nature mental health

Analysis

https://doi.org/10.1038/s44220-025-00465-9

A systematic review of computational modeling of interpersonal dynamics in psychopathology

Received: 1 November 2024

Accepted: 23 June 2025

Published online: 22 July 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Interpersonal dynamics have long been acknowledged as critical for the development and treatment of mental health problems. While recent computational approaches have been argued to be uniquely suited for investigating such dynamics, no systematic assessment has been made to scrutinize this claim. Here we conduct a systematic review to assess the utility of computational modeling in the field of interpersonal psychopathology. Candidate studies (k = 4,208), including preprints and conference manuscripts, were derived from five databases (MEDLINE, Embase, PsycINFO, Web of Science and Google Scholar) up to May 2025. A total of 58 studies met inclusion criteria and were assessed in terms of the validity, performance and transparency of their computational modeling. Bayesian modeling was the most common approach (k = 18), followed by machine learning (k = 17), dynamical systems modeling (k = 13) and reinforcement learning (k = 10). These approaches revealed several interpersonal disruptions across various mental health conditions, including rigid social learning in mood conditions, hypo- versus hyper-mentalizing in autism versus psychotic conditions and polarized relational dynamics in personality conditions. Despite these insights, critical challenges persist, with few studies reporting comprehensive performance metrics (16%) or adopting open science practices (20%). We discuss these challenges and conclude with more optimistic messages by suggesting that when rigorously and transparently conducted, computational approaches have the potential to advance our understanding of psychopathology by highlighting the social underpinnings of both mental health and disorder.

Interpersonal dynamics refer to the ways we relate to ourselves and others. These include but are not limited to attributional statements ('this is my fault')[1], mental inferences ('you hate me')[2] and social strategies ('I help you; you help me')[3]. When adaptive, interpersonal dynamics can foster positivity, flexibility and mental wellbeing[4,5]. However, when maladaptive (for example, in the sense that they are overly skewed or rigid), such dynamics can spiral downward into various psychopathologies, including personality[6], emotional[7] and psychotic pathologies[8].

Although at face value elusive (given their intersubjective nature), interpersonal dynamics can be experimentally examined, particularly with the use of computational tools. For example, reinforcement learning and Bayesian modeling can be used to examine whether participants hold rigid beliefs about themselves or others (because they cannot update them in light of disconfirming evidence)[9]. Moreover, dynamical systems can be used to examine how patient–therapist interactions evolve over time, converging in positive relational states (for example, states wherein both parties are well regulated)[10]. Finally, more exploratory, data-driven approaches can be leveraged to uncover linguistic markers that predict successful psychotherapy (for example, nonjudgemental communicating from the clinician's end)[11]. Together, these advances are part of computational

e-mail: orestis.zavlis.23@ucl.ac.uk

https://doi.org/10.1038/s44220-025-00465-9

BOX 1

Dynamical systems

Dynamical systems outline how sets of variables evolve over time based on a set of 'rules'. These rules are installed in systems of equations: specifically, differential equations that define changes in continuous time (dt=seconds) or difference equations that define changes in discrete time (t=Monday, t+1=Tuesday and so on).

An example dynamical system from our review is the one from Liebovitch et al.[31]

$$\frac{\mathrm{dPatient}}{\mathrm{d}t} = \alpha_1 + m_1\mathrm{Patient} + f_1(\mathrm{Therapist})$$

$$\frac{\mathrm{dTherapist}}{\mathrm{d}t} = \alpha_2 + m_2\mathrm{Therapist} + f_2(\mathrm{Patient}),$$

which outlines how the emotional state of a patient (dPatient) and a therapist (dTherapist) evolve over continuous time (dt) based on three influences: first, a parameter alpha (a), which denotes someone's baseline emotional state; second, a parameter mu (m), which denotes the extent to which an emotion carries over from a previous time point to the next; finally, an influence function (f), which denotes how the therapist and patient emotionally influence each other. The same system can be expressed in discrete terms, that is, $\mathrm{Patient}_{t+1} = \alpha_1 + m_1\mathrm{Patient}_t + f_1(\mathrm{Therapist}_t)$.

In our review, this dynamical system has been used, both theoretically[33] and empirically[34], to identify attracting states (for example, positive emotional states toward which the patient–therapist dyad gravitate) and also repelling states (for example, negative emotional states that the patient–therapist dyad avoid). Moreover, other dynamical systems have been employed to examine whether therapist–patient synchronicities (for example, synchronizing breathing patterns) characterize effective psychotherapy[39].

Key strengths of this approach is that it can model bidirectional influences between interacting agents and predict how their relational system (that is, dyadic influences) will evolve over time based on a set of starting conditions[10].

Key limitations of this approach include a potential oversimplification of patient–therapist interactions (to keep the model easily interpretable) and a difficulty fitting continuous dynamical systems to psychological data (because such data need to measure a process at a continuous scale, for instance, minutes or even seconds).

BOX 2

Reinforcement learning

Reinforcement learning suggests that humans, other animals and machines act so as to maximize long-term rewards. This idea of reward maximization has its origins in traditional behaviorist views in psychology (that is, operant conditioning) and can be operationalized in two ways: model-free learning and model-based learning.

An example of model-free learning is a modified Rescorla–Wagner model[79]

$$B_t^{\mathrm{self}} = B_{t-1}^{\mathrm{self}} + a\left(R_{t-1}^{\mathrm{self}} - B_{t-1}^{\mathrm{self}}\right)$$

$$B_t^{\mathrm{other}} = B_{t-1}^{\mathrm{other}} + a\left(R_{t-1}^{\mathrm{other}} - B_{t-1}^{\mathrm{other}}\right),$$

wherein beliefs about the self and others (B_t) are updated based on mismatches between prior expectations of the self and others (B_{t-1}) and real-life observations from them (R_{t-1}) weighted by an alpha parameter (a). For example, expecting a text from your friend ($B_{t-1}^{\mathrm{other}} = 10$) but not hearing from them ($R_{t-1}^{\mathrm{other}} = -10$) will make you update your belief about them in a negative manner ($R_{t-1}^{\mathrm{other}} - B_{t-1}^{\mathrm{other}} = -20$).

Model-based learning augments this simple associative learning framework by enabling agents to learn an internal model of their environment. This model allows agents to simulate future states and evaluate the utility of different actions before choosing one of them. Although more sophisticated than model-free learning, this approach was not employed by any of our reviewed studies.

A key application of our reviewed studies included estimating how patients with different diagnoses update their beliefs in the simple model-free way shown above. These studies revealed that learning parameters tend to be low ($a \approx 0$) for patients with borderline personality disorder, indicating that such patients hold rigid beliefs about others that are not updated in light of disconfirming evidence[48].

A key strength of this approach is its simplicity: model-free learning is one of the simplest algorithms to grasp, modify and fit in experimental datasets, explaining why many researchers prefer to model belief-updating in this way.

A key limitation of this approach is that it most likely does not represent the true data-generating process: humans may not simply learn in a model-free way as findings from probabilistic learning suggest (see Box 3 and validity results).

psychiatry, a growing field that aims to empirically examine and theoretically define mental disorders in terms of computational dysfunctions, rather than verbal descriptions [12–18] (Boxes 1–4).

Although the popularity of computational psychiatry is rising rapidly, no systematic assessment has been conducted on its utility and validity within the field of interpersonal dynamics. We see three reasons why such a systematic assessment is necessary. First, existing reviews on this topic have focused rather narrowly on famous computational paradigms (such as Bayesian modeling[19] and reinforcement learning[20]), not providing a systematic evaluation and integration of most available paradigms in this field of inquiry. Second, recent assessments have highlighted that computational studies in psychiatry tend to exhibit poor psychometric properties (such as low reliability and validity)[21,22], underscoring the need of systematically examining, rather than assuming, that computational modeling offers greater insights over traditional statistical perspectives. Finally, although both traditional[5,23,24] and recent theorists[25–27] have emphasized the centrality of interpersonal dynamics in mental health difficulties, no systematic attempt has been made to examine whether computational modeling that focuses on such dynamics offers especially notable insights on mental health difficulties (for example, more clinically applicable and ecologically valid findings).

In this study, we aim to address these queries by systematically reviewing extant computational modeling of interpersonal dynamics in psychopathology. To cover the entire spectrum of computational methodologies, we include both theory-driven and data-driven approaches (see Boxes 1–4 for an accessible introduction). Our primary aim is to evaluate whether these computational approaches can offer both theoretically notable and methodologically reliable insights about interpersonal psychopathology.

Results

Figure 1 presents our study selection process. From the 3,914 unique records, 3,758 were excluded (based on title and abstract), leaving 156

https://doi.org/10.1038/s44220-025-00465-9

BOX 3

Bayesian inference

Bayesian inference suggests that human beings update their beliefs in an (approximately) Bayesian way, using Bayes theorem

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

Put as simply as possible, Bayes theorem suggests that your prior belief about a given cause, *P*(cause)=probability that humans are untrustworthy, combines with the likelihood that this cause explains an observation, *P*(observation|cause)=likelihood that my friend ignores me, given that humans are untrustworthy, to form a posterior belief: *P*(cause|observation)=humans are very untrustworthy.

Importantly, unlike reinforcement learning, beliefs here are not point estimates (that is, single values) but entire distributions that outline the probabilities of certain social states: for instance, that there is an 80% probability that others harbor extremely bad intentions, 10% probability that they harbor moderately bad intentions and 10% probability that they harbor neutral intentions. When these distributions are precise (that is, most probability is concentrated in a particular state, such as a state of untrustworthiness), beliefs are overconfident and rigid; however, when they are imprecise (that is, probability is spread relatively equally across states), beliefs are uncertain and readily amenable to change.

Bayesian models can be used to examine how agents update their beliefs in light of social evidence. For example, such models have shown that people with different mental disorders have difficulty updating beliefs about themselves and others[51]. Moreover, hierarchical approaches (that model beliefs about beliefs) have shown that people with autism have difficulty engaging in deep mental inferences ('I think that you think that I think' and so on)[55], but people with paranoia are adept in doing so, explaining why they usually converge in paranoid attributions[57].

Key strengths of this modeling approach concern its flexibility across study designs and ability to formalize belief-updating mechanisms in a relatively intuitive way that appears to fit the data better than other mechanisms (for example, model-free learning)[46,56].

Key limitations include that complex, hierarchical models may overfit datasets and be difficult to interpret in a clinically pragmatic manner (for example, 'beliefs about beliefs about beliefs about beliefs' and so on).

BOX 4

Machine learning

Machine learning is a broad class of algorithms that aim to learn from observed data to make predictions about unobserved data. These learning algorithms can be grouped under at least three categories: supervised learning, unsupervised learning and reinforcement learning (as explained in Box 2).

Supervised learning includes learning how a set of well-defined inputs, $x \epsilon \mathbf{X}$, map onto a set of well-defined outputs, $y \epsilon \mathbf{Y}$. This learning is achieved by using example $\{x_i, y_i\}$ pairs in some data to train a model that predicts **Y** from **X**. When **Y** outputs are numerical (for example, symptom scores), this process is termed prediction; when *Y* outputs are categorical (for example, diagnoses), this process is termed classification.

Unsupervised learning includes identifying hidden patterns or structures in data comprising only inputs: $\mathbf{D} = \{x_i\}_{i=1}^{n}$. These patterns may be clusters of patients who exhibit particular symptom profiles (in quantitative data) or linguistic themes that distill the main sentiments of various phrases (in qualitative data).

In our review, supervised learning was employed to either classify the diagnostic status or predict the severity of various mental disorders using social variables, revealing that social isolation is among the strongest predictors of depression[69]. Unsupervised learning was instead mainly used in the domain of natural language processing, revealing notable linguistic themes that underpin therapeutic alliance (such as words reflecting better 'goal setting' between patients and therapists)[71], as well as meaningful psychotherapy (such as words from patients reflecting improvements in ways of 'relating to oneself and others')[74,75].

The key strengths of machine learning approaches include their ability to leverage large datasets, in a data-driven exploratory manner, to reveal the most important features for the prediction or classification of mental disorders, as well as novel linguistic themes that underlie the lived experiences of patients (and therapists).

The key limitations include that machine learning models have a tendency to overfit datasets and not generalize to new contexts (implying that they might capitalize the idiosyncrasies of particular datasets rather than reveal universal patterns).

necessitating full-text assessments. From these 156 studies, 58 met our inclusion criteria.

Study description

This section outlines a theoretical synthesis of our 58 included studies (Table 1), as well as two additional studies that were considered relevant despite not meeting inclusion criteria. Please refer to Supplementary Tables 1 and 2 for more details on these studies.

Random dynamical systems (k = 13)

Dynamical systems have been used to study how therapeutic relationships evolve over time and how interpersonal asynchronies map on different psychopathologies (Box 1). Some studies have applied these systems to formalize novel relational patterns, such as the unstable relationship dynamics of borderline personality disorder[28], the social motives of autonomy versus mergence[29] and the nonlinear interactions of seminal therapeutic constructs (for example, mentalizing with self-efficacy)[30]. From these modeling attempts, the most noteworthy perhaps was the dynamical system by Liebovitch and colleagues[31], which extended past work on marital interactions[32] by modeling how patients and therapists emotionally influence each other. This model is noteworthy because it was employed in both simulations (to illustrate cases of influential psychotherapy)[33] and empirical investigations (to illustrate that, regardless of theoretical orientation, influential therapists exhibit strikingly similar relational dynamics[10] because they tend to move patients from negative to positive relational states)[34].

Beyond therapeutic relationships, three studies applied dynamical systems to reveal various pathogenic coordination markers, such as detached patterns in infants with autism[35], motor incoordination in those with psychosis[36] and misaligned coregulation in those with borderline personality vis-à-vis their partners[37]. Finally, three other studies applied dynamical systems on second-by-second data from psychotherapy, showing that therapeutic success is based on both deterministic elements (for example, the patient consistently being focused on their therapist[38] or consistently returning to their baseline arousal because their therapist regulates them[39]) and stochastic elements (for example, the patient shifting from their regular breathing pattern when discussing emotionally notable matters)[40].

https://doi.org/10.1038/s44220-025-00465-9

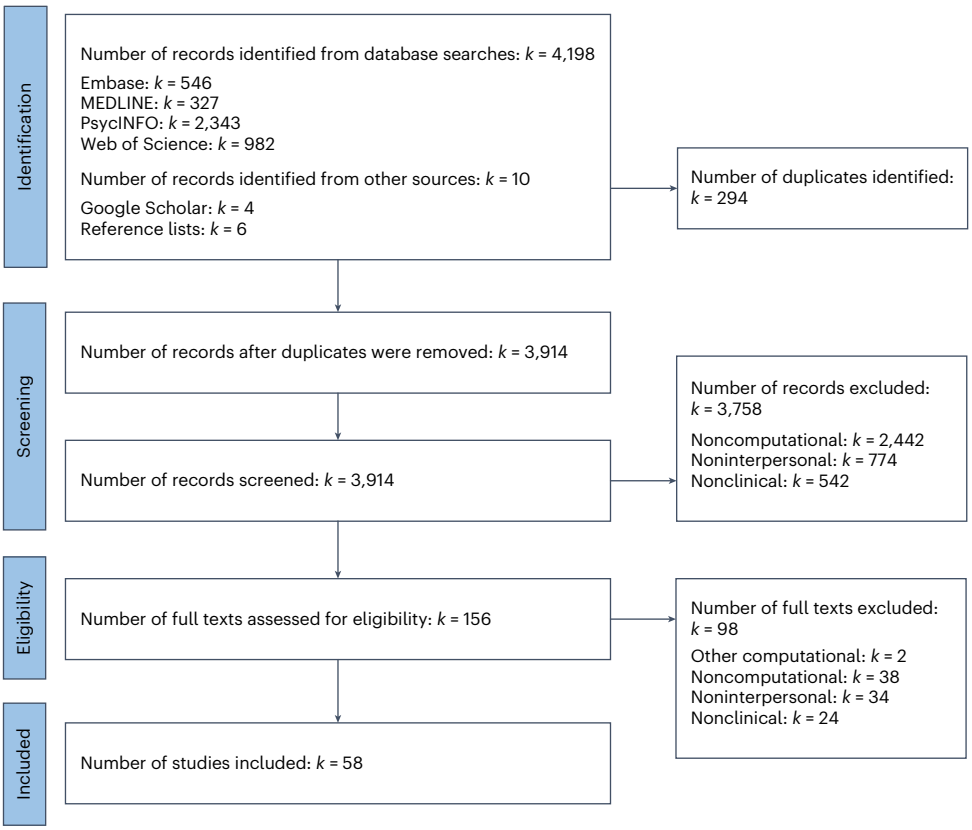

Fig. 1 | PRISMA flowchart. Out of 3,914 unique records, 58 met our inclusion criteria and were included in our study.

Reinforcement learning (k = 10)

Reinforcement learning has been used in a predominantly empirical way to decipher both transdiagnostic and disorder-specific patterns of social learning (Box 2). Transdiagnostically, studies have implicated low self-esteem to blunted social learning (for example, difficulty updating negative self-beliefs)[41,42] and adolescent experiences of relational trauma to unstable and credulous social learning (for example, 'This person was untrustworthy but may now be trustworthy')[43]. Disorder-specific findings mirrored these patterns by showing that depression is typified by overly rigid social learning (the extreme of low self-esteem)[44,45] while paranoia is associated with overly uncertain social learning (the extreme opposite of credulity)[46]. Finally, two studies[47,48] revealed that borderline personality disorder is characterized by a paradoxical combination of heightened social sensitivity and also rigid social learning: that is, an over-reaction to social information yet also a resistance to update social beliefs in light of that information. Interestingly, a third study[49] indicated that this paradoxical combination of social sensitivity with rigidity might be explained by a pattern of self-other mergence: people with borderline personality disorder tend to assume that others hold the same beliefs as them, implying that they may underestimate the need to update their beliefs about others even in light of new information from them.

Approximate bayesian inference (k = 18)

Bayesian models have been used both empirically (to understand how humans form social beliefs) and theoretically (to understand similar dynamics in simulation studies) (Box 3). Empirically, many studies converged in showing that various mental disorders (from anxiety[50] to borderline personality disorder[51] and psychotic disorders[52–54]) are typified by difficulties in updating beliefs about themselves and others. Other studies have extended this line of research by investigating meta-cognitive beliefs: that is, beliefs about beliefs[9]. Such studies have showcased that people with autism are typified by an inability to engage in this form of 'deep' social reasoning[55] ('I think that you think that I think' and so on), as well as that their social difficulties, emerge mainly in situations that necessitate this form of reasoning (a problem known as 'double-empathy')[56]. Conversely, people with paranoia exhibit the exact opposite difficulties: that is, they think too deeply about mental states and tend to conclude that others have ulterior motives[57]. Finally, people with a diagnosis of 'personality disorder' were shown to exhibit both overly deep[58] (and polarized[59]) and overly shallow[60] (and reactive[61]) mental inferences, which impair how they cooperate with others[62].

Based on these and related findings, several simulation studies have attempted to provide a more formal way of understanding these psychopathologies. One simulation study, for instance, has argued that 'personality disorders' can be more specifically understood as 'relational disorders' precisely because their main impairments include maladaptive ways of thinking about and relating to oneself and others[63]. By contrast, other studies have attempted to explain mental health problems by appealing to the free energy principle: that is, the idea that humans act so as to minimize the discrepancy between their subjective beliefs and their perceived reality. These studies used computer simulations to show that conditions such as depression and psychopathy emerge when people fail to update their beliefs about themselves in light of contradictory social evidence, leading to overly deflated self-perceptions in depression[64] and overly inflated self-perceptions in psychopathy[65].

Machine learning (k = 17)

Machine learning studies focused on three distinct topics: classifying, predicting or linguistically exploring psychological phenotypes (Box 4). Although classification studies revealed novel variables that could classify attachment styles (for example, anxiety from posting more emotional social media posts and avoidance from receiving more likes on such posts)[66] and psychosis (for example, from social and cognitive functioning)[67], they were generally deemed of low quality because they were predicated on small sizes (for example, 30–90 participants per group). By contrast, prediction studies were of higher quality and

revealed, consistently, that relational difficulties are the strongest predictors of disorders of emotion: for example, experiences of victimization were the strongest predictors of adolescent suicide attempts[68], social isolation was the strongest predictor of middle-age depression (N = 67,603)[69], and parental support was the strongest predictor of adolescent depression (N = 2445)[70]. Finally, studies analyzing natural language revealed linguistic markers that index strong therapeutic alliances (such as nonjudgemental communication from therapists[11] or better goal setting between patients and therapists[71]) and illustrated that these markers can be leveraged to identify relational ruptures that were missed by therapists[72] (see also reliability results and another study showing more modest performance[73]). Importantly, one last study applied natural language processing on the lived experiences of 2,908 patients, showing that most patients point to relational functioning as the most important aim of psychotherapy[74,75].

Economic models (k = 2)

Although outside the scope of this review, two notable studies using economic models were considered relevant. These studies examined ways of navigating relationships, showing that acting 'unfairly' and being 'closed socially' are common in those with psychopathy[76] and borderline personality[77], respectively.

Study evaluation

This section reports key information on the risk of bias, validity, performance and transparency of reviewed studies (Table 2). Refer to Supplementary Tables 3–10 for more details.

Risk of bias. Of 48 empirical studies, 18 (38%) exhibited high risk, 15 (31%) moderate risk and 15 low risk of bias. High risk included all supervised machine learning, 4/7 unsupervised machine learning[11,71,73,78] and 4/8 dynamical systems studies[10,34,35,40]; moderate risk 1/7 unsupervised machine learning[72], 7/10 reinforcement learning[41,42,45,47,48,79,43] and 7/13 Bayesian studies[53,55,56,59–61,80]; and low risk 2/7 unsupervised machine learning[74,81]; 4/8 dynamical systems[36–39], 3/10 reinforcement learning[44,46,82] and 6/13 Bayesian studies[50–52,54,58,62].

Validity. Our validity assessment revealed four types of model: (1) data-driven models (5/9 dynamical systems[35–38,40], 9/18 Bayesian models[50–54,58,62,80] and notably all reinforcement learning models[41–48,79,82]), (2) theory-driven models (2/9 dynamical systems[28,30] and 3/19 Bayesian models[64,65,83]), (3) theory-driven models with strong generative validity (2/19 Bayesian models[57,63]) and (4) excellent models scoring high on all types of validity (2/9 dynamical systems[29,31] and 4/10 Bayesian models[55,60,61,84]).

Performance. From the 31 empirical theory-driven studies, 19 (61%) reported at least one performance metric, with 5 reporting only one metric (high bias)[37,50,53,56,60], 9 only two metrics (moderate bias)[38,41,42,52,55,61,62,79,85] and 5 three or more metrics (low bias)[44–46,51,59]. Critically, no theory-driven study reported reliability measures, and only three performed parameter recoverability tests. From the 17 data-driven (machine learning) studies, 14 (87%) reported at least one performance metric, with only 6 reporting discrimination metrics[66,67,69,71–73]. While all machine learning studies performed internal validation, only one performed external validation[86] and none assessed for calibration or clinical utility (net benefit).

Transparency. Of all 58 studies, only 12 (21%) made their code publicly available[28,38,44,46,50–52,54,61,63,74,82]. Of 48 empirical studies, only 9 (19%) made their data publicly available (these also included their code)[44,46,50–52,54,61,74,82], and only 2 (4%) preregistered their hypotheses[46,54].

Discussion

In this systematic review, we evaluated the utility of 60 computational studies in informing interpersonal dynamics of psychopathology. We found that theory-driven models were able to formalize historically elusive concepts (such as mentalizing), while data-driven models were able to systematically map these concepts on various psychopathology problems (Table 1). Despite this progress, several issues persisted regarding the performance, transparency and commensurability of computational approaches. In our discussion, we contextualize these challenges and provide a roadmap for addressing them in future research (Table 2).

Challenges faced by social computational psychiatry

Our analysis revealed three core challenges faced by social computational psychiatry (Table 2). First, in line with prior critiques[21], reporting practices were incomplete in theory-driven empirical studies (with no studies reporting task reliability and only three studies reporting parameter recoverability) and supervised machine learning studies (with no studies conducting calibration or external validation, limiting clinical application). Second, transparency was worryingly low, with only two studies preregistering hypotheses and few studies sharing their code (21%) and data (19%) online, pointing toward serious concerns on the reproducibility and replicability of computational patterns. Finally, method integration was low, with only two studies combining computational approaches[46,56], suggesting that there may be a risk of the field fragmenting into siloed research lines.

A roadmap forward

Although these challenges may paint a rather pessimistic view of the field so far, we believe that there exist at least four ways in which our field can productively move forward. A first way is to enhance transparency by embracing open science practice. Although transparency could be enforced in top-down ways (for example, funding agencies mandating open data and journals requiring at least open code), a more sustainable and arguably effective way may be to promote a culture of openness in a bottom-up manner[87]. We suggest that one tangible way of achieving this end is by embracing tutorial papers. Indeed, tutorial papers on Bayesian modeling already illustrate the transformative power of transparency: by releasing open-source code, expert teams on Bayesian methodologies (such as active inference[88] and hierarchical Bayesian approaches[89,90]) have lowered entry barriers and enabled many researchers to use complex models in diverse research contexts. Moreover, by detailing best research practices in particular computational approaches, tutorial papers can also enhance reproducibility by specifying precise reporting guidelines (for example, consistent model performance metrics and documentation of methodological pipelines). Importantly, these transparent research practices are not simply altruistic but have tangible benefits: in particular, they correlate with 'increased citations, media attention, potential collaborators, job opportunities and funding opportunities'[91]. These patterns highlight that open science practice is not a zero-sum game: it can benefit everyone by enabling researchers to come together and examine joint research goals using each other's computational frameworks.

A second way forward is to improve the reliability and validity of computational tasks by standardizing them and tailoring them to particular relational processes[26]. For instance, instead of having general tasks that are used to examine various psychological concepts, researchers could develop tasks that are specific to particular social concepts, such as epistemic trust (for example, participants judging trustworthiness) or mentalizing (for example, participants inferring mental states)[9,21,27]. In this way, computational modeling could be more effectively applied to decipher various computational problems that speak to well-defined psychological concepts. For instance, regarding epistemic trust, Bayesian modeling could be used to quantify belief rigidity (for example, rigidly expecting betrayal even after observing benevolence)[58], reinforcement learning could dissect maladaptive learning (for example, learning to trust 'bad agents' in volatile social contexts)[43] and dynamical systems could reveal unstable relational

https://doi.org/10.1038/s44220-025-00465-9

Table 1 | Theoretical synthesis of reviewed studies

Theme	Illustrative examples	Theoretical implications
Dynamical systems		
Reciprocal coupling	Mutual patient–therapist emotional influence marks strong alliance; unidirectional dominance signals rupture[10,31]	Alliance strength emerges when patient–therapist dyads are mutually influential
Interpersonal synchrony	Schizophrenia involves intact spontaneous but impaired intentional motor synchrony[36]. Personality pathology involves impaired spontaneous heart-rate synchrony[37]	Psychopathology can manifest as a lack of intentional or spontaneous interpersonal coordination
Determinism versus stochasticity	Effective therapy involves consistent affect stabilization with occasional perturbations (for example, novel emotional insights)[39,40]	Successful interventions balance between stable regulation and stochastic perturbation
Reinforcement learning		
Transdiagnostic social learning	Low self-esteem is linked to slower, more volatile social learning (as well as altered self-referential brain activity)[41,42]	Common interpersonal vulnerabilities emerge from a difficulty in updating social beliefs
Blurred self–other boundaries in borderline personality	People with borderline personality over-weigh social feedback yet fail to update beliefs based on disconfirming evidence[47,48]. Self–other belief confusion might underpin this rigidity[49]	Borderline pathology arises from a network of relational difficulties including extreme social sensitivity, belief rigidity and also blurred self–other boundaries
Blunted learning in depression	People with depression but not anxiety exhibit reduced social-prediction-error learning[44], which predicts real-world dwelling in negative social settings[45]	Depressive rigidity stems from a difficulty in updating beliefs, even in light of disconfirming evidence
Bayesian inference		
Depth of mental inferences	Deep mental inferences of the sort 'I think that you think that I think, and so on' are hypo-active in autism[55] but hyper-active in paranoia[57]	Various relational problems arise from either too little or too much thinking about other's mental states
Personality disorders as relational disorders	Generative models of the social order are disrupted in those with personality disorders[63], leading to polarized[84] and unstable[58] mental inferences	Personality disorders can be more specifically understood as relational disorders because they disrupt ways of relating to oneself and others
Free energy as a unifying framework	Psychopathy[65], attachment insecurity[83] and depression[64] can be modeled as alterations in the process of minimizing free energy (difference between perceived reality and personal beliefs)	Psychopathology can be viewed as a problem in the way people minimize the difference between what they perceive and what they believe
Machine learning		
Classification of attachment styles	Digital footprints such as emotional posts versus receiving likes or comments can classify attachment anxiety versus avoidance[66]	High-dimensional data contain latent relational signatures, supporting a data-driven taxonomy of interpersonal styles and vulnerabilities
Prediction of psychosocial outcomes	Low social support and isolation emerge as the strongest predictors of adolescent[70] and mid-life depression[69]	Predictive models underscore the centrality of interpersonal problems in mood disorders
Language as computational phenotype	Therapist nonjudgmental language patterns index alliance strength[11] and patient's narratives suggest that relational functioning is more important than symptom relief[74,75]	Natural language carries quantifiable markers of relational processes, positioning linguistic analyses as windows into therapeutic dynamics

patterns (for example, cyclical shifts of trusting/idealizing or distrusting/devaluing others)[28,63]. Crucially, machine learning could further enhance the ecological validity of these mechanisms by systematically linking them to real-life difficulties (for example, relational ruptures in psychotherapy)[73]. To ensure, however, that this translational potential is realized, computational tasks need to be modified to reflect, in a standardized manner, real-life social problems that matter most to patients[74].

A third and related way forward is to embrace the complementary nature of computational approaches and integrate them in both empirical and theoretical research. As mentioned earlier, only two studies in our review applied multiple computational methods in their investigations[46,56]. Although this lack of cross-method integration reflects siloed research traditions (whereby specific computational frameworks are prioritized for specific benefits), we believe that these frameworks hold untapped potential when applied in tandem to the same datasets. To illustrate, we consider the study by Barnby, Mehta and Moutoussis[46], which applied both reinforcement learning and Bayesian modeling to assess various computational problems in paranoia[46]. By jointly modeling priors, learning rates and social volatility, the authors showed that paranoia emerges not from a single computational parameter but rather from a pattern of parameters: specifically, a pattern that entailed negative beliefs about other's intentions, hypersensitivity to threatening behavior and perceptions of the social world as unstable[46]. These findings illustrate that social problems cannot be reduced to individual computational parameters because they are likely governed by a network of many such parameters[92]. Accordingly, our recommendation for future research is that a comprehensive understanding of mental disorders lies precisely in our capacity to apply diverse computational models to the same datasets to map how their parameters jointly generate mental health problems.

Importantly, such an integration could enable the construction of formal theories that are not limited to single mathematical paradigms but rather integrate concepts from various such paradigms. Indeed, although existing theories have posited notable social processes, including rigid beliefs, unstable learning rates and relational attractors (Table 1), a crucial limitation is that they do not formalize how these processes are interlinked–for instance, do prior beliefs promote learning asymmetries, which in turn stabilize rigid relational dynamics? Encouragingly, some evidence from our review implies that such an integration is possible: for example, both Bayesian and reinforcement learning studies on social learning have converged empirically to indicate that people with various 'interpersonal disorders' tend to be closed socially because they cannot update their negative beliefs in light of disconfirming evidence[48,54,58,59,77]. Despite these advances, however, no concepts beyond social learning have been tested using diverse methodologies[36,37,41,42,60], implying that more work is necessary to integrate diverse computational perspectives, assess how they jointly predict mental health difficulties and build formal theories that integrate them in a truly transtheoretical manner[27].

Table 2 | Systematic evaluation and agenda for future work

Evaluation	Suggestions
Lack of transparency and openness in methodology	Make materials (at least code but also data) openly available[91]
	Write tutorial papers that detail best research practices[87]
	Delineate exploratory from confirmatory research[91]
Lack of sufficiently similar studies to enable meta-analysis	Focus on replicating and then extending existing findings[22]
	Conduct theory-driven, scoping reviews to integrate similar findings[9]
	Harmonize theory with methods for specific research questions[109]
Incomplete psychometric reporting and weak validity and reliability	Report multiple performance metrics in machine learning research[106]
	Report task reliability and recoverability metrics in experimental research[21]
	Standardize tasks and tailor them to particular psychological processes[27]
Lack of cross-method integration in empirical research and theory-building	Apply multiple computational methods on the same dataset[46]
	Scrutinize research questions using diverse computational methods[46]
	Build formal theories that integrate different computational perspectives
Absence of patient narratives in computational modeling	Develop computational paradigms that reflect specific patient difficulties[26]
	Use natural language processing to systematize patient narratives[74]
	Build formal theories inspired by patient narratives[63]

Finally, we wish to highlight that such theories should not be limited to theory-driven computational processes but should also aim to incorporate data-driven patterns reflecting patients' actual lived experiences[75]. Although historically qualitative research has been perceived as subjective and anecdotal, recent advances in natural language processing have been embraced by our field to systematically extract meaningful themes from a wealth of unstructured narratives[93,94]. As our review illustrates, such approaches have revealed several notable insights, including that patients value relational functioning more than symptom relief[74,75], as well as that they build stronger alliances with therapists who understand them more accurately[11]. These findings imply that social relatedness is not a mere epiphenomenon of mental disorder but may instead be constitutive of said disorder: it defines what it means to be human and should perhaps have primacy in understanding both mental health and illness[26].

Implications, limitations and future directions

Our systematic analysis has a number of implications for future work, including the pressing need to address issues of reporting, transparency and theory-building in the field of social computational psychiatry (see Table 2 for an outline). Beyond these implications, though, at least two limitations must be noted. First, although we aimed to cover all computational models in the social field of computational psychiatry, we realized along the way that some under-represented models (namely, economic models) were missed. Future reviews could estimate the prevalence of these less widespread models and examine whether they hold utility in informing us about interpersonal psychopathology problems. Second, although we qualitatively synthesized findings from 58 studies, we could not meta-analyze them because of their vast differences in task design, computational analyses and reported effect size. As suggested earlier, studies need to converge in their methodologies to enable researchers to meta-analyze specific findings that speak about specific social difficulties (for example, trust, reciprocity, mentalizing and so on).

Conclusion

Our review suggests that computational modeling has a potential to advance our understanding of interpersonal psychopathology. Although some challenges in methodology and transparency do exist, we are hopeful that by systematically outlining this field, we can inform readers about its key insights (Boxes 1–4 and Table 1), main shortcomings (Table 2), as well as ways of productively engaging with the field moving forward (Table 2).

Methods

Our systematic review was prospectively registered (PROSPERO CRD42024488821) and adhered to the PRISMA guidelines[95]. Key methodological details are outlined in the next sections and are further elaborated in our preregistration.

Study search and selection

Five databases (MEDLINE, Embase, PsycINFO, Web of Science and Google Scholar) were systematically searched for eligible studies from their inception to 10 June 2025. Utilizing these databases for a systematic search has been shown to capture over 90% of psychological research[96]. No geographical or publication type restrictions were set[97]. Only public studies were searched. Relevant studies were inspected for further references.

Studies qualified for inclusion if they: (1) were written in English, (2) were either empirical or theoretical, (3) employed any of the following computational frameworks (Bayesian, reinforcement learning, dynamical systems and machine learning), (4) examined interpersonal dynamics and (5) examined any psychopathology. Studies were excluded if they: (1) did not examine any psychopathology (2) did not assess interpersonal dynamics or (3) did not employ any computational methodologies (see preregistration).

Two reviewers (O.Z. and C.F.) independently screened all retrieved records against the inclusion criteria. The same reviewers extracted data from included papers and reported them in Supplementary Table 1. Any disagreements regarding the inclusion of a study were resolved via full-text review and discussion by all authors.

Study evaluation

Included studies were examined in terms of their overall risk of bias and validity, performance and transparency in their computational modeling.

Risk of bias. Risk of bias was examined using the NIH risk assessment tool[98] (for theory-driven empirical studies) and the PROBAST risk assessment tool[99] (for data-driven empirical studies). Both instruments examine biases in sampling, measurement and analysis and accommodate our diverse study designs (see Supplementary Information 3 for details).

Validity. Validity was examined using the Validity Appraisal Guide for Computational Models (VAG-CM)[100], a tool that quantifies three types of validity in theory-driven modeling, namely, face, predictive and construct validity types, which were hereby renamed (without altering their meaning) as empirical, theoretical and generative types. Empirical validity examines how well a model fits on relevant datasets; theoretical validity evaluates whether computer simulations can yield patterns that resemble the outcomes of actual interventions; and generative validity tests whether data-generating processes are specific enough to illuminate the inner workings of psychopathology. Together, these validity types provide a thorough assessment of the validity of theory-driven modeling. Importantly, although the VAG-CM does not address the validity of data-driven modeling, it is worth noting that, by definition, such modeling scores highly on only empirical validity given its focus on predictive potency.

Performance. Performance was examined by tracking key metrics for theory-driven and data-driven models. For theory-driven models, we evaluated test–retest reliability (the consistency of parameter estimates across repeated assessments)[101], parameter recoverability (the maximum reliability achievable under controlled experimental conditions)[102] and model fit (for instance, the Bayesian and Akaike Information Criteria)[103]. For data-driven (machine learning) models, we assessed internal validation (the extent to which a model overfits a dataset)[104], external validation (the extent to which a model generalizes on external datasets)[105] and predictive performance[106], using established metrics that quantify discrimination (how well the model differentiates clinical from nonclinical populations), calibration (how well the predicted scores align with observed outcomes) and net benefit (the potential benefits versus harms of using the model)[107].

Transparency. Finally, the transparency of computational modeling was examined by tracking whether our reviewed studies endorsed three open science practices: (1) open data, (2) open code and (3) preregistered protocols. The centrality of these practices in transparent reporting is supported by comprehensive reviews on the topic[91,108].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data extracted from all included studies are contained in Supplementary Table 1.

Code availability

Code availability is not applicable, as this study is a systematic review.

References

1. Heider, F. Social perception and phenomenal causality. *Psychol. Rev.* **51**, 358–374 (1944).
2. Fonagy, P. Thinking about thinking: some clinical and theoretical considerations in the treatment of a borderline patient. *Int. J. Psychoanal.* **72**, 639–656 (1991).
3. Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).
4. Horowitz, L. M. & Strack, S. *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions* (John Wiley & Sons, 2010).
5. Wright, A. G., Pincus, A. L. & Hopwood, C. J. Contemporary integrative interpersonal theory: integrating structure, dynamics, temporal scale, and levels of analysis. *J. Psychopathol. Clin. Sci.* **132**, 263 (2023).
6. Wright, A. G., Ringwald, W. R., Hopwood, C. J. & Pincus, A. L. It's time to replace the personality disorders with the interpersonal disorders. *Am. Psychol.* **77**, 1085 (2022).
7. Hames, J. L., Hagan, C. R. & Joiner, T. E. Interpersonal processes in depression. *Ann. Rev. Clin. Psychol.* **9**, 355–377 (2013).
8. Zhou, L. et al. What do four decades of research tell us about the association between childhood adversity and psychosis: an updated and extended multi-level meta-analysis. *Am. J. Phys.* **182**, 360–372 (2025).
9. Barnby, J. M., Dayan, P. & Bell, V. Formalising social representation to explain psychiatric symptoms. *Trends Cogn. Sci.* **27**, 317–332 (2023).
10. Baker, A. Z., Peluso, P. R., Freund, R., Diaz, P. & Ghaness, A. Using dynamical systems mathematical modeling to examine the impact emotional expression on the therapeutic relationship: a demonstration across three psychotherapeutic theoretical approaches. *Psychother. Res.* **32**, 223–237 (2022).
11. Martinez, V. R. et al. Identifying therapist and client personae for therapeutic alliance estimation. In *Proc. Interspeech 2019* 1901–1905 (International Speech Communication Association, 2019); https://doi.org/10.21437/Interspeech.2019-2829
12. Adams, R. A., Huys, Q. J. M. & Roiser, J. P. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psych.* https://doi.org/10.1136/jnnp-2015-310737 (2015).
13. Bennett, D., Silverstein, S. M. & Niv, Y. The two cultures of computational psychiatry. *JAMA Psychiatry* **76**, 563–564 (2019).
14. Friston, K. J., Stephan, K. E., Montague, R. & Dolan, R. J. Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry* **1**, 148–158 (2014).
15. Montague, P. R., Dolan, R. J., Friston, K. J. & Dayan, P. Computational psychiatry. *Trends Cogn. Sci.* **16**, 72–80 (2012).
16. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
17. Huys, Q. J. M., Browning, M., Paulus, M. P. & Frank, M. J. Advances in the computational understanding of mental illness. *Neuropsychopharmacology* **46**, 3–19 (2021).
18. Zavlis, O. Computational approaches to mental illnesses. *Nat. Rev. Psychol.* **3**, 650 (2024).
19. Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J. & Friston, K. J. Bayesian inferences about the self (and others): a review. *Conscious Cogn.* **25**, 67–76 (2014).
20. Lockwood, P. L. & Klein-Flugge, M. C. Computational modelling of social cognition and behaviour-a reinforcement learning primer. *Soc. Cogn. Affect Neurosci.* **16**, 761–771 (2021).
21. Karvelis, P., Paulus, M. P. & Diaconescu, A. O. Individual differences in computational psychiatry: a review of current challenges. *Neurosci. Biobehav. Rev.* **148**, 105137 (2023).
22. Pike, A. C. & Robinson, O. J. Reinforcement learning in patients with mood and anxiety disorders vs control individuals: a systematic review and meta-analysis. *JAMA Psychiatry* **79**, 313 (2022).
23. Sullivan H. S. *The Interpersonal Theory of Psychiatry* (W W Norton & Co; 1953).
24. Pincus, A. L. A contemporary integrative interpersonal theory of personality disorders. *APA PsycNet* https://psycnet.apa.org/record/2005-02797-006 (2005).
25. Bolis, D., Dumas, G. & Schilbach, L. Interpersonal attunement in social interactions: from collective psychophysiology to interpersonalized psychiatry and beyond. *Phil. Trans. R. Soc. B.* **378**, 20210365 (2023).
26. Rhoads, S. A., Gu, X. & Barnby, J. M. Advancing computational psychiatry through a social lens. *Nat. Mental Health.* **2**, 1268–1270 (2024).

27. Hitchcock, P. F., Fried, E. I. & Frank, M. J. Computational psychiatry needs time and context. *Ann. Rev. Psychol.* **73**, 243–270 (2022).
28. Westermann, S. & Banisch, S. A formal model of affiliative interpersonality. *Clin. Psychol. Sci.* **13**, 43–68 (2025).
29. Tschacher, W., Haken, H. & Kyselo, M. Alliance: a common factor of psychotherapy modeled by structural theory. *Front. Psychol.* https://doi.org/10.3389/fpsyg.2015.00421 (2015).
30. Schiepek, G., Aas, B. & Viol, K. The mathematics of psychotherapy: a nonlinear model of change dynamics. *Nonlinear Dyn. Psychol. Life Sci.* **20**, 369–399 (2016).
31. Liebovitch, L. S., Peluso, P. R., Norman, M. D., Su, J. & Gottman, J. M. Mathematical model of the dynamics of psychotherapy. *Cogn. Neurodyn.* **5**, 265–275 (2011).
32. Gottman, J. M., Murray, J. D., Swanson, C. C., Tyson, R. & Swanson, K. R. *The Mathematics of Marriage: Dynamic Nonlinear Models* 1st edn (The MIT Press, 2005).
33. Peluso, P. R., Liebovitch, L. S., Gottman, J. M., Norman, M. D. & Su, J. A mathematical model of psychotherapy: an investigation using dynamic non-linear equations to model the therapeutic relationship. *Psychother. Res.* **22**, 40–55 (2012).
34. Diaz, P., Peluso, P. R., Freund, R., Baker, A. Z. & Pena, G. Understanding the role of emotion and expertise in psychotherapy: An application of dynamical systems mathematical modeling to an entire course of therapy. *Front. Psychiatr.* **14**, 980739 (2023).
35. Saint-Georges, C. et al. Do parents recognize autistic deviant behavior long before diagnosis? Taking into account interaction using computational methods. *PLoS ONE* **6**, e22393 (2011).
36. Varlet, M. et al. Impairments of social motor coordination in schizophrenia. *PLoS ONE* **7**, e29772 (2012).
37. Schreiber, A. M. et al. Disrupted physiological coregulation during a conflict predicts short-term discord and long-term relationship dysfunction in couples with personality pathology. *J. Abnorm. Psychol.* **129**, 433–444 (2020).
38. Hale, W. W. & Aarts, E. Hidden Markov model detection of interpersonal interaction dynamics in predicting patient depression improvement in psychotherapy: proof-of-concept study. *J. Affect. Disord. Rep.* **14**, 100635 (2023).
39. Tschacher, W. & Haken, H. Causation and chance: detection of deterministic and stochastic ingredients in psychotherapy processes. *Psychother. Res.* **30**, 1075–1087 (2020).
40. Paz, A. et al. Intrapersonal and interpersonal vocal affect dynamics during psychotherapy. *J. Consult. Clin. Psychol.* **89**, 227–239 (2021).
41. Will, G. J., Rutledge, R. B., Moutoussis, M. & Dolan, R. J. Neural and computational processes underlying dynamic changes in self-esteem. *eLife* **6**, e28098 (2017).
42. Will, G. J. et al. Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Transl. Psychiatry* **10**, 1–14 (2020).
43. Lenow, J., Cisler, J. & Bush, K. Altered trust learning mechanisms among female adolescent victims of interpersonal violence. *J. Interpers. Violence* **33**, 159–179 (2018).
44. Safra, L., Chevallier, C. & Palminteri, S. Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Comput. Biol.* **15**, e1007224 (2019).
45. Frey, A. L., Frank, M. J. & McCabe, C. Social reinforcement learning as a predictor of real-life experiences in individuals with high and low depressive symptomatology. *Psychol. Med.* **51**, 408–415 (2021).
46. Barnby, J. M., Mehta, M. A. & Moutoussis, M. The computational relationship between reinforcement learning, social inference, and paranoia. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1010326 (2022).
47. Shapiro-Thompson, R. et al. Modulation of trust in borderline personality disorder by script-based imaginal exposure to betrayal. *J. Personal Disord.* **37**, 508–524 (2023).
48. Fineberg, S. K. et al. Differential valuation and learning from social and nonsocial cues in borderline personality disorder. *Biol Psychiatry* **84**, 838–845 (2018).
49. Story, G. et al. 367. A computational signature of self-other mergence in borderline personality disorder. *Biol. Psychiatry* **93**, S242 (2023).
50. Lamba, A., Frank, M. J. & FeldmanHall, O. Anxiety impedes adaptive social learning under uncertainty. *Psychol. Sci.* **31**, 592–603 (2020).
51. Henco, L. et al. Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1008162 (2020).
52. Barnby, J. M., Raihani, N. & Dayan, P. Knowing me, knowing you: interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition* **225**, 105098 (2022).
53. Na, S. et al. Computational mechanisms underlying illusion of control in delusional individuals. *Schizophr. Res.* **245**, 50–58 (2022).
54. Barnby, J. M., Bell, V., Mehta, M. A. & Moutoussis, M. Reduction in social learning and increased policy uncertainty about harmful intent is associated with pre-existing paranoid beliefs: evidence from modelling a modified serial dictator game. *PLoS Comput. Biol.* (2020).
55. Yoshida, W. et al. Cooperation and heterogeneity of the autistic mind. *J. Neurosci.* **30**, 8815–8818 (2010).
56. Forgeot d'Arc, B., Devaine, M. & Daunizeau, J. Social behavioural adaptation in autism. *PLoS Comput. Biol.* **16**, e1007700 (2020).
57. Alon, N. et al. (Mal)adaptive mentalizing in the cognitive hierarchy, and its link to paranoia. *Comput. Psychiatry* **8**, 159–177 (2024).
58. Siegel, J. Z., Curwell-Parry, O., Pearce, S., Saunders, K. E. A. & Crockett, M. J. A computational phenotype of disrupted moral inference in borderline personality disorder. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* **5**, 1134–1141 (2020).
59. Story, G. W. et al. J. A social inference model of idealization and devaluation. *Psychol. Rev.* https://doi.org/10.1037/rev0000430 (2023).
60. Xiang, T., Ray, D., Lohrenz, T., Dayan, P. & Montague, P. R. Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Comput. Biol.* **8**, e1002841 (2012).
61. Hula, A., Vilares, I., Lohrenz, T., Dayan, P. & Montague, P. R. A model of risk and mental state shifts during social interaction. *PLoS Comput. Biol.* **14**, e1005935 (2018).
62. Xiao, F. et al. Understanding guilt-related interpersonal dysfunction in obsessive-compulsive personality disorder through computational modeling of two social interaction tasks. *Psychol. Med.* **53**, 5569–5581 (2023).
63. Zavlis, O., Fonagy, P., Moutoussis, M. & Story, G. W. A generative model of personality disorder as a relational disorder. *OSF* https://osf.io/wh6na/download (2024).
64. Constant, A., Hesp, C., Davey, C. G., Friston, K. J. & Badcock, P. B. Why depressed mood is adaptive: a numerical proof of principle for an evolutionary systems theory of depression. *Comput. Psychiatr.* **5**, 60–80 (2021).
65. Prosser, A., Friston, K. J., Bakker, N. & Parr, T. A Bayesian account of psychopathy: a model of lacks remorse and self-aggrandizing. *Comput. Psychiatry* **2**, 92–140 (2018).

66. Kang, B. et al. Towards understanding relational orientation: attachment theory and facebook activities. In *Proc. 18th ACM Conference on Computer Supported Cooperative Work and Social Computing* 1404–1415 (2015); https://doi.org/10.1145/2675133.2675211
67. Doborjeh, Z. et al. Investigation of social and cognitive predictors in non-transition ultra-high-risk' individuals for psychosis using spiking neural networks. *Schizophrenia* **9**, 1–10 (2023).
68. Haghish, E. F. et al. Unveiling adolescent suicidality: holistic analysis of protective and risk factors using multiple machine learning algorithms. *J. Youth Adolescence.* **53**, 507–525 (2024).
69. Handing, E. P., Strobl, C., Jiao, Y., Feliciano, L. & Aichele, S. Predictors of depression among middle-aged and older men and women in Europe: a machine learning approach. *Lancet Reg. Health Eur.* **18**, 100391 (2022).
70. Wang, C. et al. Risk and protective factors of depression in family and school domains for chinese early adolescents: an association rule mining approach. *Behav. Sci.* **13**, 893 (2023).
71. Atzil-Slonim, D. et al. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy* **58**, 324–339 (2021).
72. Tsakalidis, A. et al. Automatic identification of ruptures in transcribed psychotherapy sessions. In *Proc. 7th Workshop on Computational Linguistics and Clinical Psychology: Improving Access* (ed. Association for Computational Linguistics) 122–128 (2021); https://doi.org/10.18653/v1/2021.clpsych-1.15
73. Goldberg, S. B. et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. *J. Counsel. Psychol.* **67**, 438–448 (2020).
74. Zavlis, O., Fonagy, P. & Luyten, P. The most important aims of psychotherapy: to love, to work, and to find meaning. *Lancet Psychiatry* **12**, 173–174 (2025).
75. Ladmanová, M. et al. Client-identified outcomes of individual psychotherapy: a qualitative meta-analysis. *Lancet Psychiatry* **12**, 18–31 (2025).
76. Driessen, J. M. A., van Baar, J. M., Sanfey, A. G., Glennon, J. C. & Brazil, I. A. Moral strategies and psychopathic traits. *J. Abnorm. Psychol.* **130**, 550–561 (2021).
77. Barnby, J. M. et al. Self-other generalisation shapes social interaction and is disrupted in borderline personality disorder. *eLife* https://doi.org/10.7554/eLife.104008.1 (2025).
78. Tasca, A. N. et al. Detecting defense mechanisms from Adult Attachment Interview (AAI) transcripts using machine learning. *Psychother. Res.* **33**, 757–767 (2023).
79. Story, G. W. et al. A computational signature of self-other mergence in borderline personality disorder. *Transl. Psychiatry* **14**, 473 (2024).
80. Thomas, L., Lockwood, P. L., Garvert, M. M. & Balsters, J. H. Contagion of temporal discounting value preferences in neurotypical and autistic adults. *J. Autism Dev. Disord.* **52**, 700–713 (2022).
81. Xu, Y. et al. Detecting premature departure in online text-based counseling using logic-based pattern matching. *Internet Interv.* **26**, 100486 (2021).
82. Contreras-Huerta, L. S., Lockwood, P. L., Bird, G., Apps, M. A. J. & Crockett, M. J. Prosocial behavior is associated with transdiagnostic markers of affective sensitivity in multiple domains. *Emotion* **22**, 820–835 (2022).
83. Cittern, D., Nolte, T., Friston, K. & Edalat, A. Intrinsic and extrinsic motivators of attachment under active inference. *PLoS ONE* https://doi.org/10.1371/journal.pone.0193955 (2018).
84. Story, G. W. et al. A social inference model of idealization and devaluation. *Psychol. Rev.* **131**, 749–780 (2024).
85. Barnby, J. M. et al. Paranoia, sensitization and social inference: findings from two large-scale, multi-round behavioural experiments. *R. Soc. Open Sci.* **7**, 191525 (2020).
86. Antonucci, L. A. et al. Machine learning-based ability to classify psychosis and early stages of disease through parenting and attachment-related variables is associated with social cognition. *BMC Psychol.* **9**, 1–87 (2021).
87. Hesse, B. W. Can psychology walk the walk of open science? *Am. Psychol.* **73**, 126 (2018).
88. Smith, R., Friston, K. J. & Whyte, C. J. A step-by-step tutorial on active inference and its application to empirical data. *J. Math. Psychol.* **107**, 102632 (2022).
89. Daunizeau, J. et al. Observing the observer (I): meta-Bayesian models of learning and decision-making. *PLoS ONE* https://doi.org/10.1371/journal.pone.0015554 (2010).
90. Daunizeau, J. et al. Observing the observer (II): deciding when to decide. *PLoS ONE* https://doi.org/10.1371/journal.pone.0015555 (2010).
91. McKiernan, E. C. et al. How open science helps researchers succeed. *eLife* **5**, e16800 (2016).
92. Durstewitz, D., Huys, Q. J. & Koppe, G. Psychiatric illnesses as disorders of network dynamics. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* **6**, 865–876 (2021).
93. Nour, M. M. & Huys, Q. J. Natural language processing in psychiatry: a field at an inflection point. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* **8**, 979–981 (2023).
94. Malgaroli, M., Hull, T. D., Zech, J. M. & Althoff, T. Natural language processing for mental health interventions: a systematic review and research framework. *Transl. Psychiatry* **13**, 309 (2023).
95. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int. J. Surg.* **88**, 105906 (2021).
96. Bramer, W. M., Rethlefsen, M. L., Kleijnen, J. & Franco, O. H. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst. Rev.* **6**, 245 (2017).
97. Mayo-Wilson, E. et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *J. Clin. Epidemiol.* **91**, 95–110 (2017).
98. Study quality assessment tools. *National Heart, Lung, and Blood Institute* https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools (2021).
99. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).
100. Nunes, A. et al. A critical evaluation of dynamical systems models of bipolar disorder. *Transl. Psychiatry.* **12**, 416 (2022).
101. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
102. Mkrtchian, A., Valton, V. & Roiser, J. P. Reliability of decision-making and reinforcement learning computational parameters. *Comput. Psychiatry* **7**, 30–46 (2023).
103. Palminteri, S., Wyart, V. & Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
104. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal–external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
105. Altman, D. G. & Royston, P. What do we mean by validating a prognostic model? *Stat. Med.* **19**, 453–473 (2000).
106. Seyedsalehi, A. & Lennox, B. Predictive tools in psychosis: what is 'good enough'? *Nat. Rev. Neurol.* https://doi.org/10.1038/s41582-023-00787-1 (2023).

107. Moons, K. G. M. et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1–W73 (2015).
108. Vicente-Saez, R. & Martinez-Fuentes, C. Open Science now: a systematic literature review for an integrated definition. *J. Bus. Res.* **88**, 428–436 (2018).
109. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).

Acknowledgements
The authors received no specific funding for this work.

Author contributions
O.Z., G.S., P.F. and M.M. jointly conceived and preregistered the study. O.Z. and C.F. independently screened papers. O.Z. wrote the main paper and its Supplementary Information, under the supervision of M.M. G.S., C.F., P.F. and M.M. edited the paper extensively thereafter and helped address all reviewer comments.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-025-00465-9.

Correspondence and requests for materials should be addressed to Orestis Zavlis.

Peer review information *Nature Mental Health* thanks Andreea Diaconescu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Orestis Zavlis[1] ✉, Giles Story[2], Claire Friedrich[3], Peter Fonagy[1] & Michael Moutoussis[4]

[1]Unit of Psychoanalysis, Department of Psychology and Language Sciences, University College London, London, UK. [2]Max Planck-University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, UK. [3]Department of Psychiatry, University of Oxford, Oxford, UK. [4]Institute of Neurology, Department of Imaging Neuroscience, University College London, London, UK.
✉e-mail: orestis.zavlis.23@ucl.ac.uk

nature portfolio

Corresponding author(s): Orestis Zavlis

Last updated by author(s): 15/06/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A (there was no data collection in this systematic review)
Data analysis	N/A (there were no statistical analyses in this systematic review)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data extracted from all included studies are contained in Table S1.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The reviewed studies only considered biological sex, not gender identity. This was stated in Table S1.
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	Relevant population characteristics from reviewed studies are included in Table S1.
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Qualitative systematic review
Research sample	N/A
Sampling strategy	N/A
Data collection	N/A
Timing	N/A
Data exclusions	N/A
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A