Our model clarifies the nature of this MI pattern by suggesting that it occurs when the salience of stimuli is inflated. In this view, the notion of reactivity is distinct from its close relative, transience, because the former can be casted as a form of evaluative certainty (rooted in a trait of sensitivity; Linehan, 1987) and the latter can be casted as a form of evaluative inconsistency (rooted in a trait of self-instability; Deutsch, 1942).

### **Combined mood states**

So far, we have explored each MI type in isolation. However, our three MI types can also be considered in combination in the sense that they co-occur within the same person (with the exception of emotional rigidity and transience, which are polar opposites). In this subsection, we briefly turn to two such combinations: emotional reactivity with either bipolar rigidity (to explain “mixed affective states”) or emotional transience (to explain “severe BPD”).

First, consider the scenario of combined bipolar rigidity and emotional reactivity (Fig. 4d), which ensues when a low learning rate ( $\alpha = .001$ ) and a moderate reference point ( $\mu_0 = 60$ ) are paired with a high weight

parameter ( $\pi = 0.1$ ). Given  $\alpha = .001$  and ( $\mu_0 = 60$ ), this scenario reproduces the previous case of bipolar rigidity: Affective values are extremely positive/negative during high-/low-value contexts, reflecting mania/depression, respectively. However, because of the strong weight parameter ( $\pi = 0.1$ ), another pattern is now evident: Affective values oscillate greatly within the moderate context, reflecting high reactivity to all moderate stimuli. Thus, states of depression and mania here are interjected by another state of extreme reactivity to moderate stimuli.

We suggest that this extreme reactivity to varying stimuli may reflect what clinicians have termed as “mixed states.” Although the definition of such states has always been hazy, research in the past decade has demonstrated that one of their defining features is “irritability” (Vieta & Valentí, 2013, p. 33), which “includes a feeling that one’s emotional responses are unjustified or disproportionate to the immediate source” (Barata et al., 2016, p. 170). Indeed, both clinical observations and psychometric analyses have converged on similar models that place emotional responsivity, not valence, at the heart of mixed states (Henry et al., 2003, 2010; Pacchiarotti et al., 2013; Perugi & Akiskal, 2005; Swann et al., 2013), leading to the conclusion that “mixed states may be better defined by emotional hyper-reactivity, meaning that patients feel emotions with a greater intensity and depending on the environmental context” (Henry et al., 2007, p. 37). Our model is remarkably consistent with these perspectives because it suggests that mixed states can be formalized as fast oscillations between extreme emotions—oscillations that occur in response to varying events. Our model is also consistent with evidence suggesting that mixed states are linked to worse clinical outcomes (Swann et al., 2009, 2013; Vieta & Valentí, 2013) by revealing that depressive and manic episodes are magnified in bipolar agents who exhibit mixed states (see Fig. 4d).

Our other case of mixed affective states is one in which a high learning rate ( $\alpha = 0.99$ ) and a moderate starting reference ( $\mu_0 = 60$ ) are paired with a high weight parameter ( $\pi = 0.1$ ) to produce an amalgam of emotional transience and reactivity, which exhibits by far the most unstable affective dynamics (Fig. 3d). In brief, this picture depicts an agent who experiences both brief emotions (because of inconsistent evaluations) and extreme emotions (because of reactive evaluations). Note that this fast-paced instability occurs across all contexts, suggesting that this agent is similarly unstable across all life situations.

This scenario of fast-paced instability mirrors severe borderline psychopathology. Indeed, researchers sometimes distinguish between two types of BPD based on severity. The less severe type (so-called “as-if” character or “quiet” BPD) is characterized by an unstable identity,

internalizing psychopathology, and coping only through intense attachment to others (Deutsch, 1942; Sherwood & Cohen, 1994). Conversely, the more severe type is typified by externalizing psychopathology, strong emotional reactivity, and impulsivity (Arntz et al., 2003; Bales et al., 2012). Our model is in line with these perspectives because it distinguishes emotional transience from reactivity, placing the former on a spectrum of mood transience (in line with the less-severe type) and the latter on separate spectrum of personality pathology (in line with the more-severe type). Our model further implies that the combination of both types (i.e., both transience and reactivity) yields the most unstable BPD profile, in line with empirical observations showing that severe cases of BPD exhibit both internalizing and externalizing psychopathology (Gunderson et al., 2018).

To summarize, in this section, we have shown how emotional reactivity can combine with bipolar rigidity and borderline transiency to exacerbate mood problems: specifically, by generating patterns that resemble mixed episodes (in bipolar disorder) and extreme emotional oscillations (in BPD). In the next and final section, we outline more formally the main predictions of our theory.

## Model Predictions

In this final section, we outline the core predictions of our theory and illustrate how they can help advance the theoretical status of MI regardless of whether they prove to be right. In doing so, we focus on a brief exposition of each prediction and its corresponding empirical examination (aimed to corroborate or refute the prediction). For more details on how to test each prediction, interested readers are referred to the Supplemental Material available online.

Our model and its overarching theory make at least three novel predictions. The first prediction is that mood valence will be predicated entirely on prediction errors. Although this prediction is inspired by past reinforcement-learning research, we note that it forms one of the primary tests of our model for the following reason: Our model assumes that neutral affective experiences are independent of the objective value of a reward. Thus, a student who expects a high grade of 99/100 and receives exactly that high grade of 99/100 will be just as “satisfied” as the student who expects a low grade of 10/100 and receives exactly that low grade of 10/100. Although this is a rather strong and novel assumption, we believe it is worth examining for the following reason: If it is supported, it would provide evidence in favor of a simple prediction-error way of understanding mood valence; if it is refuted, however, it would offer preliminary evidence for a more expanded

way of understanding mood valence (one whereby valence is determined by both prediction errors and raw value of rewards). In the Supplemental Material, we illustrate in more detail how researchers can test this prediction of our model by pinning two alternative, nested models against each other, yielding novel insights on mood valence regardless of which model is supported.

The second prediction of our model is clinical and suggests that our MI parameters will robustly predict specific mental-health problems. Although one way to test this hypothesis is with a case-control design (whereby our formal parameters are compared in a discrete manner across clinical and nonclinical samples), we believe that a dimensional approach is more robust, especially in light of recent findings implying that computational parameters are normally distributed and can predict mental-health problems across clinical and nonclinical samples (e.g., Story et al., 2024). On that basis, we suggest that a better test of our clinical predictions is to (a) recruit a diverse general-population sample (or a transdiagnostic clinical sample) who will complete a (b) set of self-reported scales (measuring psychopathology symptoms) and (b) a relevant monetary-reward experiment (see the Supplemental Material) to derive our computational parameters. Hypotheses in this setting would be as follows:

*Hypothesis 1:* Reference points will be positively associated with perfectionism, narcissism, and neuroticism (or depressed mood), highlighting that people with these problems may have higher standards for reward.

*Hypothesis 2:* Reference points will be negatively associated with (hypo-)mania.

*Hypothesis 3:* Learning rates will be positively associated with identity disturbance, suggesting that emotional instability in “borderline personality” may be rooted in self-instability.

*Hypothesis 4:* Learning rates will be negatively associated with anhedonia and depressed mood, suggesting that rigid mood states are based on an inability to adapt to different settings.

*Hypothesis 5:* Weight parameters will be positively associated with antagonism in relational pathology and aberrant salience in psychosis, formalizing their strong reactivity to external stimuli.

*Hypothesis 6:* Weight parameters will be negatively associated with apathy, formalizing its nonreactivity to external stimuli.

In our Supplemental Material, we flesh out the details of these hypotheses and comment specifically on the theoretical value that could be gained if they are supported or refuted by future research.

Finally, we highlight one last and rather tentative prediction: that is, that our computational parameters will also be associated with particular “longitudinal affects.” To elaborate, our model makes the following exploratory but clear predictions:

*Prediction 1:* Learning rates,  $\alpha$ , will be negatively related to measures of mood rigidity, such as high affect autocorrelations.

*Prediction 2:* Learning rates,  $\alpha$ , will be positively related to measures of mood transience, such as the mean square of successive differences.

*Prediction 3:* Weight parameters,  $\pi$ , will be positively related to measures of mood reactivity, such as regression slopes of how life stressors predict certain affects.

Although these predictions are explicit, we emphasize that we view them mostly as exploratory and hypothesis-generating because at the time of writing, there are several conceptual and methodological uncertainties surrounding the fields of both computational and longitudinal modeling. Specifically, there is limited evidence that cognitive mechanisms from controlled laboratory settings (e.g., “learning rates” or “sensitivity parameters”) map on mood dynamics from real-life settings (Bennett et al., 2019; Huys et al., 2021; Karvelis et al., 2023; Zavlis et al., 2025b). Likewise, several methodological concerns, such as limited reliability and validity in both experimental (Karvelis et al., 2023; Zavlis et al., 2025b) and longitudinal research (Jahng et al., 2008; Trull & Ebner-Priemer, 2013), limit the power of a thorough test of our model. In that sense, we believe that our hypotheses above are best regarded not as conclusive tests of our formal model but rather as scaffolds for an agenda of translational research that aims to enhance both the computational and longitudinal fields by integrating lab-based computational findings with real-life mood dynamics (see Discussion).

To summarize, we have outlined two confirmatory predictions (relating to mood valence and instability patterns) and one exploratory prediction (relating to the integration of computational and longitudinal lines of research). Although we believe that these hypotheses are plausible given prior computational work (e.g., Emanuel & Eldar, 2022), we note that even if they are refuted, they could still point to notable insights regarding the most likely mechanisms that generate disparate mood dynamics (for details, see the Supplemental Material).

## Discussion

In this article, we have outlined a formal theory of MI that can generate three distinct MI types: emotional rigidity (the tendency to experience prolonged emotions),



emotional transiency (the tendency to experience fleeting emotions), and emotional reactivity (the tendency to experience strong emotions; see Fig. 5). In the discussion that follows, we explore the contributions of this computational perspective on the theory, clinical utility, and measurement of MI before concluding on current limitations and future directions for research in this line of inquiry.

## Theory

To begin with, we note that our formal perspective could help address existing debates on the nature of MI by integrating disparate theoretical viewpoints. To date, much ink has been spilled on theoretical debates around MI, including whether it contains various subtypes (Hamaker et al., 2015), to what extent these subtypes are pathognomic or transdiagnostic (Trull et al., 2015; Tsanas et al., 2016), and how they could best be defined (Broome, He, et al., 2015; Broome, Saunders, et al., 2015). Our formal theory could help address these debates by showing how three well-known MI types (concerning rigid, transient, and strong emotions) can be generated from a single computational process: the process of evaluating stimuli. In so doing, our theory provides a formal way of thinking about different forms of MI and a pathway toward integrating longitudinal and experimental approaches to mood dynamics.

To briefly elaborate, our theory outlines how three central concepts from longitudinal research (rigidity/inertia, transience/instability, and reactivity) can be generated from three cognitive parameters (reference point, learning rate, and weight parameters), providing a blueprint for how the disparate lines of longitudinal versus cognitive research could be integrated. For example, reward-based paradigms (which are typically employed in experimental settings to quantify cognitive parameters from behaviors) could be employed to measure the learning rate (how fast participants adjust their expectations of future rewards), weight parameter (how strongly participants weigh prediction errors), and reference point (for reviews, see Hitchcock et al., 2022; Huys et al., 2021; Zavlis et al., 2025a,b). In turn, these cognitive parameters can be linked to corresponding longitudinal parameters, examining whether rigidity-inertia can be understood as a form of evaluative consistency (based on a slow learning rate), transience-instability can be understood as a form of evaluative inconsistency (based on a fast learning rate), and reactivity can be understood as a form of evaluative certainty (based on an increased weight parameter). So far, research in computational psychiatry has investigated evaluative processes in cross-sectional research, revealing notable findings, including that individuals

with major depression tend to evaluate disparate stimuli in a consistently pessimistic way (implying a low reference point; Kube, 2023) and that individuals with BPD and psychosis tend to overweigh the value of social stimuli (implying a higher weight parameter; Henco et al., 2020). Adding a temporal dimension to these analyses could enhance ecological validity by showing how aberrant computational processes are linked to mood outcomes in naturalistic settings—and how they ameliorate in therapeutic settings (Hitchcock et al., 2022; Karvelis et al., 2023; Moutoussis et al., 2018; Zavlis et al., 2025a,b).

Beyond elucidating the links between experimental and longitudinal parameters, our formal framework was able to generate novel predictions (that were not a priori specified). For example, our simulations indicated that manic/depressive episodes will be typified by a reduced ability to discriminate positive/negative stimuli, providing a novel hypothesis that was not prespecified but is still in line with existing evidence on mood disorders (see Kube, 2023). Likewise, our model predicted that severe borderline psychopathology will be typified by both transience and reactivity, mirroring traditional theories on the topic (e.g., O. Kernberg, 1967; Knight, 1953) that have not yet been precisely examined (Zavlis et al., 2025a). These hypotheses are notable because they emerged without a priori specification, highlighting the utility of formal, as opposed to verbal, theorizing in generating precise, quantifiable, and undogmatic predictions that can be more explicitly examined against empirical observations (see Fried, 2020; Haslbeck, Ryan, et al., 2021; Lewandowsky & Farrell, 2011; Robinaugh et al., 2021; Smaldino, 2020). In our Supplemental Material, we expand on these hypotheses further, commenting on how they could be tested such that they are informative regardless of whether they prove to be “right” (Nosek et al., 2019).

## Clinical practice

Beyond theory, our formal framework may also prove useful to clinical work. For example, our framework clarifies the clinical underpinnings of various maladaptive emotions by suggesting that rigid emotions stem from overly consistent evaluations (“I am worthless”; “I am incredible”), transient emotions stem from overly inconsistent evaluations (“I hate you” → “I love you”), and strong emotions stem from overly precise evaluations (“I am sure you hate me!”). Transient and rigid emotions were generated from the opposite ends of a single computational parameter, implying that they may form a continuum of internalizing psychopathology that includes emotionally rigid disorders on the one end (e.g., unipolar vs. bipolar depression) and emotionally



unstable disorders on the other (e.g., borderline personality; Barlow et al., 2021; MacKinnon & Pies, 2006; Paris et al., 2007; Tyrer, 2009). Of course, it is also possible that both sets of disorders exhibit the third type of MI (emotional reactivity), complicating diagnostic practice (e.g., the presence of emotional reactivity in bipolar cases leading to their misdiagnosis as “personality disorder” cases; Morgan & Zimmerman, 2015; Zimmerman et al., 2010). Our framework is consistent with dimensional classification systems and suggests that the continuum of emotional transience-rigidity might belong to the internalizing spectrum, whereas that of emotional reactivity may be more characteristic of the externalizing spectrum.

This dichotomy may prove fruitful in elucidating the nature of personality psychopathology by parsing it into two types: an internalizing one (typified by an inconsistency in evaluating oneself, which yields emotional transience; Deutsch, 1942) and an externalizing one (typified by an overcertainty when evaluating social stimuli, which yields emotional reactivity; Linehan, 1987). Currently, these two types are conglomerated within the general concept of “personality functioning,” which is associated with the entire tapestry of personality, including its more internalizing and externalizing aspects (Hopwood, 2025; Kerber et al., 2024). Our theory is consistent with traditional and recent perspectives that have called for the delineation of these two aspects (and the move away from the stigmatizing label “personality disorder”) by suggesting that to the extent that someone’s “personality” problems concern inconsistent evaluations of “the self,” they may be more appropriately considered as “internalizing mood problems” (Barlow et al., 2021; Bowen et al., 2012; Ellard et al., 2010; Ormel et al., 2013), whereas to the extent that they concern reactive evaluations of others, they may be better considered as “externalizing social problems” (Hopwood, 2024; Wright et al., 2022; Zavlis, 2023, 2024b; Zavlis et al., 2025a).

Beyond diagnostic matters, we note that several predictions of our framework align with existing evidence on therapeutic processes, implying that our parameters could be leveraged to formalize mechanisms of therapeutic change. For example, our theoretical prediction that rigid mood disorders (i.e., depression) ameliorate with increased learning of positive reference points may formalize recent theories suggesting that antidepressants work by enhancing neuroplasticity for positive information (Page et al., 2024). Likewise, our prediction that transient emotions ameliorate with more consistent evaluations could formalize the evidence-based observation that borderline instability subsides when patients adopt more integrated and consistent ways (i.e., reference points) of viewing themselves and others (O. F. Kernberg

et al., 2008). Finally, our conjecture that reactive emotions subside when patients adopt more uncertain evaluations could formalize the idea that hyper-mentalizing states ameliorate when patients are encouraged to engage in more “nuanced” and “uncertain” mentalizing (Bateman & Fonagy, 2016). Together, these patterns illustrate that it might be possible to link well-defined computational parameters to therapeutic processes, examining them based on experimental paradigms from the nascent field of computational psychotherapy (see Moutoussis et al., 2018; R. Smith et al., 2020).

## Measurement

This brings us to the final possible contribution of our theory: the longitudinal measurement of mood dynamics. Contemporary research on mood dynamics is primarily based on “clock time,” examining how emotions vary either within days or across days/weeks (e.g., Dejonckheere et al., 2019). Although somewhat informative, such approaches have been criticized because clock time does not necessarily map onto psychological mechanisms (for a commentary, see Hopwood et al., 2022). For instance, research on BPD has showcased that its remarkable instability is better predicted by fluctuating social stressors (Lazarus et al., 2014; Sadikaj et al., 2013) rather than by the passage of time (Russell et al., 2007; Trull et al., 2008). Likewise, recent experience-sampling studies have illustrated that mood problems are typically preceded by important life events, such as daily social interactions (Benson et al., 2019), but also strong life stressors, such as divorce (Bleidorn et al., 2020) or unemployment (Luhmann et al., 2014). Our theory is consistent with these event-contingent approaches (Moskowitz & Sadikaj, 2012) because it suggests that affect is interwoven with life events, implying that a deeper insight of the former could perhaps be achieved only through concomitant measuring and modeling of the latter.

Indeed, focusing on life events, rather than clock time, may reap additional benefits, including informing researchers about the number, frequency, and timing of longitudinal assessments. As an example, consider research on mood reactivity, which has indicated that mood gets triggered differently across clinical populations—for example, people with borderline traits are more reactive to less affiliative people (Sadikaj et al., 2013) and people with narcissistic traits are more reactive to assertive people (Wright et al., 2017). This research implies that “timing” the assessments so that these disparate triggers get captured is vital when examining reactive dynamics. The same event-contingent approaches might be necessary when investigating transience and reactivity, which, by definition, vary,

respectively, based on fast-scale events (e.g., daily hassles; Mauss et al., 2005) and longer-scale events (e.g., divorce; Bleidorn et al., 2020). Our theory is consistent with this evidence and embraces recent calls suggesting that “the spacing of assessments in longitudinal designs should be based on a formal theoretical model about the underlying psychological process” (Hopwood et al., 2022, p. 888).

Nevertheless, although our theory presents an advancement in this sense, it does not yet elucidate the timescales of particular mood dynamics (given insufficient empirical evidence). In that sense, more work is needed on event-contingent approaches to uncover the kinds of life events that trigger particular kinds of affects (see Moskowitz & Sadikaj, 2012). Such research could start with pilot longitudinal investigations that oversample life events to adjudicate how many and which ones are, in fact, needed for the modeling of specific mood patterns (for a discussion, see Boker & Nesselroade, 2002). Researchers could then more specifically test how frequently specific event-affect contingencies occur in people’s lives and whether those contingencies vary across psychiatric disorders (Hopwood et al., 2022). Ultimately, the understanding of mood timescales will be predicated on the synergistic use of data-driven models (that uncover event-affect patterns) and theory-driven models (that instantiate those patterns in well-defined formal theories; Haslbeck, Ryan, et al., 2021; Ryan et al., 2025).

### ***Limitations and future directions***

Despite the strengths of our framework, several of its limitations must also be acknowledged. First and foremost, we have favored model parsimony rather than complexity. As an example, we have assumed that our only dynamic (time-varying) parameter was the reference point. Arguably, however, the learning rate and weight parameters could have also been time-varying (with their own update equations), yielding interesting affective dynamics; for instance, with higher learning for positive information (Pulcu & Browning, 2017) or situational, rather than pervasive, reactive patterns (Bellemare et al., 2023). However, it is not clear how (or whether) such dynamics inform MI, which is why we have opted for fixing these parameters to first explore whether any insights could be garnered with a more parsimonious model. Future work may wish to expand our framework by adding dynamic equations for weight or learning, examining whether they can add further insights on MI. In the same spirit, future work may wish to also expand our framework by including computations for different kinds of evaluations, including

evaluations of human behavior (which can yield social emotions, such as anger) or evaluations for future prospects (which can yield future-oriented emotions, such as anxiety; for details, see Emanuel & Eldar, 2022).

A second and related limitation is that our model treats valence as a bipolar dimension, varying from positive to negative emotion. By contrast, some researchers have long noted that positive and negative affect could exist as two independent dimensions (see Diener & Emmons, 1984; Warr et al., 1983; Watson et al., 1999). Following this work, an interesting research avenue could be to extend our model by distinguishing positive and negative emotions. Still, we note that the unidimensional approach employed here is already sufficient to explain a wealth of mood dynamics. Moreover, recent work on mood dynamics has supported this assumption by illustrating that maladaptive ways of feeling reflect “a more bipolar experience of positive and negative affect, reflecting reduced emotional complexity and flexibility” (Dejonckheere et al., 2018). This research suggests that maladaptive mood patterns could be captured by simpler models that treat valence in more “unidimensional” and “inflexible” terms, precisely as our framework suggests (see also Brose et al., 2015; Feldman, 1995; Rafaeli et al., 2007).

A final limitation of our work may concern its purely theoretical nature. Indeed, here, we have focused on simulations to examine whether a simple computational model of evaluation can shed light on key aspects of MI. Still, although we have not conducted any empirical tests, we note that our model is firmly grounded on previous empirical research. For example, the notion that emotion is fundamentally reference-dependent has been supported by various studies illustrating how emotions are dependent not on life outcomes per se (e.g., a student’s grade on a school test) but rather on the discrepancy between those outcomes and reference outcomes (e.g., the mismatch between a student’s actual grade and desired grade; Eldar et al., 2016; Emanuel & Eldar, 2022; Otto & Eichstaedt, 2018; Villano et al., 2020). Likewise, the idea that evaluation is central to emotion is a core tenet of validated cognitive models (Koszegi & Rabin, 2006; Louie et al., 2013, 2014, 2015; Rigoli, 2019; Stewart et al., 2006, 2015). In that sense, plenty of research already supports the central tenets of our framework, implying that future work may wish to examine its more novel predictions, including the link between our cognitive parameters and (a) specific clinical problems and (b) longitudinal affect patterns. To these ends, we refer readers to our Supplemental Material and GitHub repository (<https://github.com/OrestisZavlis/MoodInstability>), which provide more information on how to test our theoretical predictions.

## Conclusion

To conclude, we have developed a theory that formally integrates three well-known MI types: emotional rigidity, transience, and reactivity. Our theory illustrates how specific evaluative processes (i.e., evaluative consistency, inconsistency, and certainty) can generate specific emotional experiences (respectively, emotional rigidity, transiency, and reactivity). In that sense, our theory suggests that the disparate lines of experimental versus longitudinal research can be fruitfully united by examining how well-defined computational parameters relate to longitudinal mood patterns. Future work is necessary to examine these computational-longitudinal linkages and enrich them by elucidating the timescale of their dynamics.

## Transparency

Action Editor: Aleksandra Kaurin

Editor: Jennifer L. Tackett

Author Contributions

**Orestis Zavlis:** Conceptualization; Formal analysis; Investigation; Methodology; Project administration; Software; Visualization; Writing – original draft; Writing – review & editing.

**Richard P. Bentall:** Conceptualization; Investigation; Writing – review & editing.

**Peter Fonagy:** Conceptualization; Investigation; Writing – review & editing.

**Francesco Rigoli:** Conceptualization; Formal analysis; Investigation; Methodology; Project administration; Software; Supervision; Writing – review & editing.

## Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.


## Open Practices

This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



## ORCID iDs

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

Peter Fonagy  <https://orcid.org/0000-0003-0229-0091>

Francesco Rigoli  <https://orcid.org/0000-0003-2233-934X>

## Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/21677026251363862>

## Notes

1. Because of its conceptual heterogeneity, mood instability is also known as “emotional instability” and “affective instability.”

Here, we privilege the term “mood instability” to refer to the instability in one’s mood more generally (i.e., over a longer time frame). Our use of the term “emotional” is to refer to the more specific instances of this general mood instability, including the emotional rigidity, emotional transience, and emotional reactivity (that may ensue over a shorter time frame). Finally, the term “affect” is used throughout to refer to affective dynamics in a nonspecific way.

2. Although precision is usually defined as the inverse of variation ( $1/\sigma^2$ ), here, we defined it as the inverse of standard deviation:  $1/\sigma$ .

## References

- Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2016). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, 87, 53–63. <https://doi.org/10.1136/jnnp-2015-310737>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Arntz, A., Van Den Hoorn, M., Cornelis, J., Verheul, R., Van Den Bosch, W. M. C., & De Bie, A. J. H. T. (2003). Reliability and validity of the borderline personality disorder severity index. *Journal of Personality Disorders*, 17(1), 45–59. <https://doi.org/10.1521/pedi.17.1.45.24053>
- Bales, D., Van Beek, N., Smits, M., Willemsen, S., Busschbach, J. J. V., Verheul, R., & Andrea, H. (2012). Treatment outcome of 18-month, day hospital Mentalization-Based Treatment (MBT) in patients with severe borderline personality disorder in the Netherlands. *Journal of Personality Disorders*, 26(4), 568–582. <https://doi.org/10.1521/pedi.2012.26.4.568>
- Barata, P. C., Holtzman, S., Cunningham, S., O’Connor, B. P., & Stewart, D. E. (2016). Building a definition of irritability from academic definitions and lay descriptions. *Emotion Review*, 8(2), 164–172. <https://doi.org/10.1177/1754073915576228>
- Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1), 173–196.
- Barlow, D. H., Curren, A. J., & Woodard, L. S. (2021). Neuroticism and disorders of emotion: A new synthesis. *Current Directions in Psychological Science*, 30(5), 410–417.
- Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1–23.
- Bateman, A., & 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>
- Bellemare, M. G., Dabney, W., & Rowland, M. (2023). *Distributional reinforcement learning*. MIT Press.
- Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The two cultures of computational psychiatry. *JAMA Psychiatry*, 76(6), 563–564.

- Benson, L., English, T., Conroy, D. E., Pincus, A. L., Gerstorf, D., & Ram, N. (2019). Age differences in emotion regulation strategy use, variability, and flexibility: An experience sampling approach. *Developmental Psychology, 55*(9), 1951–1964. <https://doi.org/10.1037/dev0000727>
- Black, D. W., Blum, N., Letuchy, E., Doebbeling, C. C., Forman-Hoffman, V. L., & Doebbeling, B. N. (2006). Borderline personality disorder and traits in veterans: Psychiatric comorbidity, healthcare utilization, and quality of life along a continuum of severity. *CNS Spectrums, 11*(9), 680–689. <https://doi.org/10.1017/S1092852900014772>
- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J., Hennecke, M., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., & Orth, U. (2020). Longitudinal experience-wide association studies—A framework for studying personality change. *European Journal of Personality, 34*(3), 285–300.
- Boker, S. M., & Nesselroade, J. R. (2002). A method for modeling the intrinsic dynamics of intraindividual variability: Recovering the parameters of simulated oscillators in multi-wave panel data. *Multivariate Behavioral Research, 37*(1), 127–160.
- Bolton, D., & Hill, J. (2004). *Mind, meaning and mental disorder: The nature of causal explanation in psychology and psychiatry*. Oxford University Press.
- Bortolla, R., Cavicchioli, M., Fossati, A., & Maffei, C. (2020). Emotional reactivity in borderline personality disorder: Theoretical considerations based on a meta-analytic review of laboratory studies. *Journal of Personality Disorders, 34*(1), 64–87.
- Bowen, R., Balbuena, L., Leuschen, C., & Baetz, M. (2012). Mood instability is the distinctive feature of neuroticism. Results from the British Health and Lifestyle Study (HALS). *Personality and Individual Differences, 53*(7), 896–900. <https://doi.org/10.1016/j.paid.2012.07.003>
- Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.
- Bringmann, L. F., Ferrer, E., Hamaker, E. L., Borsboom, D., & Tuerlinckx, F. (2018). Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model. *Multivariate Behavioral Research, 53*(3), 293–314. <https://doi.org/10.1080/00273171.2018.1439722>
- Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment, 23*(4), 425–435.
- Broome, M. R., He, Z., Iftikhar, M., Eyden, J., & Marwaha, S. (2015). Neurobiological and behavioural studies of affective instability in clinical populations: A systematic review. *Neuroscience & Biobehavioral Reviews, 51*, 243–254. <https://doi.org/10.1016/j.neubiorev.2015.01.021>
- Broome, M. R., Saunders, K. E. A., Harrison, P. J., & Marwaha, S. (2015). Mood instability: Significance, definition and measurement. *British Journal of Psychiatry, 207*(4), 283–285. <https://doi.org/10.1192/bjp.bp.114.158543>
- Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between-person and within-person structures of affect are a matter of degree. *European Journal of Personality, 29*(1), 55–71.
- Bruch, E., & Feinberg, F. (2017). Decision-making processes in social contexts. *Annual Review of Sociology, 43*(1), 207–227. <https://doi.org/10.1146/annurev-soc-060116-053622>
- Bylsma, L. M., Morris, B. H., & Rottenberg, J. (2008). A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review, 28*(4), 676–691.
- Chow, S.-M., Ram, N., Boker, S. M., Fujita, F., & Clore, G. (2005). Emotion as a thermostat: Representing emotion regulation using a damped oscillator model. *Emotion, 5*(2), 208–225. <https://doi.org/10.1037/1528-3542.5.2.208>
- Clark, J. E., Watson, S., & Friston, K. J. (2018). What is mood? A computational perspective. *Psychological Medicine, 48*(14), 2277–2284. <https://doi.org/10.1017/S0033291718000430>
- Cowdry, R. W., Gardner, D. L., O’Leary, K. M., Leibenluft, E., & Rubinow, D. R. (1991). Mood variability: A study of four groups. *The American Journal of Psychiatry, 148*(11), 1505–1511.
- Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B., & Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology, 114*(2), 323–341. <https://doi.org/10.1037/pspp0000186>
- Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour, 3*(5), 478–491. <https://doi.org/10.1038/s41562-019-0555-0>
- 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>
- Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology, 47*(5), 1105–1117. <https://doi.org/10.1037/0022-3514.47.5.1105>
- Dimeff, L., & Linehan, M. M. (2001). Dialectical behavior therapy in a nutshell. *The California Psychologist, 34*(3), 10–13.
- Durstewitz, D., Huys, Q. J. M., & Koppe, G. (2021). Psychiatric illnesses as disorders of network dynamics. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6*(9), 865–876. <https://doi.org/10.1016/j.bpsc.2020.01.001>
- Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology, 118*(1), 195–202. <https://doi.org/10.1037/a0014868>
- Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364–370.
- Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications, 6*, Article 6149. <https://doi.org/10.1038/ncomms7149>
- Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences, 20*(1), 15–24.
- Ellard, K. K., Fairholme, C. P., Boisseau, C. L., Farchione, T. J., & Barlow, D. H. (2010). Unified protocol for the



- transdiagnostic treatment of emotional disorders: Protocol development and initial outcome data. *Cognitive and Behavioral Practice*, 17(1), 88–101.
- Emanuel, A., & Eldar, E. (2022). Emotions as computations. *Neuroscience & Biobehavioral Reviews*, 144, Article 104977. <https://doi.org/10.1016/j.neubiorev.2022.104977>
- Ernst, A. F., Timmerman, M. E., Ji, F., Jeronimus, B. F., & Albers, C. J. (2024). Mixture multilevel vector-autoregressive modeling. *Psychological Methods*, 29(1), 137–154. <https://doi.org/10.1037/met0000551>
- Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, 69(1), 153–166. <https://doi.org/10.1037/0022-3514.69.1.153>
- 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. 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](https://doi.org/10.1016/S2215-0366(14)70275-5)
- Gilbert, P., Allan, S., Nicholls, W., & Olsen, K. (2005). The assessment of psychological symptoms of patients referred to community mental health teams: Distress, chronicity and life interference. *Clinical Psychology & Psychotherapy*, 12(1), 10–27. <https://doi.org/10.1002/cpp.426>
- Goodwin, F. K., Jamison, K. R., & Ghaemi, S. N. (2007). *Manic-depressive illness: Bipolar disorders and recurrent depression* (2nd ed.). Oxford University Press.
- Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nature Reviews Disease Primers*, 4, Article 18029. <https://doi.org/10.1038/nrdp.2018.29>
- Guthrie, C., Dormann, C., & Voelkle, M. C. (2020). Reciprocal effects between job stressors and burnout: A continuous time meta-analysis of longitudinal studies. *Psychological Bulletin*, 146(12), 1146–1173. <https://doi.org/10.1037/bul0000304>
- Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, 7(4), 316–322. <https://doi.org/10.1177/1754073915590619>
- Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2021). A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research*, 56(1), 120–149.
- 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>
- Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M.-L., Hörmann, S., Hennings, J., Hasan, A., Papazova, I., Strube, W., Bolis, D., Schilbach, L., & Mathys, C. (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLOS Computational Biology*, 16(9), Article e1008162. <https://doi.org/10.1371/journal.pcbi.1008162>
- Henry, C., M'Bailara, K., Desage, A., Gard, S., Misdrahi, D., & Vieta, E. (2007). Towards a reconceptualization of mixed states, based on an emotional-reactivity dimensional model. *Journal of Affective Disorders*, 101(1–3), 35–41. <https://doi.org/10.1016/j.jad.2006.10.027>
- Henry, C., M'Bailara, K., Lépine, J.-P., Lajnef, M., & Leboyer, M. (2010). Defining bipolar mood states with quantitative measurement of inhibition/activation and emotional reactivity. *Journal of Affective Disorders*, 127(1–3), 300–304.
- Henry, C., Swendsen, J., Van den Bulke, D., Sorbara, F., Demotes-Mainard, J., & Leboyer, M. (2003). Emotional hyper-reactivity as a fundamental mood characteristic of manic and mixed states. *European Psychiatry*, 18(3), 124–128.
- Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation*, 33(2), 398–446.
- Hills, P., & Argyle, M. (2001). Emotional stability as a major dimension of happiness. *Personality and Individual Differences*, 31(8), 1357–1364. [https://doi.org/10.1016/S0191-8869\(00\)00229-4](https://doi.org/10.1016/S0191-8869(00)00229-4)
- Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational psychiatry needs time and context. *Annual Review of Psychology*, 73(1), 243–270.
- Hochschild Tolpin, L., Cimboric Gunthert, K., Cohen, L. H., & O'Neill, S. C. (2004). Borderline personality features and instability of daily negative affect and self-esteem. *Journal of Personality*, 72(1), 111–138. <https://doi.org/10.1111/j.0022-3506.2004.00258.x>
- 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., Bleidorn, W., & Wright, A. G. (2022). Connecting theory to methods in longitudinal research. *Perspectives on Psychological Science*, 17(3), 884–894.
- Hu, Y., Boker, S., Neale, M., & Klump, K. L. (2014). Coupled latent differential equation with moderators: Simulation and application. *Psychological Methods*, 19(1), 56–71. <https://doi.org/10.1037/a0032476>
- Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, 46(1), 3–19. <https://doi.org/10.1038/s41386-020-0746-4>
- 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>
- Huys, Q. J. M., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, 24(6), 544–551. <https://doi.org/10.1016/j.neunet.2011.03.001>
- Isaksson, J., Van't Westeinde, A., Cauvet, É., Kuja-Halkola, R., Lundin, K., Neufeld, J., Willfors, C., & Bölte, S. (2019).



- Social cognition in autism and other neurodevelopmental disorders: A co-twin control study. *Journal of Autism and Developmental Disorders*, 49(7), 2838–2848. <https://doi.org/10.1007/s10803-019-04001-4>
- Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, 13(4), 354–375. <https://doi.org/10.1037/a0014173>
- Johnson, S. L., Morriss, R., Scott, J., Paykel, E., Kinderman, P., Kolamunnage-Dona, R., & Bentall, R. P. (2011). Depressive and manic symptoms are not opposite poles in bipolar disorder: Depression and mania. *Acta Psychiatrica Scandinavica*, 123(3), 206–210. <https://doi.org/10.1111/j.1600-0447.2010.01602.x>
- Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, 107(38), 16489–16493. <https://doi.org/10.1073/pnas.1011492107>
- Kalisch, R., Cramer, A. O. J., Binder, H., Fritz, J., Leertouwer, Ij., Lunansky, G., Meyer, B., Timmer, J., Veer, I. M., & Van Harmelen, A.-L. (2019). Deconstructing and reconstructing resilience: A dynamic network approach. *Perspectives on Psychological Science*, 14(5), 765–777. <https://doi.org/10.1177/1745691619855637>
- Kanske, P., Schulze, L., Dziobek, I., Scheibner, H., Roepke, S., & Singer, T. (2016). The wandering mind in borderline personality disorder: Instability in self- and other-related thoughts. *Psychiatry Research*, 242, 302–310. <https://doi.org/10.1016/j.psychres.2016.05.060>
- Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*, 148, Article 105137. <https://doi.org/10.1016/j.neubiorev.2023.105137>
- Kaufman, E. A., & Meddaoui, B. (2021). Identity pathology and borderline personality disorder: An empirical overview. *Current Opinion in Psychology*, 37, 82–88. <https://doi.org/10.1016/j.copsyc.2020.08.015>
- Keltner, D., & Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion*, 13(5), 467–480.
- 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. (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.
- Knight, R. P. (1953). Borderline states. *Bulletin of the Menninger Clinic*, 17(1), 1–12.
- Kockler, T. D., Santangelo, P. S., Eid, M., Kuehner, C., Bohus, M., Schmaedeke, S., & Ebner-Priemer, U. W. (2022). Self-esteem instability might be more characteristic of borderline personality disorder than affective instability: Findings from an e-diary study with clinical and healthy controls. *Journal of Psychopathology and Clinical Science*, 131(3), 301–313. <https://doi.org/10.1037/abn000073>
- Koenigsberg, H. W. (2010). Affective instability: Toward an integration of neuroscience and psychological perspectives. *Journal of Personality Disorders*, 24(1), 60–82. <https://doi.org/10.1521/pedi.2010.24.1.60>
- Koszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4), 1133–1165. <https://doi.org/10.1093/qje/121.4.1133>
- Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, 13(6), 1132–1141. <https://doi.org/10.1037/a0033579>
- Krizan, Z., & Johar, O. (2015). Narcissistic rage revisited. *Journal of Personality and Social Psychology*, 108(5), 784–801. <https://doi.org/10.1037/pspp0000013>
- Kube, T. (2023). Biased belief updating in depression. *Clinical Psychology Review*, 103, Article 102298. <https://doi.org/10.1016/j.cpr.2023.102298>
- Lange, J., Dalege, J., Borsboom, D., van Kleef, G. A., & Fischer, A. H. (2020). Toward an integrative psychometric model of emotions. *Perspectives on Psychological Science*, 15(2), 444–468.
- Lapate, R. C., & Heller, A. S. (2020). Context matters for affective chronometry. *Nature Human Behaviour*, 4(7), 688–689.
- Larsen, R. J. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *Journal of Personality and Social Psychology*, 52(6), 1195–1204. <https://doi.org/10.1037/0022-3514.52.6.1195>
- Larsen, R. J., & Diener, E. (1987). Affect intensity as an individual difference characteristic: A review. *Journal of Research in Personality*, 21(1), 1–39.
- Lazarus, S. A., Cheavens, J. S., Festa, F., & Zachary Rosenthal, M. (2014). Interpersonal functioning in borderline personality disorder: A systematic review of behavioral and laboratory-based assessments. *Clinical Psychology Review*, 34(3), 193–205. <https://doi.org/10.1016/j.cpr.2014.01.007>
- Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Sage.
- Linehan, M. M. (1987). Dialectical behavior therapy for borderline personality disorder: Theory and method. *Bulletin of the Menninger Clinic*, 51(3), 261–276.
- Linehan, M. M., & Chen, E. Y. (2005). Dialectical behavior therapy for eating disorders. In A. Freeman, S. H. Felgoise, C. M. Nezu, A. M. Nezu, & M. A. Reinecke (Eds.), *Encyclopedia of cognitive behavior therapy* (pp. 168–171). Springer.
- Loas, G. (1996). Vulnerability to depression: A model centered on anhedonia. *Journal of Affective Disorders*, 41(1), 39–53. [https://doi.org/10.1016/0165-0327\(96\)00065-1](https://doi.org/10.1016/0165-0327(96)00065-1)
- Louie, K., Glimcher, P. W., & Webb, R. (2015). Adaptive neural coding: From biological to behavioral decision-making. *Current Opinion in Behavioral Sciences*, 5, 91–99. <https://doi.org/10.1016/j.cobeha.2015.08.008>
- Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National*

- Academy of Sciences*, 110(15), 6139–6144. <https://doi.org/10.1073/pnas.1217854110>
- Louie, K., LoFaro, T., Webb, R., & Glimcher, P. W. (2014). Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. *The Journal of Neuroscience*, 34(48), 16046–16057. <https://doi.org/10.1523/JNEUROSCI.2851-14.2014>
- Lucas, R. E., Diener, E., Grob, A., Suh, E. M., & Shao, L. (2000). Cross-cultural evidence for the fundamental features of extraversion. *Journal of Personality and Social Psychology*, 79(3), 452–468. <https://doi.org/10.1037//0022-3514.79.3.452>
- Luhmann, M., Orth, U., Specht, J., Kandler, C., & Lucas, R. E. (2014). Studying changes in life circumstances and personality: It's about time. *European Journal of Personality*, 28(3), 256–266.
- MacKinnon, D. F., & Pies, R. (2006). Affective instability as rapid cycling: Theoretical and clinical implications for borderline personality and bipolar spectrum disorders. *Bipolar Disorders*, 8(1), 1–14.
- Mansell, W., Morrison, A. P., Reid, G., Lowens, I., & Tai, S. (2007). The interpretation of, and responses to, changes in internal states: An integrative cognitive model of mood swings and bipolar disorders. *Behavioural and Cognitive Psychotherapy*, 35(5), 515–539. <https://doi.org/10.1017/S1352465807003827>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Marwaha, S., He, Z., Broome, M., Singh, S. P., Scott, J., Eyden, J., & Wolke, D. (2014). How is affective instability defined and measured? A systematic review. *Psychological Medicine*, 44(9), 1793–1808. <https://doi.org/10.1017/S0033291713002407>
- Marwaha, S., Parsons, N., & Broome, M. (2013). Mood instability, mental illness and suicidal ideas: Results from a household survey. *Social Psychiatry and Psychiatric Epidemiology*, 48(9), 1431–1437. <https://doi.org/10.1007/s00127-013-0653-7>
- Marwaha, S., Parsons, N., Flanagan, S., & Broome, M. (2013). The prevalence and clinical associations of mood instability in adults living in England: Results from the Adult Psychiatric Morbidity Survey 2007. *Psychiatry Research*, 205(3), 262–268. <https://doi.org/10.1016/j.psychres.2012.09.036>
- Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2), 175–190. <https://doi.org/10.1037/1528-3542.5.2.175>
- McLaren, V., Gallagher, M., Hopwood, C. J., & Sharp, C. (2022). Hypermentalizing and borderline personality disorder: A meta-analytic review. *American Journal of Psychotherapy*, 75(1), 21–31. <https://doi.org/10.1176/appi.psychotherapy.20210018>
- Mitchell, P. B., & Malhi, G. S. (2004). Bipolar depression: Phenomenological overview and clinical characteristics. *Bipolar Disorders*, 6(6), 530–539. <https://doi.org/10.1111/j.1399-5618.2004.00137.x>
- Moeller, F. G., Barratt, E. S., Dougherty, D. M., Schmitz, J. M., & Swann, A. C. (2001). Psychiatric aspects of impulsivity. *American Journal of Psychiatry*, 158(11), 1783–1793. <https://doi.org/10.1176/appi.ajp.158.11.1783>
- 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>
- Morgan, T. A., & Zimmerman, M. (2015). Is borderline personality disorder underdiagnosed and bipolar disorder overdiagnosed? In L. W. Choi-Kain & J. G. Gunderson (Eds.), *Borderline personality and mood disorders* (pp. 65–78). Springer. [https://doi.org/10.1007/978-1-4939-1314-5\\_4](https://doi.org/10.1007/978-1-4939-1314-5_4)
- Moskowitz, D. S., & Sadikaj, G. (2012). Event-contingent recording. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 160–175). The Guilford Press.
- 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, 50–73. [https://doi.org/10.1162/CPSY\\_a\\_00014](https://doi.org/10.1162/CPSY_a_00014)
- Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica: Journal of the Econometric Society*, 29(3), 315–335. <https://doi.org/10.2307/1909635>
- Myin-Germeys, I., Krabbendam, L., Delespaul, P., & Van Os, J. (2003). Do life events have their effect on psychosis by influencing the emotional reactivity to daily life stress? *Psychological Medicine*, 33(2), 327–333.
- Myin-Germeys, I., Peeters, F., Havermans, R., Nicolson, N., DeVries, M. W., Delespaul, P., & Van Os, J. (2003). Emotional reactivity to daily life stress in psychosis and affective disorder: An experience sampling study. *Acta Psychiatrica Scandinavica*, 107(2), 124–131.
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., Van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), 815–818. <https://doi.org/10.1016/j.tics.2019.07.009>
- Ormel, J., Jeronimus, B. F., Kotov, R., Riese, H., Bos, E. H., Hankin, B., Rosmalen, J. G., & Oldehinkel, A. J. (2013). Neuroticism and common mental disorders: Meaning and utility of a complex relationship. *Clinical Psychology Review*, 33(5), 686–697.
- Otto, A. R., & Eichstaedt, J. C. (2018). Real-world unexpected outcomes predict city-level mood states and risk-taking behavior. *PLOS ONE*, 13(11), Article e0206923. <https://doi.org/10.1371/journal.pone.0206923>
- Oud, J. H., & Jansen, R. A. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, 65, 199–215.
- Pacchiarotti, I., Nivoli, A. M. A., Mazzarini, L., Kotzalidis, G. D., Sani, G., Koukopoulos, A., Scott, J., Strejilevich, S., Sánchez-Moreno, J., Murru, A., Valentí, M., Girardi, P., Vieta, E., & Colom, F. (2013). The symptom structure of bipolar acute episodes: In search for the mixing link. *Journal of Affective Disorders*, 149(1–3), 56–66. <https://doi.org/10.1016/j.jad.2013.01.003>

- Page, C. E., Epperson, C. N., Novick, A. M., Duffy, K. A., & Thompson, S. M. (2024). Beyond the serotonin deficit hypothesis: Communicating a neuroplasticity framework of major depressive disorder. *Molecular Psychiatry*, 29(12), 3802–3813. <https://doi.org/10.1038/s41380-024-02625-2>
- Paris, J., Gunderson, J., & Weinberg, I. (2007). The interface between borderline personality disorder and bipolar spectrum disorders. *Comprehensive Psychiatry*, 48(2), 145–154. <https://doi.org/10.1016/j.comppsy.2006.10.001>
- Perugi, G., & Akiskal, H. S. (2005). Emerging concepts of mixed states: A longitudinal perspective. In A. Marneros & F. Goodwin (Eds.), *Bipolar disorders: Mixed states, rapid cycling and atypical forms* (pp. 45–60). Cambridge University Press.
- Pulcu, E., & Browning, M. (2017). Affective bias as a rational response to the statistics of rewards and punishments. *eLife*, 6, Article e27879. <https://doi.org/10.7554/eLife.27879>
- Rafaeli, E., Rogers, G. M., & Revelle, W. (2007). Affective synchrony: Individual differences in mixed emotions. *Personality and Social Psychology Bulletin*, 33(7), 915–932.
- Revelle, W. (1997). Extraversion and impulsivity: The lost dimension. In H. Nyborg (Ed.), *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty* (pp. 189–212). Pergamon/Elsevier Science Inc.
- Rigoli, F. (2019). Reference effects on decision-making elicited by previous rewards. *Cognition*, 192, Article 104034. <https://doi.org/10.1016/j.cognition.2019.104034>
- Rigoli, F. (2022a). Prisoner of the present: Borderline personality and a tendency to overweight cues during Bayesian inference. *Personality Disorders*, 13(6), 609–618. <https://doi.org/10.1037/per0000549>
- Rigoli, F. (2022b). When all glasses look half empty: A computational model of reference dependent evaluation to explain depression. *Journal of Cognitive Psychology*, 34(8), 1022–1031. <https://doi.org/10.1080/20445911.2022.2107650>
- Rigoli, F., & Martinelli, C. (2021). A reference-dependent computational model of anorexia nervosa. *Cognitive, Affective, & Behavioral Neuroscience*, 21(2), 269–277. <https://doi.org/10.3758/s13415-021-00886-w>
- Rigoli, F., & Martinelli, C. (2023). A computational theory of evaluation processes in apathy. *Current Psychology*, 42, 26163–26172. <https://doi.org/10.1007/s12144-022-03643-5>
- Rigoli, F., Martinelli, C., & Pezzulo, G. (2021). The half-empty/full glass in mental health: A reference-dependent computational model of evaluation in psychopathology. *Clinical Psychological Science*, 9(6), 1021–1034. <https://doi.org/10.1177/2167702621998344>
- 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>
- Rottenberg, J. (2005). Mood and emotion in major depression. *Current Directions in Psychological Science*, 14(3), 167–170. <https://doi.org/10.1111/j.0963-7214.2005.00354.x>
- Rottenberg, J., Gross, J. J., & Gotlib, I. H. (2005). Emotion context insensitivity in major depressive disorder. *Journal of Abnormal Psychology*, 114(4), 627–639. <https://doi.org/10.1037/0021-843X.114.4.627>
- Russell, J. J., Moskowitz, D., Zuroff, D. C., Sookman, D., & Paris, J. (2007). Stability and variability of affective experience and interpersonal behavior in borderline personality disorder. *Journal of Abnormal Psychology*, 116(3), 578–588. <https://doi.org/10.1037/0021-843X.116.3.578>
- Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, 111(33), 12252–12257.
- Ryan, O., Dablander, F., & Haslbeck, J. (2025). Towards a generative model for emotion dynamics. *Psychological Review*, 132(2), 416–441. <https://doi.org/10.1037/rev0000513>
- Sadikaj, G., Moskowitz, D., Russell, J. J., Zuroff, D. C., & Paris, J. (2013). Quarrelsome behavior in borderline personality disorder: Influence of behavioral and affective reactivity to perceptions of others. *Journal of Abnormal Psychology*, 122(1), 195–207. <https://doi.org/10.1037/a0030871>
- Santangelo, P., Reinhard, I., Mussgay, L., Steil, R., Sawitzki, G., Klein, C., Trull, T. J., Bohus, M., & Ebner-Priemer, U. W. (2014). Specificity of affective instability in patients with borderline personality disorder compared to post-traumatic stress disorder, bulimia nervosa, and healthy controls. *Journal of Abnormal Psychology*, 123(1), 258–272. <https://doi.org/10.1037/a0035619>
- Santangelo, P. S., Kockler, T. D., Zeitler, M.-L., Knies, R., Kleindienst, N., Bohus, M., & Ebner-Priemer, U. W. (2020). Self-esteem instability and affective instability in everyday life after remission from borderline personality disorder. *Borderline Personality Disorder and Emotion Dysregulation*, 7(1), Article 25. <https://doi.org/10.1186/s40479-020-00140-8>
- Santangelo, P. S., Reinhard, I., Koudela-Hamila, S., Bohus, M., Holtmann, J., Eid, M., & Ebner-Priemer, U. W. (2017). The temporal interplay of self-esteem instability and affective instability in borderline personality disorder patients' everyday lives. *Journal of Abnormal Psychology*, 126(8), 1057–1065. <https://doi.org/10.1037/abn0000288>
- Sharp, C. (2014). The social-cognitive basis of BPD: A theory of hypermentalizing. In C. Sharp & J. L. Tackett (Eds.), *Handbook of borderline personality disorder in children and adolescents* (pp. 211–225). Springer New York. [https://doi.org/10.1007/978-1-4939-0591-1\\_15](https://doi.org/10.1007/978-1-4939-0591-1_15)
- Sherwood, V. R., & Cohen, C. P. (1994). *Psychotherapy of the quiet borderline patient: The as-if personality revisited*. J. Aronson.
- Silk, J. S. (2019). Context and dynamics: The new frontier for developmental research on emotion regulation. *Developmental Psychology*, 55(9), 2009–2014. <https://doi.org/10.1037/dev0000768>
- Smaldino, P. E. (2020). How to build a strong theoretical foundation. *Psychological Inquiry*, 31(4), 297–301. <https://doi.org/10.1080/1047840X.2020.1853463>



- Smillie, L. D. (2013). Extraversion and reward processing. *Current Directions in Psychological Science*, 22(3), 167–172. <https://doi.org/10.1177/0963721412470133>
- Smith, M. M., Sherry, S. B., Chen, S., Saklofske, D. H., Flett, G. L., & Hewitt, P. L. (2016). Perfectionism and narcissism: A meta-analytic review. *Journal of Research in Personality*, 64, 90–101.
- Smith, M. M., Sherry, S. B., Chen, S., Saklofske, D. H., Mushquash, C., Flett, G. L., & Hewitt, P. L. (2018). The perniciousness of perfectionism: A meta-analytic review of the perfectionism–suicide relationship. *Journal of Personality*, 86(3), 522–542.
- Smith, R., Lane, R. D., Nadel, L., & Moutoussis, M. (2020). A computational neuroscience perspective on the change process in psychotherapy. In R. D. Lane (Ed.), *Neuroscience of enduring change* (pp. 395–432). Oxford University Press. <https://doi.org/10.1093/oso/9780190881511.003.0015>
- Solomon, D. A., Leon, A. C., Endicott, J., Coryell, W. H., Mueller, T. I., Posternak, M. A., & Keller, M. B. (2003). Unipolar mania over the course of a 20-year follow-up study. *American Journal of Psychiatry*, 160(11), 2049–2051. <https://doi.org/10.1176/appi.ajp.160.11.2049>
- Stein, K. F. (1996). Affect instability in adults with a borderline personality disorder. *Archives of Psychiatric Nursing*, 10(1), 32–40.
- Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, 53(1), 1–26. <https://doi.org/10.1016/j.cogpsych.2005.10.003>
- Stewart, N., Reimers, S., & Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, 61(3), 687–705. <https://doi.org/10.1287/mnsc.2013.1853>
- Stiglmayr, C., Grathwol, T., & Bohus, M. (2001). States of aversive tension in patients with borderline personality disorder: A controlled field study. In J. Fahrenberg & M. Myrtek (Eds.), *Progress in ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 135–141). Hogrefe & Huber Publishers.
- Stokes, P. R. A., Yalin, N., Mantingh, T., Colasanti, A., Patel, R., Bellivier, F., Leboyer, M., Henry, C., Kahn, J.-P., Etain, B., & Young, A. H. (2020). Unipolar mania: Identification and characterisation of cases in France and the United Kingdom. *Journal of Affective Disorders*, 263, 228–235. <https://doi.org/10.1016/j.jad.2019.11.024>
- Story, G. W., Ereira, S., Valle, S., Chamberlain, S. R., Grant, J. E., & Dolan, R. J. (2024). A computational signature of self-other mergence in borderline personality disorder. *Translational Psychiatry*, 14, Article 473. <https://doi.org/10.1038/s41398-024-03170-w>
- Strickland, J. C., & Johnson, M. W. (2021). Rejecting impulsivity as a psychological construct: A theoretical, empirical, and sociocultural argument. *Psychological Review*, 128(2), 336–361. <https://doi.org/10.1037/rev0000263>
- Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, 24(2), 127–136.
- Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.
- Swann, A. C., Lafer, B., Perugi, G., Frye, M. A., Bauer, M., Bahk, W.-M., Scott, J., Ha, K., & Suppes, T. (2013). Bipolar mixed states: An international society for bipolar disorders task force report of symptom structure, course of illness, and diagnosis. *American Journal of Psychiatry*, 170(1), 31–42. <https://doi.org/10.1176/appi.ajp.2012.12030301>
- Swann, A. C., Steinberg, J. L., Lijffijt, M., & Moeller, G. F. (2009). Continuum of depressive and manic mixed states in patients with bipolar disorder: Quantitative measurement and clinical features. *World Psychiatry*, 8(3), 166–172. <https://doi.org/10.1002/j.2051-5545.2009.tb00245.x>
- Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Gotlib, I. H. (2010). Maladaptive coping, adaptive coping, and depressive symptoms: Variations across age and depressive state. *Behaviour Research and Therapy*, 48(6), 459–466. <https://doi.org/10.1016/j.brat.2010.01.007>
- Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Gotlib, I. H. (2012). The everyday emotional experience of adults with major depressive disorder: Examining emotional instability, inertia, and reactivity. *Journal of Abnormal Psychology*, 121(4), 819–829. <https://doi.org/10.1037/a0027978>
- Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, 9(1), 151–176. <https://doi.org/10.1146/annurev-clinpsy-050212-185510>
- Trull, T. J., Lane, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emotion Review*, 7(4), 355–361. <https://doi.org/10.1177/1754073915590617>
- Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., Piasecki, T. M., & Watson, D. (2008). Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, 117(3), 647–661. <https://doi.org/10.1037/a0012532>
- Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Palmius, N., Osipov, M., Clifford, G. D., Goodwin, G. M., & De Vos, M. (2016). Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *Journal of Affective Disorders*, 205, 225–233. <https://doi.org/10.1016/j.jad.2016.06.065>
- Tyrer, P. (2009). Why borderline personality disorder is neither borderline nor a personality disorder: Borderline personality disorder wrongly named. *Personality and Mental Health*, 3(2), 86–95. <https://doi.org/10.1002/pmh.78>
- Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the Brownian motion. *Physical Review*, 36(5), 823–841.
- Vaidyanathan, U., Hall, J. R., Patrick, C. J., & Bernat, E. M. (2011). Clarifying the role of defensive reactivity deficits in psychopathy and antisocial personality using startle reflex methodology. *Journal of Abnormal Psychology*, 120(1), 253–258. <https://doi.org/10.1037/a0021224>
- Vallacher, R. R., Read, S. J., & Nowak, A. (Eds.). (2017). *Computational social psychology*. Routledge, Taylor & Francis Group.

- Vanhasbroeck, N., Ariens, S., Tuerlinckx, F., & Loossens, T. (2021). Computational models for affect dynamics. In C. E. Waugh & P. Kuppens (Eds.), *Affect dynamics* (pp. 213–260). Springer Nature Switzerland AG. [https://doi.org/10.1007/978-3-030-82965-0\\_10](https://doi.org/10.1007/978-3-030-82965-0_10)
- Vieta, E., & Valentí, M. (2013). Mixed states in DSM-5: Implications for clinical care, education, and research. *Journal of Affective Disorders*, 148(1), 28–36. <https://doi.org/10.1016/j.jad.2013.03.007>
- Villano, W. J., Otto, A. R., Ezie, C., Gillis, R., & Heller, A. S. (2020). Temporal dynamics of real-world emotion are more strongly linked to prediction error than outcome. *Journal of Experimental Psychology: General*, 149(9), 1755–1766. <https://doi.org/10.1037/xge0000740>
- Vinckier, F., Rigoux, L., Oudiette, D., & Pessiglione, M. (2018). Neuro-computational account of how mood fluctuations arise and affect decision making. *Nature Communications*, 9(1), Article 1708. <https://doi.org/10.1038/s41467-018-03774-z>
- Warr, P. B., Barter, J., & Brownbridge, G. (1983). On the independence of positive and negative affect. *Journal of Personality and Social Psychology*, 44(3), 644–651. <https://doi.org/10.1037/0022-3514.44.3.644>
- Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, 76(5), 820–838. <https://doi.org/10.1037/0022-3514.76.5.820>
- Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *eLife*, 6, Article e28098. <https://doi.org/10.7554/eLife.28098>
- Wirth, M., Voss, A., Wirth, S., & Rothermund, K. (2022). Affect dynamics and well-being: Explanatory power of the model of intraindividual variability in affect (MIVA). *Cognition and Emotion*, 36(2), 188–210. <https://doi.org/10.1080/02699931.2021.1993148>
- Wittgenstein, L. (2010). *Philosophical investigations* (G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans.; Rev. 4th ed.). Wiley-Blackwell.
- Wright, A. G., 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>
- Wright, A. G., Stepp, S. D., Scott, L. N., Hallquist, M. N., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology*, 126(7), 898–910. <https://doi.org/10.1037/abn0000286>
- Zavlis, O. (2023). Complex relational needs impede progress in NHS Talking Therapies (IAPT): Implications for public mental health. *Frontiers in Public Health*, 11, Article 1270926. <https://doi.org/10.3389/fpubh.2023.1270926>
- Zavlis, O. (2024a). Computational approaches to mental illnesses. *Nature Reviews Psychology*, 3, 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., Matheou, A., & Bentall, R. (2023). Identifying the bridge between depression and mania: A machine learning and network approach to bipolar disorder. *Bipolar Disorders*, 25(7), 571–582. <https://doi.org/10.1111/bdi.13316>
- Zavlis, O., Moutoussis, M., Fonagy, P., & Story, G. (2025a). A generative model of personality disorder as a relational disorder. *Journal of Psychopathology and Clinical Science*. Advance online publication. <https://doi.org/10.1037/abn0001010>
- Zavlis, O., Story, G., Friedrich, C., Fonagy, P., & Moutoussis, M. (2025b). A systematic review of computational modeling of interpersonal dynamics in psychopathology. *Nature Mental Health*, 3, 932–942. <https://doi.org/10.1038/s44220-025-00465-9>
- Zimmerman, M., Galione, J. N., Ruggero, C. J., Chelminski, I., Young, D., Dalrymple, K., & McGlinchey, J. B. (2010). Screening for bipolar disorder and finding borderline personality disorder. *The Journal of Clinical Psychiatry*, 71(09), 1212–1217. <https://doi.org/10.4088/JCP.09m05161yel>