Computational modeling of interpersonal dynamics in psychopathology: A systematic review and agenda for future work

Orestis Zavlis, Giles Story, Peter Fonagy, Michael Moutoussis

ABSTRACT

Interpersonal dynamics have long been acknowledged as instrumental to the generation and alleviation of psychopathology problems. Recent computational approaches have been argued to be uniquely suited for investigating such dynamics either with exploratory models (machine learning) or confirmatory models (generative modeling). However, the utility of such models to the study of interpersonal problems has not yet been scrutinized. We thus conducted a systematic review to assess the validity, reliability, and openness of computational models to the study of interpersonal psychopathology problems. Candidate studies (n = 2,944), including peer-reviewed conference manuscripts, were derived from five databases (MEDLINE, Embase, PsycINFO, Web of Science, Google Scholar) up to September 2024. A total of 55 studies met inclusion criteria and were assessed for their computational approach, validity, reliability, and openness. Results indicated that Bayesian modeling was the most commonly used approach (k=19), followed by machine learning (k=16), dynamical systems modeling (k=10), and reinforcement learning (k=10). Quality assessments revealed considerable heterogeneity in the validity of these approaches, with some scoring high primarily on empirical validity (reinforcement learning), others on theoretical validity (dynamical systems), and yet others on generative validity (Bayesian models). Finally, and most strikingly, few studies reported comprehensive performance metrics and even fewer studies adopted open science practices (specifically, 2 pre-registered their hypotheses and 8 shared their data and code online). We discuss these matters and conclude with more optimistic messages regarding how, when rigorously and openly conducted, computational approaches have the potential to advance the field of interpersonal psychopathology by enabling us to formalize and examine historically elusive social concepts (like mental representations of the self and others) and their role in psychopathology problems.

INTRODUCTION

Interpersonal dynamics refer to the ways we relate to 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 (i.e., in the sense that they are overly skewed or rigid), such dynamics can spiral downward into various psychopathologies, including personality,[6] emotional,[7] and eating psychopathologies.[8]

Although at face value intractable (given their intersubjective nature), interpersonal dynamics can be amenable to psychological experimentation, particularly via the use of computational tools. For example, Bayes rule can be used to obtain the most likely states that underlie a set of observations (such as the probability that a participant has low 'self-esteem' based on their evaluations).[9] Similarly, mismatches between what participants 'expect' and what they 'get' in reinforcement learning paradigms can be used to explore how they learn about themselves and others (revealing, for instance, that people with borderline personality disorder have difficulty learning about others because they place too much emphasis on themselves).[10] Finally, these approaches can be extended over time, casting humans as agents who act so as to minimize their 'free energy' (a statistical quantity of predictive error);[11] or behave according to 'laws' in differential and difference equations[12] (see Box 1 for a primer on these methods). These advances are part of *computational psychiatry*, a burgeoning field that aims to empirically examine and theoretically define mental disorders in terms of computational dysfunctions, rather than verbal descriptions.[13–19]

Although the popularity of computational psychiatry is rising rapidly, no systematic assessment has been conducted on its validity and reliability within the field of interpersonal dynamics. Moreover, previous reviews of computational modeling in related fields have focused mainly on Bayesian[20] and reinforcement learning paradigms[21]. Since then, however, computational psychiatry has expanded its scope to include additional theory-driven methods (such as dynamical systems) and data-driven methods (such as machine learning).[18] An up-to-date review of all such methods, therefore, is warranted. A related concern, more recently expressed, is about the effectiveness of these computational approaches over traditional statistical ones. On this, a recent review by Karvelis and colleagues (2023)[22] reported that computational studies tend to exhibit poor psychometric qualities (e.g., low reliability), implying that a systematic review of the extant computational literature is long overdue.

In this paper, we present the first systematic review of computational modelling of interpersonal dynamics in psychopathology. To cover the entire spectrum of computational methods, we include both theory-driven (generative) and data-driven (machine learning) approaches. Our primary aim is to evaluate whether these computational methods offer novel insights into the study of interpersonal dynamics. To this aim, we outline four specific objectives. First, to parse out the methodological heterogeneity of this field, we estimate the frequency with which the four main computational psychiatry approaches are utilised: (1) Dynamical Systems, (2) Approximate Bayesian models, (3) Reinforcement Learning models, and (4) Machine Learning algorithms. Second, to examine the validity of these approaches in both theoretical and applied settings, we employ a validated instrument that quantifies their empirical, theoretical, and generative validity.[23] Third, in response to recent concerns regarding the reliability of these models,[22] we assess the performance metrics for both generative (computational) and discriminative (machine learning) models. Finally, to examine the accessibility of these models, we track the adoption of open science practices, which are critical for ensuring the broad dissemination and validation of computational models across diverse studies and research contexts.

METHODS

Our systematic review was prospectively registered (PROSPERO registration number CRD42024488821) and adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[23] Key methodological details are outlined in the following sections and are further elaborated within our preregistration and Supplement.

Study search and selection

Five databases were searched for eligible studies from their inception to September 1st, 2024: MEDLINE, Embase, PsycINFO, Web of Science, Google Scholar. Utilizing these databases for a systematic search has been shown to capture over 90% of psychological research.[24] No geographical or publication type restrictions were made (see Supplement SI). Non-public studies were not searched.[25] Relevant studies were inspected for further references.

Studies qualified for inclusion if they: (1) were written in English; (2) were either empirical or theoretical in nature; (3) employed any of the following four computational frameworks (approximate Bayesian, reinforcement learning, dynamical systems, machine learning); (4) examined relational dynamics (that is, dynamics involving a person in relation to themselves or others, including artificial agents or human agents); and (5) were conducted within a psychopathology setting (see preregistration).

Two reviewers independently screened the retrieved records against the inclusion criteria, starting with titles and abstracts, and proceeding to full texts when necessary. The same reviewers then extracted and reported pertinent data (see next section) from included full-texts in Table 1. Any disagreements regarding the inclusion of a study were resolved through a full-text review and discussion by all authors.

Data extraction

Included studies were assessed to identify key theoretical, methodological, and clinical characteristics, which are summarized within Table 1. Given the diversity in study designs, computational models, and clinical outcomes, formal assessments of publication bias were not feasible. Nonetheless, three types of quality assessments were feasible: (1) validity, (2) reliability, and (3) open science practice.

Validity

To assess the validity of generative models, we employed the validity appraisal guide for computational models (VAG-CM), a validated instrument for evaluating three types of validity in disparate computational models which can vary in many ways, including in the phenomena they aim to explain (e.g., neuronal versus psychological dynamics) and the mathematical frameworks they employ (e.g., Bayesian versus dynamical systems models).[26] The VAG-CM, therefore, accommodates our inclusion of various computational frameworks (with the only exception being machine learning, the validity of which cannot be evaluated with this instrument given its data-driven nature).

The VAG-CM assesses three types of validity which can be defined as follows.[a] First, *empirical validity* can be defined as the extent to which the computational model can be *fitted* on data to explain a phenomenon of interest. (For example, if a model aims to explain borderline personality, then its empirical validity depends on its ability to 'fit well' data from this psychopathology.[27]) Second, *theoretical validity* can be defined as the extent to which a simulated intervention yields patterns that resemble the outcomes of an actual intervention. (For instance, if a model parameter is intended to quantify emotional sensitivity, then its appropriate tuning is supposed to alleviate simulated borderline psychopathology.[28]) Finally, *generative validity* can be defined as the degree to which the model's data-generating process

is specific enough to provide insights into the mechanistic underpinnings of a phenomenon of interest. (For example, if a model is designed to capture the psychology (but not biology) of emotional instability, then its mathematical architecture must be specific enough to inform us about the psychological mechanisms that underlie emotional instability.[29] More importantly, if this model architecture is shown to outperform alternative ones (via model comparison procedures), then its data-generating process is considered more valid.

[a]Although the authors of the VAG-CM guide name the three validity types as *face*, *predictive*, and *construct*, we chose to rename them (without altering their meaning) as *empirical*, *theoretical*, and *generative*, respectively, for at least two reasons. First, the former nomenclature has a long history in the psychometric science and a clear statistical interpretation, which does not exactly apply to generative modeling. Second, the latter nomenclature appears to be closer to the (computational) meaning of each validity type.[26]

Reliability

The reliability of computational modeling was quantified through various performance metrics, which differed across generative and discriminative models. For generative models, performance metrics included (1) test-retest reliability, (2) parameter recoverability, and (3) fit indices. First, test-retest reliability is usually quantified through the intraclass correlation coefficient (ICC), which is the ratio of between-subjects variance to total variance.[30] Due to the different operationalizations of the ICC, and the potential biases in some methods (such as Pearson's correlation coefficient, which assumes normality), it is recommended to specify how ICC estimates were calculated, including both statistical and software details.[31]

Second, parameter recoverability refers to the ‘maximum’ reliability that a computational model can achieve in a specific experimental setting.[22] This ‘upper bound’ of reliability can be derived by simulations that pin model’s predictions against simulated task versions. Recent simulation studies recommend the absolute ICC as a measure of parameter recoverability, because it is less sensitive to errors compared to traditional measures like Person and rank correlations.[32]

Finally, fit indices include (but are not limited to) information criteria, such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC), which are the most frequently employed in computational studies.[27]

For discriminative (here, machine learning, ML) models, guidelines recommend the reporting of (1) internal validation, (2) external validation, and (3) multiple measures of model performance.[33–35] First, internal validation aims to test the extent to which parameter estimates are contaminated by idiosyncrasies in measurement.[36] Common approaches here include cross-validation (i.e., splitting the data into training and testing subsamples) and bootstrapping (i.e., re-estimating the model using different subsets of data).[35] Internal validation is considered essential during the initial stages of model development.[34]

Second, external validation examines the generalizability of parameter estimates in external datasets. These external datasets could be based on similar (patient) populations but different time points (temporal validation) or different locations (geographical validation).[37,38] Although not essential during the initial stages of model development, external validation is considered vital for subsequent stages of thorough validation.[36]

Finally, ML performance outcomes include discrimination (how well the model differentiates clinical from nonclinical populations), calibration (how well do predicted scores align with the observed ones), and net benefit (the potential benefits against the potential harms of using the model). Example metrics include, but are not limited to, sensitivity, specificity, false positive/negative rates, accuracy, and F1 scores. Established guidelines recommend the reporting all of these metrics because together they paint a more comprehensive picture of model performance.[34,39]

Open Science

A recent systematic review defined Open Science as the *"transparent and accessible knowledge that is shared and developed through collaborative networks"*.[40] Such knowledge includes, but is not limited to, (1) open data, (2) open code, and (3) pre-registered protocols.[41] We sought to examine whether computational studies embrace these practices by reporting their study materials openly and transparently on Open Science platforms.

RESULTS

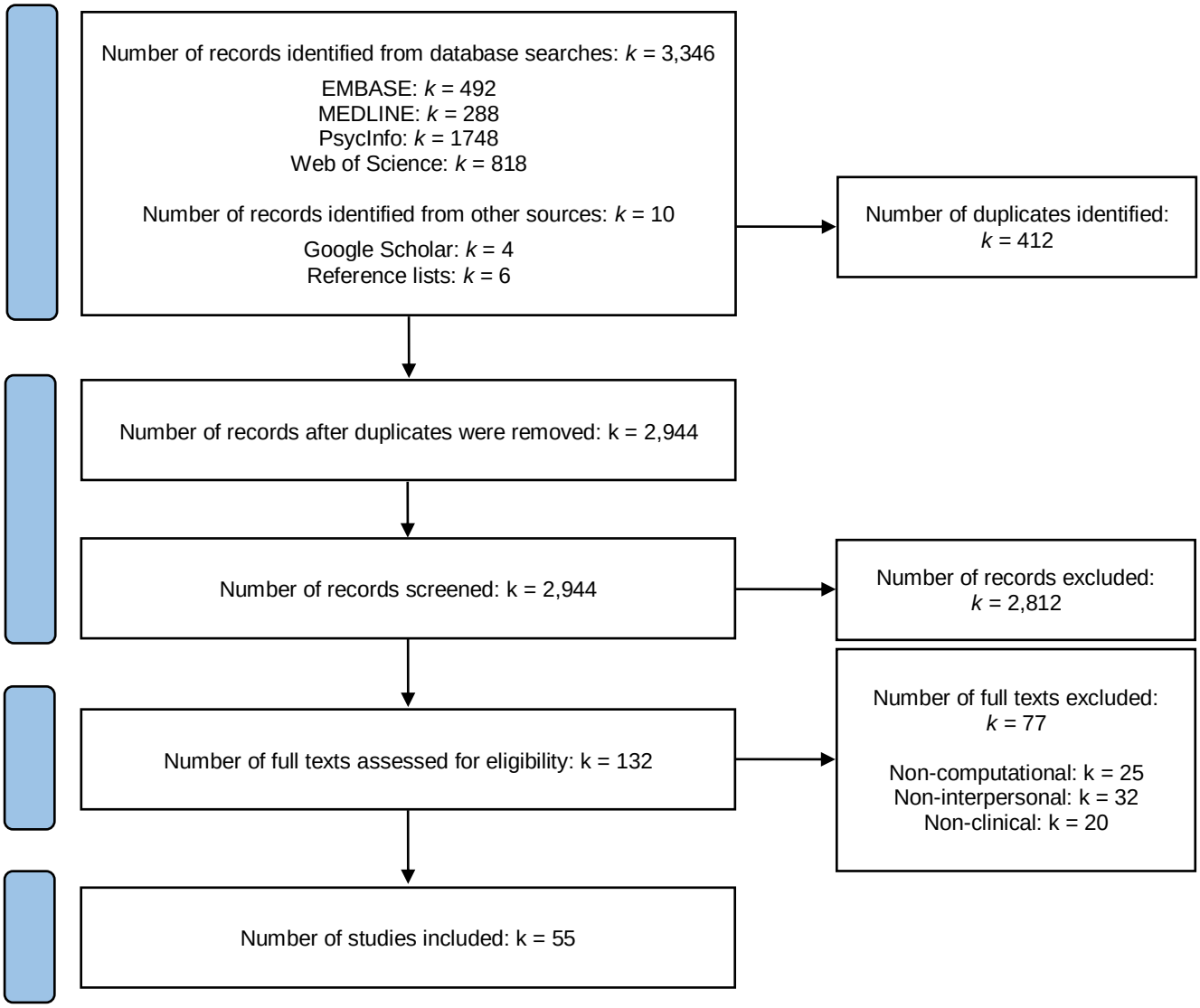
Figure 1 presents our study selection process using a PRISMA flow diagram. Overall, we obtained 3,356 records. Upon removing duplicate entries, 2,944 studies were screened against inclusion criteria. From these, 2,812 studies were excluded, leaving 132 necessitating full text assessment. Table 1 presents the final 55 studies that met our inclusion criteria.

[TABLE 1 about here]

Study characteristics

Results indicate an exponential growth of computational studies of interpersonal dynamics in the fields of psychopathology. Although the earliest study dates back to 2010, most studies were published between 2014 and 2019. The period 2020-2023, in particular, saw the highest volume of studies (44, 80%), reflecting the recent popularity of computational psychiatry.[18] Bayesian modeling was the most commonly used approach (k=19), followed by machine learning (k=16), dynamical systems modeling (k=10), and reinforcement learning (k=10). Interestingly, most computational modeling remained at the psychological level (k=50, 91%), with only 5 (9%) studies being applied at the biological level.[42–46] Below, we outline these patterns in more detail and discuss their emerging clinical themes.

Figure 1. PRISMA flowchart.



Random Dynamical Systems (k=10)

Therapeutic relationship (k=6). Six studies employed dynamical systems modeling to formalize the therapeutic relation between patients and therapists. Liebovitch et al. (2011)[47] modified the difference equations of Gottman and colleagues,[48] which were meant for marriage partners, to apply them to the therapeutic encounter. The outcome was a set of two differential equations that described how the therapists' emotional state influences the patient (and vice versa). By numerically integrating these equations, the authors simulated realistic social situations; for instance, attractors wherein the therapeutic dyad gets 'stuck' in either predominantly positive or predominantly negative relational states.[47] Inspired by this work, three studies applied Liebovitch et al.'s two-dimensional system in either *in silico* or *in vitro* investigations. First, Peluso et al. (2012)[49] simulated data to reveal three realistic relational patterns: (1) the highly responsive client, (2) the highly (emotionally) affected therapist, and (3) the highly influential therapist-client dyad (which gives rise to complex system dynamics). Second, Baker et al. (2021)[50] applied the two-dimensional system on time-series data, demonstrating that clinicians with different theoretical orientations exhibit similar relational dynamics during the treatment of the same two patients. Finally, Diaz et al. (2023)[51] extended these patterns by showing that the therapeutic relation may start with a negative relational attractor but unfold into a positive one over the course of several sessions.

Beyond Liebovitch et al.'s (2011) two-dimensional system,[47] two more dynamical models of therapeutic relations were detected. The first one by Schiepek et al. (2016)[52] was a complex system of five nonlinear difference equations (with 16 functions) that accounted for a range of psychotherapeutic phenomena, including self-efficacy, mentalizing, and emotion regulation. The second one by Tschacher et al. (2015)[53] was a unidimensional system, which argued that human relationships are characterized by two opposing tendencies (distinction; participation) and that therapy operates by altering the patient's attractors of these states, allowing them to more flexibly individuate while also socially participate.

***Interpersonal synchrony* (*k*=4).** Despite the history of dynamical systems modeling in the field of interpersonal synchrony, only four studies were found to employ this methodology within a psychopathology setting. The first study was detected in 2012 and involved the use of coupled oscillators to study both unintentional (that is, spontaneous) and intentional (that is, instructed) motor coordination dynamics in individuals with schizophrenia.[42] The results here revealed that the schizophrenia group showed intact unintentional coordination; but severely impaired intentional coordination, which was attributed to low information transmission from clinical to nonclinical participants. A second study applied the Fokker-Planck equation to show that less determinism in respiratory synchrony (across client and therapist) was related to better client-rated progress and therapist-rated alliance.[43] Relatedly, a third investigation indicated that the relational instabilities of people with personality disorder are associated with a disrupted physiological co-regulation between them and their romantic partners.[44] Finally, these patterns were extended in a psychotherapy study, which indicated that therapists who achieve greater synchrony with their clients (by 'pulling' their client's vocal arousals toward their own baseline) achieve superior session outcomes.[45]

Approximate Bayesian Inference (k=19)

Social Inferences (k=15). Fifteen studies assessed social inferences (probabilistic statements about the self and other) in individuals with various psychiatric disorders. Three of these studies investigated social inferences in individuals with autism spectrum disorders (hereon ASD). The first study was an early investigation by Yoshida and colleagues (2010),[54] which indicated that ASD individuals exhibit impaired mentalizing and iterative planning abilities ('I think that you think that I think' and so on) in a social interaction task. A second study by d'Arc, Devaine, and Daunizeau (2020)[55] extended these patterns by showing that they were only evident when ASD people were playing against artificial agents with sophisticated mental models (who shifted their behavioural repertoire across the task). Interestingly, when ASD individuals were playing against non-sophisticated artificial agents (who showed more habitual behavioural responses), they outperformed nonclinical controls, suggesting that they may only be impaired in social situations where their relational style stands in contrast to that of their partner.[55] Further support that the ASD phenotype is not entirely socially impaired came from a final study, which indicated that patterns of social contagion were equally impactful on ASD and (matched) nonclinical populations.[56]

Three studies investigated similar social inferences in nonclinical individuals who scored high in paranoid or delusional traits. Two studies came from Barnby and colleagues who indicated that, compared to individuals with low paranoia, those with high paranoia exhibit (1) greater uncertainty about others' actions[57] and (2) impaired updating of social representations.[58] A final study extended these patterns by indicating that individuals who score highly on delusional traits are more likely to believe that they can control others, particularly in unstable environments.[59]

Six studies examined social inferences in individuals with personality disorders. Two of these studies examined cooperation patterns in economic tasks, revealing that people with borderline personality disorder (BPD) exhibit (1) low 'depth' of mentalizing (that is, low reasoning of the sort 'I think that you think that I think'),[60] (2) an inability to detect irritation in their interlocutors,[61] and (3) low levels of guilt.[61] Relatedly, a third study indicated that people with obsessive compulsive personality disorder also tend to exhibit low levels of *guilt aversion* during social cooperation.[62] The final three studies examined moral inferences in individuals with BPD and psychopathy, revealing that the former have a general propensity to either idealize or devalue other agents,[63] as well as be certain and rigid when doing so;[64] while the latter are characterized by a pattern of low inequality, but not guilt, aversion.[65]

Three final studies examined social inferences in individuals with internalizing disorders. The first one indicated that anxiety impairs social inferences under conditions of uncertainty (by impairing social learning).[66] The second one applied a hidden Markov model on non-verbal behaviours to reveal that dynamics of hyperfocus during psychotherapy are associated with better treatment outcomes in people with depression.[67] A final study compared individuals with depression, BPD, and psychosis in social and non-social tasks, revealing that the latter (BPD and psychosis) are more interpersonally sensitive because they tend to weigh social information more strongly when making judgements of others.[68]

Free Energy perspectives (k=3). Although many studies verbally discussed the Free Energy principle, only three of them applied it mathematically to interpersonal dynamics. All studies employed a partially-observable Markov decision model that minimized Free Energy (a statistical quantity of uncertainty) in order to explain: (1) psychopathy (by casting it as a remorseless and aggrandizing self),[69] (2) attachment (by casting it as a process that minimizes epistemic uncertainty regarding caregiver responsivity),[70] and (3) depression (by casting it as a disorder of social alienation).[71] All of these studies were based on computer simulations.

Interpersonal synchrony. One final study used Markov modeling of videotaped interactions between infants and their caregivers, revealing that deviant autistic behaviours appear before 18 months of age and can be accurately identified by parental responses.[72]

Reinforcement Learning (k=10)

Transdiagnostic social learning. Five studies examined social learning, transdiagnostically, on psychopathology-relevant endophenotypes. Two of these indicated that *low self-esteem* is associated with slower social learning and more volatile self-beliefs, which mapped onto a dimension of interpersonal vulnerability, psychologically, and key brain regions (such as insula), neurobiologically.[46,73] A related study on adolescent victims of *interpersonal violence* revealed opposite patterns: assault frequency was related to high social learning (of trusted human faces) and this association was moderated by a high preference of stochasticity, pointing to patterns of credulity.[74] A third study indicated that markers of *emotional sensitivity*, but not psychiatric symptomatology, were predictive of prosocial tendencies (learning to avoid harming others).[75] Finally, a study on *paranoia* employed both Bayesian and reinforcement learning modeling to show that paranoia is associated with a general uncertainty (and thus impaired learning) over the state of the world.[76]

Social sensitivities in borderline personality disorder. Three studies examined social learning in BPD, revealing patterns of social sensitivity. The first study by Fineberg et al. (2018)[77] suggested that BPD patients tend to weigh social cues more heavily and expect greater volatility in a social evaluation task, two computational patterns that might explain BPD's interpersonal problems. A second study replicated these patterns in an imagery task,[78] while a final study revealed that patients with BPD have difficulty separating themselves from others because they attribute to others beliefs that they hold about themselves.[10]

Slower social learning in depression. Two final studies examined how internalizing psychopathology impacts social learning, revealing that people who score high on depression, but not anxiety, exhibit slower social learning,[79] a computational deficit that was associated with greater time spent in negative social situations (beyond depressive scores).[80]

Machine Learning (k=16)

Classification studies (k=2). Two studies used machine learning methods in a classificatory way. The first one attempted to classify attachment styles based on Facebook posts (from 640 users), revealing interesting predictors of attachment anxiety (being 'highly responsive') and avoidance (receiving many likes and comments).[81] The second study showed that psychosis can be classified modestly well using attachment variables, though its sample was small (N=34 clinical and N=71 nonclinical individuals).[82]

Predictive modeling studies (k=7). Seven studies used relatively large samples to examine psychosocial predictors of various interpersonal conditions. One study revealed that interpersonal factors (along with cognitive and emotional ones) were important prospective predictors of risk for psychosis in a sample of 90 ultra-high-risk patients vs 81 controls.[83] Another study on a representative Norwegian sample (N=173,644) showed that interpersonal variables are the strongest predictors of adolescent suicidal attempts.[84] Three studies assessed depression, indicating that factors relating to social connectedness are important predictors; in particular, social isolation was the strongest predictor of middle-age depression (N=67,603),[85] whereas parental support was the strongest predictor (even more so than peer support) for adolescent depression, with N=5952[86] and N=2445.[87] Finally, two studies assessed antisocial behaviour, revealing that early emotional and physical abuse,[88] as well as current relational problems (like interpersonal mistrust and alexithymia)[89] are its strongest predictors.

Natural language programming (k=7). Seven studies applied machine learning methods to analyse natural language (NLP). Four studies focused on therapeutic alliance, either in an unsupervised way (to reveal linguistic themes that underlie it),[90,91] or in a supervised way (to examine words/phrases that predict it).[92,93] Results here pointed to interesting linguistic themes of alliance rupturing (like markers denoting 'communication' and 'goal-setting' difficulties),[91] which in some cases were able to identify client-reported ruptures that were unidentified by therapists[93] (but see also section on reliability).

Three additional studies used NLP to identify premature departures in online therapy (achieving relatively high classification power),[94] defense mechanisms in therapy (achieving modest performance, which was no better than human coding),[95] and attachment styles (which were better classified using sex-specific models).[96]

Quality Assessment

Validity. Our validity assessment revealed four classes of models. First, models that scored highly only on empirical validity, because they were used purely in a data-driven way. These included three dynamical models,[42,44,45] nine Bayesian models,[56,57,59,62,64,67,68,72,97] and notably all reinforcement learning models[46,63,73–80] (patterns that we comment on in our Discussion).

A second category concerned the opposite pattern: primarily theory-driven models that scored highly only on theoretical validity (because they were only based on simulations). One dynamical system was identified in this category[52] and three Bayesian ones concerning Free Energy perspectives.[69–71]

A third category involved models that scored highly on empirical validity (because they were strongly data-driven) but also generative validity (because they were also assessed empirically vis-à-vis alternative models). This class included three Bayesian models,[55,58,66] which were all shown to be superior to reinforcement learning alternatives.

A final category concerned excellent models that scored highly on all types of validity, because they were created in a theory-driven manner, generated precise predictions (through computer simulations), and then examined for empirical predictions. These models included the dynamical systems by Liebovitch et al.[47,49] and Tschacher et al.[43,53], and five Bayesian models that were examined both in silico and in vitro.[54,60,61,65,98]

Reliability. Seven studies were purely theoretical (i.e., non-empirical) so reliability assessments were not applicable for them. One study was presented in a conference abstract, so a complete reliability assessment was not feasible.[10] Finally, 15 studies (27%), of which 13 were generative and 2 discriminative, did not report any performance metrics (Table 1).

Regarding generative models, 19 out of 32 empirical ones (59%) reported at least one performance metric. First, 5 out of these 19 studies (26%) reported only one performance metric, such as R^2 or a single information criterion, which may indicate high risk of bias.[44,55,59,60,66] Second, 8 out of these 19 studies (42%) reported only two performance metrics, with the most common ones being log-likelihoods and R^2, which are not sufficient on their own to evaluate model performance.[46,54,58,61,62,67,73,97] Finally, 6 out of 19 studies (32%) reported three or more metrics, indicating higher quality (particularly since some of these compared alternative models).[63,65,68,76,79,80] Nevertheless, no study reported task or test-retest reliability measures and only three studies performed parameter recoverability tests, suggesting that few computational investigations address reliability issues directly.[65,79,80]

Regarding machine learning (ML) studies, 14 out of 16 (87%) reported at least one performance metric. First, all 14 studies conducted internal validation, with the most common method being cross-validation with an 80-20% data split. Second, only one study conducted external validation to validate a psychosis prediction model.[82] Third, all studies reported at least one discrimination metric; however, 6 of them reported only one or two such metrics (focusing primarily on accuracy and F1 scores), suggesting potential risk of bias.[81,83,85,91–93] Finally, no studies performed calibration or examined the net benefit of their models, a trend that raises concerns with respect to the clinical utility of ML models.

Open Science. Only a few studies endorsed open science practices. More specifically, out of 48 empirical studies, 9 (19%) made their code publicly available,[57,58,61,66–68,75,76,79] 8 (17%) made their data (as well as code) publicly available,[57,58,61,66,68,75,76,79] and only 2 (4%) pre-registered their hypotheses.[57,76]

DISCUSSION

In this systematic review, we evaluated the utility of 55 computational studies in informing interpersonal dynamics of psychopathology. We found that the most frequently employed models were Bayesian, followed by machine learning ones and dynamical and reinforcement learning ones. These computational frameworks have been argued to be particularly well-suited for capturing intersubjective dynamics by formalizing (and thereby specifying) their nature. Indeed, the reviewed studies revealed notable interpersonal themes across various psychopathologies (themes that we summarize in Figure 1 and later on expand). Despite this progress, several issues remain regarding the reliability and validity of computational models, as well as the transparency with which those models are reported. In the following discussion, we expand on these matters, outline the strengths and limitations of existing research, and propose a roadmap for future work in this line of inquiry.

Issues facing interpersonal computational psychiatry

To begin with, we note some stark results from our quality assessment. First, the majority (81%) of empirical studies using generative modeling did not adequately report reliability metrics. Specifically, no study reported task reliability and only three studies conducted parameter recoverability tests. Previous commentaries have noted a tendency among experimentalists to overlook fundamental psychometric reporting.[99,100] Our findings here echo these concerns and align with previous research that has called for the routine reporting of (at least) task reliability and parameter recoverability metrics.[22,100] Such routine reporting is particularly important for the field of computational psychiatry, which has the tendency to modify tasks to suit well-defined hypotheses tests but in so doing distort their psychometric properties (even minor task modifications can dramatically affect established reliability metrics).[22,101] In that sense, routine reporting of task reliability, alongside sensitivity probing across task variabilities, is necessary.

Our second striking result is that similar reporting issues are present in ML studies. Although most (87%) ML studies reported at least one performance metric, none conducted calibration or reported more sophisticated and clinically relevant performance metrics (like net benefit). Likewise, only one ML study conducted external validation.[82] Of course, it can be argued that most ML studies did not perform comprehensive model evaluations because they aimed at a data-driven *exploration* of phenomena rather than the building of clinically useful models to *predict* those phenomena. However, this appears to be the case only for specific studies that were conducted in a predominantly unsupervised (and thus exploratory) way (for instance, the natural language processing ones). For other ML studies, it was unclear whether their focus was on clinical prediction modeling or causal inference (the two subtypes of supervised ML). This conceptual ambiguity is particularly concerning, given that some ML studies drew conclusions that were either inherently predictive (e.g., "*the autoEnsemble algorithm … could be implemented towards early identification of at-risk adolescents*")[84] or causal (e.g., "*when these factors are combined, the risk of depression in adolescents increases*").[87] In that sense, ML studies may predominantly suffer conceptual problems, implying that they could be improved through clearer framing of research questions and more sound interpretation of estimated effects.[102–104]

These conceptual problems lead us to the final issue at hand: open science practices. To elaborate, only 2 out of 52 empirical studies were pre-registered and few studies made their code (16%) and data (14%) publicly available. Alarmingly, none of the ML studies endorsed any open science practices, raising serious reproducibility concerns. It has been argued, for example, that many optimistic ML patterns (e.g., near perfect accuracy rates) or inconsistencies across their validation studies might be due to data pre-processing differences or data leakage problems (i.e., training data being leaked into testing data).[105,106] Crucially, these issues remain undetected unless researchers openly and transparently share their

analytical practices (i.e., open code). These standards are equally, if not more, relevant for generative models: Generative models are inherently idiosyncratic, as they posit *unique* data-generating mechanisms that are best understood by their innovators. Consequently, the only way these models can be tested by other research teams is if their dissemination is 'open enough' to allow for their precise reproduction (i.e., at least open code).

At present, however, few journals mandate open code practices, a limitation that may reduce researchers' perceived incentives for sharing their code publicly. We suggest that one of promoting research openness is through the publication of tutorial papers. Indeed, several tutorial papers already exist on computational methods that range from Bayesian modeling[107] to dynamical systems modeling[108] and Active Inference.[109] Publishing such papers, as well as endorsing related open science practices, has been found to be related to various researcher-level benefits, including 'increases in citations, media attention, potential collaborators, job opportunities and funding opportunities.'[41] In that sense, computational researchers could benefit by making a habit of open science practices, including writing tutorial papers so that *many labs* can test and validate their computational models.

A roadmap forward

The aforementioned problems have likely painted a pessimistic view of the field, to date. However, it must be acknowledged that several studies were of excellent quality, both in their reporting practices and the validity of their computational approaches. Below, we outline the main research themes from these studies, focusing on their key findings as well as their possible extensions by future studies (see Figure 1).

First, **dynamical systems (DS)** have been used to formalize ***human relationships*** as systems that evolve through time according to a set of rules (Figure 1). This conception of relationships can be traced back to the work of Gottman and colleagues, who were the first to use discrete equations to formally outline how two married partners emotionally influence each other, over time.[48] Since then, various research teams have extended Gottman's model or created their own models to formalize the relation between patients and therapists. [49–51] This research has demonstrated that it is possible to distil the multicoloured interactions between patients and therapists into *phase plots*, which depict the attractor states of the therapeutic field: that is, the relational states to which the patient-therapist dyad is drawn. Beyond human relationships, DS modeling has been used to summarize the physiological and vocal dynamics of interacting agents, showing that their coordination is crucial for successful interpersonal interactions.[42–45] Taken together, this field illustrates that it is possible to extract key dynamical properties of social transactions at multiple levels of analysis (linguistic, behavioural, biological, etc). At the same time, however, this field is currently limited empirically because it has rarely linked these dynamical properties to clinically relevant outcomes (e.g., symptoms or treatment outcomes). In that sense, a future goal of dynamical systems might be to specifically examine the above dynamics in a confirmatory way, not only to probe their clinical implications, but also to provide support for their data-generating implementations (and thereby enhance both their empirical and generative validity).

A second theme from our reviewed studies involves the use of **Bayesian** and **reinforcement learning (RL) models** to formalize ***mental representations*** of the self and others (as well as how these representations change over time, i.e., ***social learning***). This field of social representation modeling emerged from the confluence of several social disciplines (including behaviourism, social neuroscience, and neuroeconomics), which investigated how humans learn from rewards and punishments using mainly reinforcement learning models.[110] Our review revealed that interpersonal research has adopted these models to uncover notable learning themes in psychopathology, such as blunted social learning in depression (but not anxiety),[79] erroneous social learning in borderline personality disorder (e.g., updating representations of others based on self-representations),[10] and unstable (i.e., both increased and decreased) trust learning in adolescents with a history of interpersonal abuse.[74] The RL models of these studies are useful empirically, because they can be readily applied to experimental data to estimate how humans update single-point values (e.g., a self-esteem value) via associative learning.[111] At the same time, however, these models have been challenged theoretically (and score low on both theoretical and generative validities), because an increasing number of studies has questioned their data-generating mechanisms.[55,58,66]

Indeed, three of our reviewed studies have directly compared the performance of reinforcement learning models vis-à-vis Bayesian alternatives, showing that the latter provide a better fit to the data.[55,58,66] These results add to the generative validity of Bayesian modeling by implying that self-other mental representations are better understood as probability distributions that are updated via (approximate) Bayesian inference, not as point values that are updated via associative learning (see Figure 1).[9] Conceptualizing social representations in this way has enabled researchers to formalize (and better tap into) traditionally elusive social concepts, like *mentalizing* (by casting at as a form of deep reasoning: 'I think that you think that I think' and so on),[60] *irritability* (by casting it as a disregard of prior beliefs about someone else and reduced mentalizing with them),[61] and *moral strategizing* (by casting it as the type of utility that an agent tries to maximize in a social interaction).[63–65] In turn, these formal concepts have been linked to psychiatric symptomatologies, shedding light on their causal origins. For example, while psychological research has identified mentalizing impairments in both individuals with autism and those with borderline personality disorder, computational research has shown that the latter may only face those problems when they are feeling 'irritated'.[61] Findings like these illustrate that Bayesian modeling could disentangle the causal origins of similar symptoms in different psychiatric conditions.[110] However, such findings remain limited empirically because their data-generating mechanisms have only been investigated cross-sectionally, to date. It is thus crucial for contemporary Bayesian models to be extended over time in order to explore how their dynamics evolve either in response to life stressors[112] or therapeutic interventions.[113]

This latter point on intervention brings us to the final and perhaps most clinically relevant research theme of our reviewed studies: the use of **natural language processing (NLP)** to reveal ***linguistic themes*** in therapeutic relationships. NLP techniques improve upon traditional and labour-intensive qualitative methods by identifying statistical regularities in word co-occurrence patterns and distilling them into themes (aka sentiments) in unsupervised or (semi-)supervised ways.[113] Research has used such NLP methodologies to distil therapeutic alliance linguistically, identifying interesting themes that might underpin its rupturing (for instance, communication difficulties).[91] Notably, in some cases, NLP methods were able to leverage clients' speech to identify alliance ruptures that were not picked up by therapists.[93] However, despite these promising advances, it must be noted that most studies attempting to classify other psychological variables, like defense mechanisms[95] or attachment styles[96], from patients' speech have not exceeded human-coding reliability. Thus, while encouraging, these NLP findings will need to be replicated with larger sample sizes, other languages, and varied NLP models (including the more recent large language models) in order to be ultimately considered clinically useful.[113,114]

Conclusion

To conclude, our systematic review has identified notable computational themes relating to interpersonal dynamics in psychopathology. Although many studies in this area of inquiry exhibited conceptual and pragmatic problems, several were conducted with sufficient rigour to offer a promising path forward. These studies, in particular, illustrated that computational models can be leveraged to formalize how humans re-present themselves and others, how they interpersonally transact with others, and how they communicate with them. Together, these patterns suggest that computational models hold the potential to offer a more precise understanding of historically intractable intersubjective dynamics. However, for this potential to be realized, computational researchers must adopt transparent, comprehensive, and open reporting practices.

References

1. Heider F. Social perception and phenomenal causality. *Psychological Review*. 1944;51(6):358-374. doi:10.1037/h0055425

2. Fonagy P. Thinking about thinking: Some clinical and theoretical considerations in the treatment of a borderline patient. *International Journal of psychoanalysis*. 1991;72(4):639-656.

3. Nowak MA, Sigmund K. Tit for tat in heterogeneous populations. *Nature*. 1992;355(6357):250-253.

4. Horowitz LM, Strack S. *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions*. John Wiley & Sons; 2010.

5. Wright AG, Pincus AL, Hopwood CJ. Contemporary integrative interpersonal theory: Integrating structure, dynamics, temporal scale, and levels of analysis. *Journal of Psychopathology and Clinical Science*. 2023;132(3):263.

6. Wright AG, Ringwald WR, Hopwood CJ, Pincus AL. It's time to replace the personality disorders with the interpersonal disorders. *American Psychologist*. 2022;77(9):1085.

7. Hames JL, Hagan CR, Joiner TE. Interpersonal processes in depression. *Annual review of clinical psychology*. 2013;9:355-377.

8. Hartmann A, Zeeck A, Barrett MS. Interpersonal problems in eating disorders. *International journal of eating disorders*. 2010;43(7):619-627.

9. Low AAY, Hopper WJT, Angelescu I, Mason L, Will GJ, Moutoussis M. Self-esteem depends on beliefs about the rate of change of social approval. *Sci Rep*. 2022;12(1):6643. doi:10.1038/s41598-022-10260-6

10. Story G, Ereira S, Valle S, Chamberlain S, Grant J, Dolan R. 367. A Computational Signature of Self-Other Mergence in Borderline Personality Disorder. *Biological Psychiatry*. 2023;93(9):S242. doi:10.1016/j.biopsych.2023.02.607

11. Friston K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci*. 2010;11(2):127-138. doi:10.1038/nrn2787

12. Durstewitz D, Huys QJM, Koppe G. Psychiatric Illnesses as Disorders of Network Dynamics. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2021;6(9):865-876. doi:10.1016/j.bpsc.2020.01.001

13. Adams RA, Huys QJM, Roiser JP. Computational Psychiatry: towards a mathematically informed understanding of mental illness. *J Neurol Neurosurg Psychiatry*. Published online July 8, 2015:jnnp-2015-310737. doi:10.1136/jnnp-2015-310737

14. Bennett D, Silverstein SM, Niv Y. The two cultures of computational psychiatry. *JAMA psychiatry*. 2019;76(6):563-564.

15. Friston KJ, Stephan KE, Montague R, Dolan RJ. Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*. 2014;1(2):148-158. doi:10.1016/S2215-0366(14)70275-5

16. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. *Trends in Cognitive Sciences*. 2012;16(1):72-80. doi:10.1016/j.tics.2011.11.018

17. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci*. 2016;19(3):404-413. doi:10.1038/nn.4238

18. Huys QJM, Browning M, Paulus MP, Frank MJ. Advances in the computational understanding of mental illness. *Neuropsychopharmacol*. 2021;46(1):3-19. doi:10.1038/s41386-020-0746-4

19. Zavlis O. Computational approaches to mental illnesses. *Nature Reviews Psychology*. Published online 2024:1-1.

20. Moutoussis M, Fearon P, El-Deredy W, Dolan RJ, Friston KJ. Bayesian inferences about the self (and others): a review. *Conscious Cogn*. 2014;25(9303140):67-76. doi:10.1016/j.concog.2014.01.009

21. Lockwood PL, Klein-Flugge MC. Computational modelling of social cognition and behaviour-a reinforcement learning primer. *Soc cogn affect neurosci*. 2021;16(8):761-771. doi:10.1093/scan/nsaa040

22. Karvelis P, Paulus MP, Diaconescu AO. Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*. 2023;148:105137. doi:10.1016/j.neubiorev.2023.105137

23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*. 2021;88:105906. doi:10.1016/j.ijsu.2021.105906

24. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev*. 2017;6(1):245. doi:10.1186/s13643-017-0644-y

25. Mayo-Wilson E, Li T, Fusco N, et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *Journal of clinical epidemiology*. 2017;91:95-110.

26. Nunes A, Singh S, Allman J, et al. A critical evaluation of dynamical systems models of bipolar disorder. *Transl Psychiatry*. 2022;12(1):416. doi:10.1038/s41398-022-02194-4

27. Palminteri S, Wyart V, Koechlin E. The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*. 2017;21(6):425-433.

28. Rigoli F. Prisoner of the present: Borderline personality and a tendency to overweight cues during Bayesian inference. *Pers disord*. 2022;13(6):609-618. doi:10.1037/per0000549

29. Zavlis O, Bentall R, Fonagy P, Anonymous. *Mood Instability: A Reference-Dependent Computational Account*. Open Science Framework; 2024. doi:10.31219/osf.io/rgv5m

30. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation–A discussion and demonstration of basic features. *PloS one*. 2019;14(7):e0219854.

31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*. 2016;15(2):155-163.

32. Mkrtchian A, Valton V, Roiser JP. Reliability of Decision-Making and Reinforcement Learning Computational Parameters. *Computational Psychiatry*. 2023;7(1):30. doi:10.5334/cpsy.86

33. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698

34. Riley RD, Windt D van der, Croft P, eds. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. First edition. Oxford University Press; 2019.

35. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer; 2009.

36. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology*. 2016;69:245-247.

37. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in medicine*. 2000;19(4):453-473.

38. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of internal medicine*. 1999;130(6):515-524.

39. Seyedsalehi A, Lennox B. Predictive tools in psychosis: what is 'good enough'? *Nat Rev Neurol*. Published online March 6, 2023. doi:10.1038/s41582-023-00787-1

40. Vicente-Saez R, Martinez-Fuentes C. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*. 2018;88:428-436. doi:10.1016/j.jbusres.2017.12.043

41. McKiernan EC, Bourne PE, Brown CT, et al. How open science helps researchers succeed. *eLife*. 2016;5:e16800. doi:10.7554/eLife.16800

42. Varlet M, Marin L, Raffard S, et al. Impairments of social motor coordination in schizophrenia. *PLoS ONE*. 2012;7(1):e29772. doi:10.1371/journal.pone.0029772

43. Tschacher W, Haken H. Causation and chance: Detection of deterministic and stochastic ingredients in psychotherapy processes. *Psychotherapy Research*. 2020;30(8):1075-1087. doi:10.1080/10503307.2019.1685139

44. Schreiber AM, Wright AGC, Beeney JE, 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*. 2020;129(5):433-444. doi:10.1037/abn0000526

45. Paz A, Rafaeli E, Bar-Kalifa E, et al. Intrapersonal and interpersonal vocal affect dynamics during psychotherapy. *Journal of Consulting and Clinical Psychology*. 2021;89(3):227-239. doi:10.1037/ccp0000623

46. Will GJ, Moutoussis M, Womack PM, et al. Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Transl Psychiatry*. 2020;10(1):96. doi:10.1038/s41398-020-0702-4

47. Liebovitch LS, Peluso PR, Norman MD, Su J, Gottman JM. Mathematical model of the dynamics of psychotherapy. *Cogn Neurodyn*. 2011;5(3):265-275. doi:10.1007/s11571-011-9157-x

48. Gottman JM, Murray JD, Swanson CC, Tyson R, Swanson KR. *The Mathematics of Marriage: Dynamic Nonlinear Models*. 1st MIT Press paperback edition. The MIT Press; 2005.

49. Peluso PR, Liebovitch LS, Gottman JM, Norman MD, Su J. A mathematical model of psychotherapy: An investigation using dynamic non-linear equations to model the therapeutic relationship. *Psychotherapy Research*. 2012;22(1):40-55. doi:10.1080/10503307.2011.622314

50. Baker AZ, Peluso PR, 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. *Psychotherapy Research*. 2022;32(2):223-237. doi:10.1080/10503307.2021.1921303

51. Diaz P, Peluso PR, Freund R, Baker AZ, 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*. 2023;14(101545006):980739. doi:10.3389/fpsyt.2023.980739

52. Schiepek G, Aas B, Viol K. The Mathematics of Psychotherapy: A Nonlinear Model of Change Dynamics. *Nonlinear Dynamics Psychol Life Sci*. 2016;20(3):369-399.

53. Tschacher W, Haken H, Kyselo M. Alliance: a common factor of psychotherapy modeled by structural theory. *Front Psychol*. 2015;6. doi:10.3389/fpsyg.2015.00421

54. Yoshida W, Dziobek I, Kliemann D, Heekeren HR, Friston KJ, Dolan RJ. Cooperation and heterogeneity of the autistic mind. *J Neurosci*. 2010;30(26):8815-8818. doi:10.1523/JNEUROSCI.0400-10.2010

55. Forgeot d'Arc B, Devaine M, Daunizeau J. Social behavioural adaptation in Autism. *PLoS Comput Biol*. 2020;16(3):e1007700. doi:10.1371/journal.pcbi.1007700

56. Thomas L, Lockwood PL, Garvert MM, Balsters JH. Contagion of Temporal Discounting Value Preferences in Neurotypical and Autistic Adults. *J Autism Dev Disord*. 2022;52(2):700-713. doi:10.1007/s10803-021-04962-5

57. Barnby JM, Bell V, Mehta MA, 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. Marinazzo D, ed. *PLoS Comput Biol*. 2020;16(10):e1008372. doi:10.1371/journal.pcbi.1008372

58. Barnby JM, Raihani N, Dayan P. Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*. 2022;225(0367541, dmh):105098. doi:10.1016/j.cognition.2022.105098

59. Na S, Blackmore S, Chung D, et al. Computational mechanisms underlying illusion of control in delusional individuals. *Schizophrenia Research*. 2022;245:50-58. doi:10.1016/j.schres.2022.01.054

60. Xiang T, Ray D, Lohrenz T, Dayan P, Montague PR. Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Comput Biol*. 2012;8(12):e1002841. doi:10.1371/journal.pcbi.1002841

61. Hula A, Vilares I, Lohrenz T, Dayan P, Montague PR. A model of risk and mental state shifts during social interaction. *PLoS Comput Biol*. 2018;14(2):e1005935. doi:10.1371/journal.pcbi.1005935

62. Xiao F, Zhao J, Fan L, et al. Understanding guilt-related interpersonal dysfunction in obsessive-compulsive personality disorder through computational modeling of two social interaction tasks. *Psychol Med*. 2023;53(12):5569-5581. doi:10.1017/S003329172200277X

63. Story GW, Smith R, Moutoussis M, et al. A social inference model of idealization and devaluation. *Psychol Rev*. 2023;(0376476, qfb). doi:10.1037/rev0000430

64. Siegel JZ, Curwell-Parry O, Pearce S, Saunders KEA, Crockett MJ. A Computational Phenotype of Disrupted Moral Inference in Borderline Personality Disorder. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2020;5(12):1134-1141. doi:10.1016/j.bpsc.2020.07.013

65. Driessen JMA, van Baar JM, Sanfey AG, Glennon JC, Brazil IA. Moral strategies and psychopathic traits. *J Abnorm Psychol*. 2021;130(5):550-561. doi:10.1037/abn0000675

66. Lamba A, Frank MJ, FeldmanHall O. Anxiety Impedes Adaptive Social Learning Under Uncertainty. *Psychol Sci*. 2020;31(5):592-603. doi:10.1177/0956797620910993

67. Hale WW, Aarts E. Hidden Markov model detection of interpersonal interaction dynamics in predicting patient depression improvement in psychotherapy: Proof-of-concept study. *Journal of Affective Disorders Reports*. 2023;14:100635. doi:10.1016/j.jadr.2023.100635

68. Henco L, Diaconescu AO, Lahnakoski JM, et al. Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. Sterzer P, ed. *PLoS Comput Biol*. 2020;16(9):e1008162. doi:10.1371/journal.pcbi.1008162

69. Prosser A, Friston KJ, Bakker N, Parr T. A Bayesian Account of Psychopathy: A Model of Lacks Remorse and Self-Aggrandizing. *Computational Psychiatry*. 2018;2(0):92. doi:10.1162/CPSY_a_00016

70. Cittern D, Nolte T, Friston K, Edalat A. Intrinsic and extrinsic motivators of attachment under active inference. Maloney LT, ed. *PLoS ONE*. 2018;13(4):e0193955. doi:10.1371/journal.pone.0193955

71. Constant A, Hesp C, Davey CG, Friston KJ, Badcock PB. Why Depressed Mood is Adaptive: A Numerical Proof of Principle for an Evolutionary Systems Theory of Depression. *Comput Psychiatr*. 2021;5(1):60-80. doi:10.5334/cpsy.70

72. Saint-Georges C, Mahdhaoui A, Chetouani M, et al. Do parents recognize autistic deviant behavior long before diagnosis? Taking into account interaction using computational methods. *PLoS ONE*. 2011;6(7):e22393. doi:10.1371/journal.pone.0022393

73. Will GJ, Rutledge RB, Moutoussis M, Dolan RJ. Neural and computational processes underlying dynamic changes in self-esteem. *eLife*. 2017;6:e28098. doi:10.7554/eLife.28098

74. Lenow J, Cisler J, Bush K. Altered Trust Learning Mechanisms Among Female Adolescent Victims of Interpersonal Violence. *J Interpers Violence*. 2015;(8700910).

75. Contreras-Huerta LS, Lockwood PL, Bird G, Apps MAJ, Crockett MJ. Prosocial behavior is associated with transdiagnostic markers of affective sensitivity in multiple domains. *Emotion*. 2022;22(5):820-835. doi:10.1037/emo0000813

76. Barnby JM, Mehta MA, Moutoussis M. The computational relationship between reinforcement learning, social inference, and paranoia. Gershman SJ, ed. *PLoS Comput Biol*. 2022;18(7):e1010326. doi:10.1371/journal.pcbi.1010326

77. Fineberg SK, Leavitt J, Stahl DS, et al. Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. *Biol Psychiatry*. 2018;84(11):838-845. doi:10.1016/j.biopsych.2018.05.020

78. Shapiro-Thompson R, Shah TV, Yi C, Jackson N, Trujillo Diaz D, Fineberg SK. Modulation of Trust in Borderline Personality Disorder by Script-Based Imaginal Exposure to Betrayal. *J Personal Disord*. 2023;37(5):508-524. doi:10.1521/pedi.2023.37.5.508

79. Safra L, Chevallier C, Palminteri S. Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Comput Biol*. 2019;15(7):e1007224. doi:10.1371/journal.pcbi.1007224

80. Frey AL, Frank MJ, McCabe C. Social reinforcement learning as a predictor of real-life experiences in individuals with high and low depressive symptomatology. *Psychol Med*. 2021;51(3):408-415. doi:10.1017/S0033291719003222

81. Kang B, Lee S, Oh A, Kang S, Hwang I, Song J. Towards Understanding Relational Orientation: Attachment Theory and Facebook Activities. In: *Proceedings of the 18th*

ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM; 2015:1404-1415. doi:10.1145/2675133.2675211

82. Antonucci LA, Raio A, Pergola G, 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*. 2021;9(1):47. doi:10.1186/s40359-021-00552-3

83. Doborjeh Z, Doborjeh M, Sumich A, et al. Investigation of social and cognitive predictors in non-transition ultra-high-risk' individuals for psychosis using spiking neural networks. *Schizophrenia (Heidelb)*. 2023;9(1):10. doi:10.1038/s41537-023-00335-2

84. Haghish EF, Nes RB, Obaidi M, et al. Unveiling Adolescent Suicidality: Holistic Analysis of Protective and Risk Factors Using Multiple Machine Learning Algorithms. *J Youth Adolescence*. 2024;53(3):507-525. doi:10.1007/s10964-023-01892-6

85. Handing EP, 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. *The Lancet Regional Health - Europe*. 2022;18:100391. doi:10.1016/j.lanepe.2022.100391

86. Hu Y, Fei J, Zheng C, et al. Parental Relationships Surpass Friendships as Predictors of Long-Term Mental Functioning: A Multilevel Analysis. *Int J Ment Health Addiction*. Published online June 19, 2023. doi:10.1007/s11469-023-01092-0

87. Wang C, Zhou T, Fu L, Xie D, Qi H, Huang Z. Risk and Protective Factors of Depression in Family and School Domains for Chinese Early Adolescents: An Association Rule Mining Approach. *Behavioral Sciences*. 2023;13(11):893. doi:10.3390/bs13110893

88. Schorr MT, Quadors Dos Santos BTM, Feiten JG, et al. Association between childhood trauma, parental bonding and antisocial personality disorder in adulthood: A machine learning approach. *Psychiatry Research*. 2021;304:114082. doi:10.1016/j.psychres.2021.114082

89. Lu H, Xie C, Lian P, Yu C, Xie Y. Psychosocial Factors Predict the Level of Aggression of People with Drug Addiction: A Machine Learning Approach. *Psychology, Health & Medicine*. 2022;27(5):1168-1175. doi:10.1080/13548506.2021.1910321

90. Martinez VR, Flemotomos N, Ardulov V, et al. Identifying Therapist and Client Personae for Therapeutic Alliance Estimation. In: *Interspeech 2019*. ISCA; 2019:1901-1905. doi:10.21437/Interspeech.2019-2829

91. Atzil-Slonim D, Juravski D, Bar-Kalifa E, et al. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy*. 2021;58(2):324-339. doi:10.1037/pst0000362

92. Goldberg SB, Flemotomos N, Martinez VR, et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*. 2020;67(4):438-448. doi:10.1037/cou0000382

93. Tsakalidis A, Atzil-Slonim D, Polakovski A, Shapira N, Tuval-Mashiach R, Liakata M. Automatic Identification of Ruptures in Transcribed Psychotherapy Sessions. In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics; 2021:122-128. doi:10.18653/v1/2021.clpsych-1.15

94. Xu Y, Chan CS, Tsang C, et al. Detecting premature departure in online text-based counseling using logic-based pattern matching. *Internet Interventions*. 2021;26:100486. doi:10.1016/j.invent.2021.100486

95. Tasca AN, Carlucci S, Wiley JC, Holden M, El-Roby A, Tasca GA. Detecting defense mechanisms from Adult Attachment Interview (AAI) transcripts using machine learning. *Psychotherapy Research*. 2023;33(6):757-767. doi:10.1080/10503307.2022.2156306

96. Gómez-Zaragozá L, Marín-Morales J, Vargas EP, Giglioli IAC, Raya MA. An Online Attachment Style Recognition System Based on Voice and Machine Learning. *IEEE J Biomed Health Inform*. 2023;27(11):5576-5587. doi:10.1109/JBHI.2023.3304369

97. Barnby JM, Deeley Q, Robinson O, Raihani N, Bell V, Mehta MA. Paranoia, sensitization and social inference: findings from two large-scale, multi-round behavioural experiments. *R Soc open sci*. 2020;7(3):191525. doi:10.1098/rsos.191525

98. Story GW, Smith R, Moutoussis M, et al. *A Social Inference Model of Idealization and Devaluation*. PsyArXiv; 2021. doi:10.31234/osf.io/yvu2b

99. Vasey MW, Dalgleish T, Silverman WK. Research on information-processing factors in child and adolescent psychopathology: A critical commentary. *Journal of Clinical Child and Adolescent Psychology*. 2003;32(1):81-93.

100. Parsons S, Kruijt AW, Fox E. Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*. 2019;2(4):378-395. doi:10.1177/2515245919879695

101. Zorowitz S, Niv Y. Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Published online 2023.

102. Andaur Navarro CL, Damen JAA, Van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*. 2023;154:8-22. doi:10.1016/j.jclinepi.2022.11.015

103. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. Published online October 20, 2021:n2281. doi:10.1136/bmj.n2281

104. Andaur Navarro CL, Damen JAA, Takada T, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*. 2023;158:99-110. doi:10.1016/j.jclinepi.2023.03.024

105. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2012;6(4):1-21.

106. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Translational psychiatry*. 2019;9(1):271.

107. Gershman SJ, Blei DM. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*. 2012;56(1):1-12. doi:10.1016/j.jmp.2011.08.004

108. van der Maas HLJ. *Complex-Systems Research in Psychology*. SFI Press; 2024.

109. Smith R, Friston KJ, Whyte CJ. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*. 2022;107:102632. doi:10.1016/j.jmp.2021.102632

110. Barnby JM, Dayan P, Bell V. Formalising social representation to explain psychiatric symptoms. *Trends in cognitive sciences*. 2023;27(3):317-332.

111. Dayan P, Niv Y. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*. 2008;18(2):185-196.

112. Hitchcock PF, Fried EI, Frank MJ. Computational psychiatry needs time and context. *Annual review of psychology*. 2022;73(1):243-270.

113. Nour MM, Huys QJ. Natural Language Processing in Psychiatry: A Field at an Inflection Point. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2023;8(10):979-981.

114. Nour MM. Adequate methodological reporting and sensitivity analyses are essential to allow reproducibility and interpretability in NLP studies. *Journal of affective disorders*. 2024;356:436-437.