

A systematic review of computational modeling of interpersonal dynamics in psychopathology

In the format provided by the
authors and unedited

Contents

Supplement I: Search Strategies	1
EMBASE	1
MEDLINE	2
PsycInfo	3
Web of Science	4
Google Scholar	4
Supplement II. Study Details	5
Table S1. Key characteristics of included studies.....	6
Table S2. Study Design.....	10
Supplement III. Risk of Bias Assessment.....	16
Table S3. NIH Quality assessment of dynamical systems studies.	17
Table S4. NIH Quality assessment of reinforcement learning studies.	18
Table S5. NIH Quality assessment of Bayesian studies.....	19
Table S6. PROBAST assessment of supervised machine learning studies.	20
Table S7. PROBAST-modified assessment of unsupervised machine learning studies.	21
Supplement IV. Validity Assessment	22
Table S8. Validity assessment for dynamical systems.....	23
Table S9. Validity assessment for reinforcement learning models.	24
Table S10. Validity assessment for Bayesian models.....	25
References.....	26

Supplement I: Search Strategies

EMBASE

Search Strategy:

- 1 interpersonal psychotherapy/ or interpersonal communication/ or interpersonal.mp. (260062)
- 2 human relation/ or relational.mp. (125026)
- 3 intrapsychic.mp. or psychoanalysis/ (36358)
- 4 ego/ or ego psychology/ or ego identity/ or ego.mp. (17209)
- 5 intrapersonal.mp. (4323)
- 6 mentalization/ or mentali*.mp. (8350)
- 7 theory of mind.mp. or "theory of mind"/ (9879)
- 8 attachment.mp. or emotional attachment/ or clinical attachment level/ (156187)
- 9 object relation.mp. or object relation/ (12892)
- 10 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 (573035)
- 11 computer simulation/ or computational.mp. or computer model/ (487975)
- 12 generative model*.mp. (3991)
- 13 mathematical model/ or mathematical model*.mp. (190052)
- 14 computational psychiatry.mp. (705)
- 15 dynamical system.mp. or nonlinear system/ (36223)
- 16 Bayes theorem/ or Bayes*.mp. (114875)
- 17 reinforcement learning.mp. (9742)
- 18 Active Inference.mp. (554)
- 19 Free Energy.mp. (48619)
- 20 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 (825973)
- 21 psychopathology.mp. or mental disease/ (348854)
- 22 mental illness.mp. (55035)
- 23 mental disorder.mp. (18524)
- 24 paranoid schizophrenia/ or schizophrenia.mp. or schizophrenia spectrum disorder/ or schizophrenia assessment/ or simple schizophrenia/ or schizophrenia/ or residual schizophrenia/ or treatment-resistant schizophrenia/ or latent schizophrenia/ or catatonic schizophrenia/ (252434)
- 25 childhood psychosis/ or psychosis.mp. or puerperal psychosis/ or depressive psychosis/ or paranoid psychosis/ or treatment-resistant psychosis/ or affective psychosis/ or menstrual psychosis/ or endogenous psychosis/ or psychosis/ or intensive care psychosis/ or acute psychosis/ or drug induced psychosis/ or manic psychosis/ or schizoaffective psychosis/ or Korsakoff psychosis/ or alcohol psychosis/ or cocaine-induced psychosis/ or methamphetamine-induced psychosis/ (157544)
- 26 bipolar*.mp. or bipolar disorder/ or bipolar depression/ or bipolar I disorder/ (156184)
- 27 mania.mp. or "mixed mania and depression"/ or bipolar mania/ or mania/ (32909)
- 28 mood disorder.mp. or mood disorder/ (64858)
- 29 depression/ or depressi*.mp. (991828)
- 30 anxiety/ or anxi*.mp. (612580)
- 31 phobia/ or phobia.mp. (33978)
- 32 personality disorder.mp. or personality disorder/ (61679)
- 33 relational disorder.mp. (23)
- 34 substance abuse/ or substance disorder.mp. (61729)
- 35 heroin dependence/ or drug dependence/ or opiate addiction/ or addict*.mp. or addiction/ (248141)
- 36 dissociative disorder.mp. or dissociative disorder/ (4931)
- 37 eating disorder.mp. or eating disorder/ (47990)
- 38 sleep disorder.mp. or sleep disorder/ (109923)
- 39 impulse control disorder/ or impulse disorder.mp. (4807)
- 40 adjustment disorder.mp. or adjustment disorder/ (5896)
- 41 autism/ or autis*.mp. (119946)
- 42 attention deficit.mp. or attention deficit hyperactivity disorder/ (94644)
- 43 hyperactivity.mp. or hyperactivity/ (131944)
- 44 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 (2288087)
- 45 10 and 20 and 44 (546)

MEDLINE

Search Strategy:

- 1 Interpersonal Relations/ or interpersonal.mp. (125629)
- 2 relational.mp. (26451)
- 3 Psychoanalytic Therapy/ or intrapsychic.mp. (16655)
- 4 ego.mp. or Ego/ (19594)
- 5 intrapersonal.mp. (3918)
- 6 Mentalization/ or mentali*.mp. (6464)
- 7 theory of mind.mp. or "Theory of Mind"/ (8023)
- 8 attachment.mp. or Object Attachment/ (146025)
- 9 object relation.mp. (135)
- 10 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 (331800)
- 11 Computer Simulation/ or computational.mp. (517405)
- 12 generative model*.mp. (3750)
- 13 mathematical model*.mp. (70435)
- 14 computational psychiatry.mp. (505)
- 15 Nonlinear Dynamics/ or dynamical system.mp. (24418)
- 16 Bayes Theorem/ or Bayes*.mp. (99081)
- 17 reinforcement learning.mp. (8523)
- 18 Active Inference.mp. (642)
- 19 Free Energy.mp. (50976)
- 20 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 (727918)
- 21 psychopathology.mp. or Psychopathology/ (46800)
- 22 mental illness.mp. (41679)
- 23 mental disorder.mp. or Mental Disorders/ (195177)
- 24 "Schizophrenia Spectrum and Other Psychotic Disorders"/ or Schizophrenia, Paranoid/ or schizophrenia.mp. or Schizophrenia, Disorganized/ or Schizophrenia, Treatment-Resistant/ or Schizophrenia, Childhood/ or Schizophrenia/ or Schizophrenia, Catatonic/ (170931)
- 25 psychosis.mp. or Psychotic Disorders/ (82683)
- 26 Bipolar Disorder/ or bipolar*.mp. (100126)
- 27 mania.mp. or Mania/ (12933)
- 28 mood disorder.mp. or Mood Disorders/ (22140)
- 29 Depressive Disorder/ or Depression/ or depressi*.mp. or Depressive Disorder, Major/ (629682)
- 30 Anxiety/ or Anxiety Disorders/ or anxi*.mp. (374120)
- 31 phobia.mp. or Phobic Disorders/ (17866)
- 32 personality disorder.mp. or Personality Disorders/ (49418)
- 33 relational disorder.mp. (12)
- 34 Substance-Related Disorders/ or substance disorder.mp. (110917)
- 35 Heroin Dependence/ or Narcotics/ or Opioid-Related Disorders/ or Behavior, Addictive/ or addict*.mp. or Alcoholism/ (205525)
- 36 dissociative disorder.mp. or Dissociative Disorders/ (4583)
- 37 eating disorder.mp. or "Feeding and Eating Disorders"/ (29960)
- 38 sleep disorder.mp. or Sleep Wake Disorders/ (34553)
- 39 impulse control disorder.mp. or "Disruptive, Impulse Control, and Conduct Disorders"/ (3146)
- 40 adjustment disorder.mp. or Adjustment Disorders/ (5342)
- 41 Autism Spectrum Disorder/ or autis*.mp. or Autistic Disorder/ (81637)
- 42 Attention Deficit Disorder with Hyperactivity/ or attention deficit.mp. (53854)
- 43 hyperactivity.mp. (73165)
- 44 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 (1656127)
- 45 10 and 20 and 44 (327)

PsycInfo

Search Strategy:

- 1 exp Interpersonal Interaction/ or interpersonal.mp. or exp Interpersonal Compatibility/ or exp Interpersonal Psychotherapy/ or exp Interpersonal Communication/ or exp Interpersonal Relationships/ or exp Interpersonal Control/ or exp Interpersonal Attraction/ or exp Interpersonal Influences/ (1052888)
- 2 exp Relational Aggression/ or relational.mp. or exp Brief Relational Therapy/ (73483)
- 3 exp Intersubjectivity/ or exp Psychoanalytic Theory/ or exp Psychoanalysis/ or intrapsychic.mp. or exp Psychodynamics/ (104377)
- 4 ego.mp. or exp Ego/ or exp Ego Identity/ (33016)
- 5 intrapersonal.mp. (7459)
- 6 exp Mentalization/ or mentalization.mp. (3754)
- 7 attachment.mp. or exp Attachment Disorders/ or exp Attachment Behavior/ or exp Attachment Theory/ or exp Attachment Style/ (67973)
- 8 object relations.mp. or exp Object Relations/ (11212)
- 9 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 (1218931)
- 10 computational.mp. or exp Computational Modeling/ or exp Simulation/ (118385)
- 11 generative model*.mp. (913)
- 12 exp Mathematical Modeling/ (44188)
- 13 computational psychiatry.mp. (368)
- 14 dynamical system.mp. (604)
- 15 exp Bayesian Analysis/ or Bayes*.mp. (19255)
- 16 exp Computational Reinforcement Learning/ or reinforcement learning.mp. (4249)
- 17 Active Inference.mp. (416)
- 18 Free Energy.mp. (480)
- 19 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 (136427)
- 20 exp Child Psychopathology/ or psychopathology.mp. or exp Adolescent Psychopathology/ or exp Psychopathology/ (83292)
- 21 mental illness.mp. (55774)
- 22 mental disorder.mp. or exp Mental Disorders/ (1116895)
- 23 exp "Schizophrenia (Disorganized Type)"/ or exp "Fragmentation (Schizophrenia)"/ or exp Undifferentiated Schizophrenia/ or schizophrenia.mp. or exp Childhood Onset Schizophrenia/ or exp Process Schizophrenia/ or exp Paranoid Schizophrenia/ or exp Catatonic Schizophrenia/ or exp Acute Schizophrenia/ or exp Schizophrenia/ (153064)
- 24 exp Reactive Psychosis/ or exp Paranoid Psychosis/ or exp Chronic Psychosis/ or psychosis.mp. or exp Postpartum Psychosis/ or exp Childhood Onset Psychosis/ or exp Affective Psychosis/ or exp Psychosis/ (153862)
- 25 bipolar*.mp. or exp Bipolar Disorder/ or exp Bipolar II Disorder/ or exp Bipolar I Disorder/ (59215)
- 26 mania.mp. or exp Mania/ (19216)
- 27 mood disorder.mp. or exp Affective Disorders/ (197569)
- 28 exp Major Depression/ or depress*.mp. (470814)
- 29 exp Anxiety Disorders/ or exp Illness Anxiety Disorder/ or exp Anxiety/ or anxiety.mp. or exp Generalized Anxiety Disorder/ (325541)
- 30 phobia.mp. or exp Phobias/ (21366)
- 31 personality disorder.mp. or exp Personality Disorders/ (50795)
- 32 personality disorder.mp. or exp Personality Disorders/ (50795)
- 33 relational disorder.mp. (28)
- 34 exp Drug Abuse/ or exp "Substance Use Disorder"/ or exp Alcoholism/ or substance disorder.mp. or exp Addiction/ (154972)
- 35 exp Alcoholism/ or exp Drug Abuse/ or exp Heroin/ or exp "Heroin Use Disorder"/ or addict*.mp. or exp Opiates/ or exp Addiction/ or exp Drug Addiction/ or exp Addiction Treatment/ or exp Cocaine/ (192682)
- 36 dissociative disorder.mp. or exp Dissociative Disorders/ (6263)
- 37 eating disorder.mp. or exp Eating Disorders/ (42544)
- 38 sleep disorder.mp. or exp Sleep Wake Disorders/ (27884)
- 39 exp Impulse Control Disorders/ or impulse control disorder.mp. (1678)
- 40 adjustment disorder.mp. or exp Adjustment Disorders/ (1871)
- 41 exp Autism Spectrum Disorders/ or autism*.mp. (77746)
- 42 exp Attention Deficit Disorder/ or exp Attention Deficit Disorder with Hyperactivity/ or attention deficit.mp. (48437)
- 43 exp Hyperactivity/ or hyperactivity.mp. (56869)
- 44 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 (1526975)
- 45 9 and 19 and 44 (2343)

Web of Science

((("interpersonal") OR ("relational") OR ("intrapsychic") OR ("psychoanaly*") OR ("ego") OR ("intrapersonal") OR ("mentali*") OR ("theory of mind") OR ("attachment") OR ("object relation"))) AND (((("computational") OR ("generative model*") OR ("mathematical model*") OR ("computational psychiatry") OR ("dynamical system") OR ("Bayes*") OR ("reinforcement learning") OR ("Active Inference") OR ("Free Energy") OR ("Artificial Intelligence") OR ("Machine Learning")) AND (("psychopathology") OR ("mental illness") OR ("mental disorder") OR ("schizo*") OR ("psychosis") OR ("psychotic") OR ("bipolar*") OR ("mania") OR ("manic") OR ("mood disorder") OR ("depressi*") OR ("anxi*") OR ("phobia") OR ("personality disorder") OR ("relational disorder") OR ("substance disorder") OR ("addict*") OR ("dissociative disorder") OR ("eating disorder") OR ("sleep disorder") OR ("impulse control disorder") OR ("adjustment disorder") OR ("autis*") OR ("attention deficit") OR ("hyperactivity"))))

Google Scholar

For Google Scholar, we conducted three independent searches: ‘Interpersonal Computational Psychiatry’, ‘Social Computational Psychiatry’, ‘Computational Psychopathology’.

Supplement II. Study Details

This section reports two tables: Table S1, which outlines the key characteristics of each reviewed study (including its main aim, results, and evaluation) and Table S2, which outlines some details on the experimental design of each study (e.g., whether humans interacted with other humans or artificial agents).

Table S1. Key characteristics of included studies.

DYNAMICAL SYSTEMS MODELING								
STUDY OVERVIEW			STUDY RESULTS			STUDY EVALUATION		
Reference	Study Aim	Study Methods	Simulation Results	Empirical Results	Implications	Validity	Performance	Transparency
THERAPEUTIC RELATIONSHIP								
1	Liebovitch et al. (2011)	Formalize psychotherapeutic relational dynamics	DS model with distinct equations for patient and therapist	The model showcased realistic relational dynamics in therapy	-	DS modeling can be used to simulate relational scenarios in therapy	Empirical=High Theoretical=High Generative=High	N/A registration N/A data No open code
2	Peluso et al. (2012)	Formalize psychotherapeutic relation	DS model with distinct equations for patient and therapist	The model discovered three types of therapeutic relating	-	DS modeling can be used to simulate therapeutic relational styles	Empirical=High Theoretical=High Generative=High	N/A registration N/A data No open code
3	Baker et al. (2021)	Examine the impact of emotional expression on therapeutic relation	DS modeling on three sessions for two patients (total 6 sessions)	-	Therapists from different psychotherapy schools exhibited similar relational dynamics	DS analyses can empirically reveal relational dynamics of psychotherapy	Empirical=High Theoretical=High Generative=High	No registration No open data No open code
4	Diaz et al. (2023)	Examine relational attractors during psychotherapy	DS modeling on six 45min sessions with one therapist and patient	-	Therapist-patient attractors shifted from positive-positive, to positive-negative, and positive-positive	DS analyses can reveal unfolding relational attractors during psychotherapy	Empirical=High Theoretical=High Generative=High	No registration No open data No open code
5	Schiepek et al. (2016)	Examine complex relational patterns during psychotherapy	Complex DS model with five nonlinear difference equations and a total of 16 functions	The model produced various psychotherapy phenomena based on cognitive, affective, and relational parameters	-	DS modeling can be used to simulate therapeutic change processes with balanced model complexity	Empirical=Low Theoretical=High Generative=Low	N/A registration N/A data No open code
6	Westermann & Banisch (2025)	Formalize interpersonal affiliation in both health and mental illness	One-dimensional DS model that includes the state of the relationship and the motives of interactants	The model produced stable, bistable, and cyclical patterns (with the latter resembling BPD's relational style).	-	DS modeling can formalize stable, complementary, and unstable affiliative processes	Empirical=Low Theoretical=High Generative=Low	N/A registration N/A data Open code: https://osf.io/wfehp
7	Tschacher, Haken, and Kyselo (2015)	Formalize therapeutic alliance	One-dimensional DS model with poles of distinction-participation	The model shows how patients could shift from maladaptive attractor states to adaptive ones during psychotherapy	-	DS modeling can be used to model the relation between two agents (rather than the two agents themselves)	Empirical=High Theoretical=High Generative=High	N/A registration N/A data No open code
INTERPERSONAL SYNCRHONY								
8	Saint-Georges et al. (2011)	Examine the utility of interaction synchrony as a diagnostic marker of ASD and ID in infants	Markov modeling of videotaped interactions between infants with ID (N=12), ASD (N=15), and HCs (N=15)	-	ASD and ID infants exhibit deviant autistic behaviours and delayed behaviours, respectively; and their parents exhibit compensatory soliciting behaviours	More credence should be given to parents as they could be the first to recognize, through their interactions, their child's developmental problems	Empirical=High Theoretical=Low Generative=Low	No registration No open data No open code
9	Varlet et al. (2012)	Examine social motor coordination in schizophrenia	DS modeling of social motor coordination in HC group (20 pairs of HCs) and SCZ group (10 HCs & 10 SCZ paired)	-	A coupled oscillator showed that SCZ were impaired in intentional, but not unintentional, motor coordination due to a delay in information transmission	DS modeling can be used to reveal relational motor signatures underlying social deficits in various interpersonal disorders	Empirical=High Theoretical=Low Generative=Low	No registration No open data No open code
10	Schreiber et al. (2020)	Examine the role of physiological alignment in relational dysfunction in people with PD (personality disorder)	Dynamical systems model of heart rate changes on 121 couples, of whom one partner had PD	-	Greater contrarian coregulation is related to short-term discord and long-term relational dysfunction in PD couples	Physiological misalignment may be one mechanism underlying relational problems in PD romantic relations	Empirical=High Theoretical=Low Generative=Low	R ² =.61 No registration No open data No open code
11	Tschacher & Haken (2019)	Formalize the process of psychotherapy change using deterministic and stochastic functions	DS modeling of heart-rate and respiratory data from 20 therapy sessions (each approx. 50mins)	-	Heart-rate and respiratory-rate attractor signatures are linked to positive therapy self-reports	DS modeling can link dynamic nonverbal data to psychotherapy outcomes	Empirical=High Theoretical=High Generative=High	No registration No open data No open code
12	Hale and Aarts (2023)	Examine whether relational dynamics can predict improvement of depression during psychotherapy	HMM of nonverbal behaviours from N=39 patients with depression	-	The HMM detected dynamics of hyperfocus which were linked to improvement in depression	HMMs are a promising tool for capturing interpersonal patterns that are important for therapeutic outcomes	Empirical=High Theoretical=Low Generative=Low	Three state HMM fitted data best: AIC=4472.76, mean R-hat=1.11 No registration No open data Open code: https://zenodo.org/records/7621460
13	Paz et al. (2021)	Examine intrapersonal and interpersonal affect dynamics during psychotherapy	DS modeling on 279 therapy sessions with 21,133 vocal arousal observations	-	Interpersonal dampening ('being pulled to the other's baseline arousal level') is associated with better session outcomes	DS modeling can link dynamic vocal arousal data to psychotherapy outcomes	Empirical=High Theoretical=Low Generative=Low	No registration No open data No open code
REINFORCEMENT LEARNING								
STUDY OVERVIEW			STUDY RESULTS			STUDY EVALUATION		
Reference	Study Aim	Study Methods	Simulation Results	Empirical Results	Implications	Validity	Reliability	Transparency
TRANSDIAGNOSTIC SOCIAL LEARNING								
14	Will et al. (2017)	Examine fluctuations in self-esteem and how they relate to psychiatric symptoms and brain activations	Exponential kernel regression modeling of a social evaluation task with 44 university students	-	Interpersonal vulnerability modulated PEs in insula-vmPFC during self-esteem updates	Learning about the self might be based on similar mechanisms as learning about others	Empirical=High Theoretical=Low Generative=Low	Best fitting model exhibited mean R ² =0.31, median R ² =0.27, and BIC=-633
15	Will et al. (2020)	Examine the neural and computational underpinnings of (low) self-esteem	Exponential kernel regression modeling of a social evaluation task with participants scoring low (N=61) and high (N=61) self-esteem	-	Low baseline esteem yielded low learning and high volatility in esteem and was related to relational vulnerability, which was linked to specific brain regions	Self-esteem might be underpinned by specific computational signatures (and neural substrates) that present a general vulnerability to psychiatric disorder	Empirical=High Theoretical=Low Generative=Low	Best fitting model exhibited mean pseudo-R ² of behaviour = 0.71, mean R ² of worth ratings = 0.24, and BIC=-1693
16	Contreras-Huerta et al. (2022)	Examine the link of prosocial behaviour and affective sensitivity	Reinforcement learning of harm aversion and prosocial effort tasks	-	Individuals with higher affective sensitivity manifested greater	Emotional sensitivity may promote prosocial	Empirical=High Theoretical=Low	No registration Open data and code: https://osf.io/hkbzg/?view

		transdiagnostically	with N=212 participants		prosocial behaviour across the two tasks	behaviour	Generative=Low	only=b6eefb6fe79d4f448b7f0adc7979834b
17	Lenow, Cisler, & Bush (2018)	Examine computational mechanisms underlying trust learning in victims of interpersonal violence	Reinforcement learning of social learning task with N=15 adolescent girls who were victims of interpersonal violence and N=17 HCs	-	Victims of interpersonal violence had higher learning of trust but only when they attributed trust inconsistently	Relational credulity may be an adaptive response to unstable social settings	Empirical=High Theoretical=Low Generative=Low	No performance metrics No registration No open data No open code
18	Barnby et al. (2022)	Examine the relationship between reinforcement learning, social inference, and paranoia	Reinforcement learning and Bayesian inference in a probabilistic reversal learning task and a repeated Dictator game with N=693	-	Paranoia was related to uncertainty in both social and non-social contexts and across diverse models/tasks	Paranoia is associated with a general uncertainty over states of the world (particularly in the social sphere)	Empirical=High Theoretical=Low Generative=High	Best fitting (Bayesian) model exhibited LL=-44.2, BIC=115, and AIC=106 Registration: aspredicted.org/ds9bf.pdf aspredicted.org/57p5e.pdf Open data and code: https://github.com/josephmbarnby/Barnby_etal_2022_ReversalLearning/
SOCIAL SENSITIVITY IN BORDERLINE PERSONALITY DISORDER								
19	Fineberg et al. (2018)	Examine social learning in patients with BPD	Reinforcement learning of a social valuation task with BPD people (N=20) and HCs (N=23)	-	BPD subjects weighted social cues more heavily but exhibited blunted learning during (social and non-social) volatility	Individuals with BPD expect volatility, which might explain their relational difficulties	Empirical=High Theoretical=Low Generative=Low	No performance metrics No registration No open data No open code
20	Shapiro-Thompson et al. (2023)	Examine whether betrayal imagery evokes mistrustful behaviour in those with BPD	Reinforcement learning of a script imagery task on people with BPD (N=21) and HCs (N=20)	-	BPD participants were more sensitive to social cues and exhibited greater negative affect	Script-driven imagery presents a useful method for probing disorder-specific social difficulties	Empirical=High Theoretical=Low Generative=Low	No performance metrics No registration No open data No open code
21	Story et al. (2023)	Examine ‘self-other’ confusion in BPD	Reinforcement learning of a false-belief task on people with BPD (N=38) and matched HCs (N=74)	-	BPD participants tend to erroneously update their beliefs of others based on beliefs of themselves	Lack of self-other differentiation may be a computational process underlying BPD	Empirical=High Theoretical=Low Generative=Low	Parameter recoverability: Pearson’s r=0.82-0.85 Cohen’s d = 0.64 No registration No open data No open code
SLOW SOCIAL LEARNING IN DEPRESSION								
22	Safra, Chevallier, & Palminteri (2019)	Examine the relation between social decision-making with depression and anxiety	Q-learning modeling of a social learning task with 200 participants	-	Depressive, not anxiety, symptoms impaired task performance only in the social context	Audience effects (being watched by others) may contribute to depressive social deficits	Empirical=High Theoretical=Low Generative=Low	PP=90±3%, EP=100% Average parameter recoverability: r = 0.73±.01 No registration Open data and code: https://rb.gv/0usysl https://rb.gv/173sxr
23	Frey, Frank, and McCabe (2021)	Examine social and non-social learning in patients with depression	Q learning modeling of a (non)social learning task with N=40 subjects scoring high and N=52 scoring low depression	-	Social learning was slower in people with high depression and predicted (across all participants) greater time spent in negative social situations	Depression may be characterized by deficits in social learning, which are directly linked to the quality of everyday experiences	Empirical=High Theoretical=Low Generative=Low	Best fitting models: AIC weights = .06/.08 {social} & .09/.08 {non-social}, <i>pseudo-R</i> ² = 0.34 {social} & 0.33 {non-social}, and parameter recoverability= Spearman rho=0.32-0.61 No registration No open data No open code

BAYESIAN MODELING								
STUDY OVERVIEW				STUDY RESULTS		STUDY EVALUATION		
Reference	Study Aim	Study Methods	Simulation Results	Empirical Results	Implications	Validity	Reliability	Transparency
AUTISM								
24	Yoshida et al. (2010)	Examine latent computational processes involved in ASD social interactions	Bayesian modeling of a social hunting game with N=17 HCs and N=17 ASD individuals	(No simulations here, but the model was based on Yoshida et al. (2008))	Impaired mentalizing and iterative planning were implicated in ASD	Computational modeling of game-theoretic tasks can reveal computational mechanisms underlying hallmark ASD symptoms	Bayesian model selection based on log-likelihoods (HC=-5776 vs ASD=-5330) HCs R ² =.23 for ToM model ASD R ² =.26 for fixed model	No registration No open data No open code
25	d’Arc, Devaine, and Daunizeau (2020)	Examine social behaviour adaptation in neurotypical and autistic individuals	Bayesian and reinforcement modeling of a dyadic competitive game with N=24 autistic and N=24 neurotypical participants playing against AI agents	-	ASD individuals do not use social knowledge to adapt to others, a deficit that predicts social symptoms and ASD diagnosis with 62% and 79% accuracy, resp.	Computational parameters that define the ASD phenotype may prove useful for guiding diagnostic and clinical practice	Classifier performance = 73-79% accuracy	No registration No open data No open code
26	Thomas et al. (2022)	Examine the contagion effect in neurotypical and autistic individuals	Bayesian modeling of a temporal discounting task with two neurotypical samples (N = 48; N = 98) and one ASD sample (N=12)	-	The contagion effect was equally evident in neurotypical and ASD populations, showing no privileged association with ASD traits	Despite widespread social impairments, ASD is characterized by a contagion of value preferences	No performance metrics	No registration No open data No open code
27	Barnby, Raihani, & Dayan (2022)	Examine how people learn about themselves and others and use those models to make social inferences	Bayesian Inference and reinforcement learning in a social value orientation task (N=697)	-	Discrepancies in self-other representations led to greater paranoid attributions of social situations	Maladaptive updating of social representations is a key computational mechanism of paranoia	Model Responsibility = 1 Log likelihood = -5.41	No registration Open data and code: https://github.com/josephmbarnby/Barnby_etal_2021_SVO
28	Na et al. (2022)	Examine computational mechanisms underlying illusion of control in individuals with high/ low delusional traits (D)	Forward thinking modeling of a two-party exchange game with low D people (N=126) and high D people (N=125)	-	Individuals with high, but not low, D exhibited illusion of control at both objective and subjective levels	People with high D are more likely to believe that they can control others, particularly in unstable environments	Elbow point of Draper’s Information Criterion = 10	No registration No open data No open code
29	Barnby et al. (2020)	Examine computational parameters of harmful intent attributions in paranoia	Bayesian inference in a serial dictator game with N=1754	-	Paranoia was associated with greater uncertainty about others’ actions and enhanced learning of harmful intent attributions	Paranoia may be underpinned by an increased need to attend to immediate social experiences	Log likelihood = -4.394 Simulated attributions are highly correlated with real attributions (rho=0.8-0.9)	Registration: http://aspredicted.org/bli_nd.php?x=8cj8zk http://aspredicted.org/bli_nd.php?x=ub9z2x Open data and code: https://osf.io/24urf/
30	Alon et al. (2024)	Formalize paranoia using Theory of Mind	Interactive POMD with different levels of mentalizing	Agents who hyper-mentalize can counter deception but also misinterpreted non-strategic behavior as intentional deception	-	Paranoia and conspiratorial thinking may stem from <i>over-application</i> of strategic reasoning to non-strategic behavior	N/A	N/A registration N/A open data No open code
31	Siegel et al. (2020)	Examine whether moral inference is disrupted in patients with BPD	Hierarchical Bayesian modeling of a moral inference task with BPD (N=20), treated (T)BPD (N=23), and HCs (N=102)	-	PERSONALITY DISORDER BPD patients exhibited more certain and rigid beliefs for harmful agents, compare to treated BPDs and HCs The model revealed that individuals with BPD scored significantly higher on computational parameters quantifying	BPD patients exhibit high certainty and rigidity in their beliefs which could be ameliorated through psychotherapy Splitting can be formalized as an HMM and applied on data to test specific hypotheses about idealization /	No performance metrics	No registration No open data No open code
32	Story et al. (2023)	Formalize splitting (that is, devaluation and idealization patterns)	HMM of moral inference task on people with BPD (N=20), treated BPD (N=23), and HCs (N=102)	The model was able to reproduce patterns of idealization and devaluation			ΔBIC=18/135, mean log LR =12.5/71.1; likelihood ratio test: χ ² (2)=25/142, <i>p</i> <.0001, Mcfadden’s pseudo-R ² for best-fitting	No registration No open data No open code

33	Xiang et al. (2012)	Examine ‘depth’ of mentalizing	Bayesian theory-of-mind model (as well as RL model) in a repeated economic exchange task with N=195 pairs (HC=102 and BPD=93)	-	splitting than HCs The model was able to capture depth of mentalizing, relate it to cognitive and neural markers, and show that it is ‘shallower’ in BPD	devaluation patterns Computational modeling of game theoretic tasks can reveal depth of mentalizing and use it to distinguish BPD from healthy controls	Empirical=High Theoretical=High Generative=High	(split HMM) model = 0.22 Averaged log likelihood = −11.92±0.27	No registration No open data No open code
34	Hula et al. (2018)	Address failures of a previous model on cooperation by adding two new parameters: risk aversion; irritation	POMDP modeling of a multi-round trust game with 93 healthy controls (HCs) playing with either 38 HCs or 55 BPDs	Including the two parameters led to better (AIC/BIC) model fit and explained irritation phenomena	Individuals with BPD have failures in detecting others’ irritability and exhibit lower guilt than HCs	Computational mechanisms relating to failed mentalizing and lower guilt can explain BPD relational ruptures	Empirical=High Theoretical=High Generative=High	Likelihood ratio tests $p < 10^{-46}$ and average BIC=21.68	No registration Open data and code: https://github.com/AndreasHula/StateShiftPaperData
35	Xiao et al. (2023)	Assess guilt-related relational responses in those with OCPD	Guilt aversion modeling of two social tasks with 46 people with OCPD and 67 HCs	-	Individuals with OCPD exhibited less guilt aversion and less guilt-induced compensation in their interactions	Individuals with OCPD are less affected by guilt when inter-relating	Empirical=High Theoretical=Low Generative=Low	Task 1 Accuracy =0.83 [95% CI (0.81–0.85)] Task 2 Accuracy = 0.85 [95% CI (0.82–0.87)]	No registration No open data No open code
36	Zavlis et al. (2024)	Formalise personality disorder as a dynamic relational disorder	Hidden Markov model of mental inference about oneself and others	Model can reproduce various personality phenomena (from borderline instability to narcissistic grandiosity)	-	Personality disorders can be more formally viewed as relational disorders: ways of relating to the self and others	Empirical=Low Theoretical=High Generative=Low	N/A	N/A registration N/A open data Open code: https://github.com/OrestisZavlis/RelationalDisorder
INTERNALIZING DISORDERS									
37	Henco et al. (2020)	Examine learning and decision-making within social and non-social domains in patients with BPD, SCZ, and MDD	Bayesian modeling (and RL alternative model) of a probabilistic learning (social & non-social) task in patients with BPD (N=28), SCZ (N=29), MDD (N=28), HC (N=31)	-	BPD participants had slower learning and exaggerated sensitivity to changes in both task domains; SCZ and BPD relied more on social vs non-social information	Patients with BPD and SCZ are sensitive to social information, which could explain their relational difficulties	Empirical=High Theoretical=Low Generative=Low	Best-fitting model exhibited PP = 0.30, XP=.91, and PXP=0.17	No registration Open data and code: https://osf.io/8kfph/
38	Lamba, Frank, and FeldmanHall (2020)	Examine how people with low versus high anxiety learn (socially) under uncertainty	Bayesian (as well as RL) modeling on a dynamic trust game with N=257 reporting low and N=97 reporting high anxiety	-	Subjects with anxiety overinvested in exploitative partners, a pattern related to two distinct computational mechanisms	Anxiety impairs social learning under conditions of uncertainty	Empirical=High Theoretical=Low Generative=High	Protected Exceedance Probability (PXP) > .56–.99	No registration Open data and code: https://osf.io/ea67f/
FREE ENERGY PERSPECTIVES									
39	Prosser et al. (2018)	Formalize psychopathy	POMD model within a Free Energy landscape	The model reproduced to key aspects of psychopathy: lacks remorse and self-aggrandizing	-	Psychopathy may emerge via hyperpriors defining the self as remorseless and aggrandizing	Empirical=Low Theoretical=High Generative=Low	N/A	N/A registration N/A data No open code
40	Cittern et al. (2018)	Formalize attachment styles	POMD model within a Free Energy landscape	The model reproduced secure, ambivalent, avoidant, disorganized attachment styles	-	Attachment styles may emerge as a way of minimizing epistemic uncertainty regarding caregiving responsivity	Empirical=Low Theoretical=High Generative=Low	N/A	N/A registration N/A data No open code
41	Constant et al. (2021)	Formalize an evolutionary systems theory of depressed mood	POMD model within a Free Energy landscape	The model was able to reproduce depressed mood and show how it can be ameliorated via social support or pharmacotherapy	-	Depression may be an adaptive response to interpersonal adversity and be ameliorated via social support or specific pharmacotherapies	Empirical=Low Theoretical=High Generative=Low	N/A	N/A registration N/A data No open code

MACHINE LEARNING									
STUDY OVERVIEW					STUDY RESULTS		STUDY EVALUATION		
Reference	Study Aim	Study Methods	Simulation Results	Empirical Results	Implications	Validity	Reliability	Transparency	
42	Kang et al. (2015)	Examine whether Facebook activities can be used to classify people’s attachment styles	SVM on 640 Facebook users and their 525,334 posts	-	CLASSIFICATION Being ‘tagged’ and highly ‘responsive’ have strong classifying utility for attachment anxiety and receiving likes and comments for attachment avoidance	Attachment styles can be deciphered through people’s self-expression and responsiveness in Facebook	N/A	10-fold CV Precision: 0.71-0.91 F1-scores: 0.61-0.86	No registration No open data No open code
43	Gómez-Zaragoza et al. (2023)	Develop an attachment-recognition model that can distinguish secure-insecure attachment styles from voice recordings	Supervised ML models (SVM, RF, LR, and KNN) were trained on 83% of audio data from 199 subjects	-	Sex-dependent models performed best with accuracy reaching 63.88% and 83.63% for females and males, respectively	ML can be used to remotely assess attachment styles	N/A	5-fold CV with 83-17% split Test Accuracy: 58.8-83.6% Test AUC: 0.58-0.83 Sensitivity: 0.55-0.89 Specificity: 0.51-0.78	No registration No open data No open code
44	Antonucci et al. (2021)	Classify psychosis using parenting and attachment variables	SVM classification of 34PSY and 71HC with external validation using two samples (60HC / 30PSY) & (26HC/13PSY)	-	BAC ranged from 44.2% in validation sample to 72.2% in main sample	Parenting and attachment variables have strong classification power in PSY patients	N/A	Nested CV Sensitivity=82.4%; Specificity=62.2% AUC=0.71 External validation on clinical and family samples	No registration No open data No open code
45	Doborjeh et al. (2023)	Identify social and cognitive predictors of (two year) outcomes relating to psychosis	Spiking neural networks were used on 90 ultra-high-risk (UHR) patients (remitters, converters, and maintained) and 81 healthy controls	-	Accuracy up to 80% and strongest predictors were cognitive, affect, and interpersonal variables	Spiking neural networks show promise in identifying patients with high risk of, but no transition to, psychosis	N/A	30x CV based on 50% split Total Accuracy: 0.78-0.80	No registration No open data No open code
46	Haghighi et al. (2023)	Identify both risk and protective factors of adolescent suicidal attempts	Five ML algorithms trained with 80-20% split on a representative adolescent Norway sample (N=173,644)	-	PREDICTION The stacked ensemble algorithm performed the best and interpersonal variables were the strongest predictors	ML can identify youth at risk for suicide using interpersonal variables	N/A	10-fold CV with 80-20 split SNS*SPC: 0.9 (0.89-0.91) AUC: 0.96 (0.96-0.97) AUCPR: 0.67 (0.65-0.69)	No registration No open data No open code
47	Handing et al. (2023)	Compare 56 risk/protective predictors of depression	RF was applied in a European representative sample of middle-aged people (N=67,603)	-	Self-rated social isolation and self-rated poor health were the strongest risk factors	ML can be used to identify strong predictors of depression and tease out their differential effect across sexes	N/A	Accuracy: M = 0.76 for females and M = 0.82 for males	No registration No open data No open code
48	Wang et al. (2023)	Examine the role of various risk/protective factors in adolescent depression	Association rule mining was used on 2445 children-adolescents from China	-	Family cohesion and peer and teacher support were the strongest predictors of	Mining association rules can be used to reveal strong predictors of youth depression	N/A	Support: 5.0-33.2% Confidence: 70.2-90.5% Lift: 1.1-3.0%	No registration No open data No open code

49	Hu et al. (2023)	Examine the role of interpersonal factors in depression & resilience	RF was applied on a cross-sectional sample of N=5952 adolescents	-	youth depression Parental support was a stronger predictor than friend support in predicting depression and resilience	Parental interpersonal support is important for youth mental functioning	N/A	No performance metrics	No registration No open data No open code
50	Schorr et al. (2021)	Examine the role of childhood trauma and parental bonding in antisocial personality disorder (ASPD)	Elastic net trained using 80-20% split on N=346 male cocaine users	-	Emotional and physical abuse and parental care and control were the most important predictors of ASPD	Early trauma and key relational variables are key predictors of ASPD	N/A	10-fold CV on 80-20% split Specificity: 85% Sensitivity: 50% AUC: 75,5%	No registration No open data No open code
51	Lu et al. (2022)	Examine the role of psychosocial factors in the prediction of aggression in people with drug addiction	GB was applied with five-fold CV to N=896 males with drug addiction	-	The most important predictors of aggression were interpersonal trust, psychological security and capital, parental conflict, and alexithymia	Interpersonal variables are key predictors of aggression in people with drug addiction	N/A	No performance metrics	No registration No open data No open code
NATURAL LANGUAGE PROCESSING									
52	Martinez et al. (2019)	Discover relational characters from narrative data ('personas') and examine their predictive utility on alliance	Unsupervised topic modeling of 1,235 transcribed sessions (N=386 patients and N=40 therapists) to discover personas and SVM for predicting alliance	-	Alliance can be explained by the interactions between the discovered personas (better than observed linguistic data)	A combination of unsupervised and supervised NLP methods can reveal important relational patterns for therapeutic alliance	N/A	CV with leave-one-therapist-out $AIC_{min} = 721.32, \chi^2(2) = 119.09, p < 0.001$ MSE = 0.69 ($\sigma=0.50$)	No registration No open data No open code
53	Atzil-Slonim et al. (2021)	Examine whether topic models can identify patient's functioning and alliance ruptures	Unsupervised LDA modeling of 873 transcribed sessions (with N=58 patients and N=52 therapists) to identify therapy topics and supervised SMLR to predict alliance ruptures	-	{Loneliness, suffering, physical issues, anger} and {communication, goal-setting, and needing help} were the strongest predictors of {low functioning} and {therapeutic ruptures}, respectively.	A combination of unsupervised and supervised NLP methods can identify important predictors of functioning and therapeutic alliance	N/A	CV on 80-20% data split Accuracy: 65% (alliance) and 75% (functioning)	No registration No open data No open code
54	Tsakalidis et al. (2021)	Examine whether NLP methods can be used to identify alliance ruptures	Skip-gram modeling of 873 transcribed sessions (with N=68 patients and N=52 therapists) and logistic regression for predicting patient and therapist rupture scores	-	The NLP model outperforms majority classifiers and captures 40% of patient-reported ruptures (ground truth) that were unidentified by therapists	NLP models can be used to identify patient-reported ruptures	N/A	CV with leave-one-patient-out (68 folds in total) F1-score: 70.9% Accuracy: 83.6%	No registration No open data No open code
55	Goldberg et al. (2020)	Predict patient-rated therapeutic alliance from session recordings	L2 regression with uni- and bi-grams of 1,235 transcribed sessions (with N=386 patients and N=40 therapists)	-	ML models with one- and two-word pairings modestly predict alliance ratings ($p=.15$)	NLP models have only modest predictive power but can reveal novel linguistic predictors of therapeutic alliance	N/A	10-fold CV MSE=0.67, Spearman's $\rho=0.15, p < .001$	No registration No open data No open code
56	Xu et al. (2021)	Identify premature departures in online text-based counselling	Pattern matching of 575 transcribed sessions (each with different patient-therapist pairs), validation on 34,821 sessions, and binary classification	-	The model found 43.5% premature departures and revealed that they were associated with lower patient-perceived helpfulness	A combination of logic rules and pattern matching can be used to detect premature departures in online counselling sessions	N/A	80-20% split Precision: 91% Recall: 94% F1-score: 92%	No registration No open data No open code
57	Tasca et al. (2023)	Develop a valid ML approach for automatically coding defense mechanisms	Five RoBERTa-based models were applied on 16,875 talk-turns between therapists and women with BED (N=92) and HC women (N=66)	-	The models were able to distinguish defenses, but no better than human coders	Further validation of NLP models is needed for categorical outcomes	N/A	10-fold CV with 5 epochs ROC-AUC=0.82, Accuracy=0.74, PR-AUC=0.60, F1=0.61, Precision=0.51, Recall=0.77, FPR=0.27, Pessimistic Accuracy=0.76	No registration No open data No open code
58	Zavlis, Fonagy, Luyten (2025)	Identify what therapy aims patients value most	Apply LDA on meta-analytic phrases from 2,908 patients' lived experience stories	-	LDA revealed three sentiments (love, work, and meaning) with love ('relational functioning') being the most reported by the patients	Relational functioning (that is, the capacity to better relate to oneself and others) might be the most preferred aim of psychotherapy	N/A	No performance metrics	No registration Open data and open code: https://osf.io/ya2r9/

ECONOMIC MODELS									
59	Driessen et al. (2021)	Examine the moral strategies of a group of non-offenders with varying degrees of psychopathy	Expected utility model of a trust game on N=86 healthy participants with N=20 scoring low, N=40 moderate, and N=26 high on total-score psychopathic traits	The model was able to produce hypothesized moral strategies: inequity aversion, guilt aversion, moral opportunism, greed, and generosity	A computational parameter quantifying inequality aversion, but not guilt, was negatively related to affective psychopathic traits as well as antisocial traits	Psychopathy may entail a reduced sense of fairness, not 'guilt' or 'greed' as traditional theories predict	Empirical=High Theoretical=High Generative=High	$\Delta BIC=-268.26/-79.12/-35.01, t(82)=-32.55/-12.28/-9.23, all p<.001$ Parameter recoverability: Person $r(86) \sim .99$	No registration No open data No open code
60	Barnby et al. (2025)	Examine self-other generalization patterns in people with BPD	Bayesian model with Fehr-Schmidt utility functions in intentions game with BPD (N=50) and HC (N=53) people	Simulations to validate model accuracy, validity, and parameter recoverability	People with BPD did not generalize their own preferences to others (and vice versa)	People with BPD may suffer from an inability to generalize information from the self to others and vice versa	Empirical=High Theoretical=High Generative=High	Model accuracy~0.8 Log-likelihood > chance Parameter recoverability: ($r = 0.88, p < 0.001$)	No registration Open data and code: https://github.com/josephmbarnby/SocialTransfer_Barnby_et al 2024

Note. DS = Dynamical Systems, HC = Healthy (or nonclinical) Controls, SCZ = Schizophrenia, ASD = Autism Spectrum Disorders, BPD = Borderline Personality Disorder, OCPD = Obsessive-Compulsive Personality Disorder, HMM = Hidden Markov Model, RL = Reinforcement Learning, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion, PXP = Protected Exceedance Probability, PP = Posterior model probability, XP = Exceedance Probability, POMD = Partially Observable Markov Decision process, PEs = Prediction Errors, vmPFC = ventromedial Prefrontal Cortex, LL = Log Likelihood. ML = Machine Learning, SVM = Support Vector Machine, NLP = Natural Language Processing, BAC = Balanced Accuracy, CV = Cross-validation, GB = Gradient Boosting, RF = Random Forest, LR = Logistic Regression. KNN = K-Nearest Neighbours.

Table S2. Study Design

DYNAMICAL SYSTEMS
<p>1. Baker et al. (Dynamical systems modeling in psychotherapy)</p> <ul style="list-style-type: none">(1) Human participants interacting with other humans (therapist-client interactions).(2) Measures bilateral social behaviours (dynamical systems modeling focuses on bidirectional, real-time emotional and relational dynamics).
<p>2. Diaz et al. (Role of emotion/expertise in psychotherapy)</p> <ul style="list-style-type: none">(1) Human participants interacting with other humans (therapist-client interactions).(2) Measures bilateral social behaviours (analyses interaction dynamics over therapy sessions using dynamical systems).
<p>3. Saint-Georges et al. (Autistic deviant behavior recognition)</p> <ul style="list-style-type: none">(1) Human participants interacting with other humans (parent-child interactions).(2) Measures bilateral social behaviours (hidden Markov modeling to analyse interaction patterns, not unilateral choices).
<p>4. Varlet et al. (Social motor coordination in schizophrenia)</p> <ul style="list-style-type: none">(1) Human participants interacting with other humans (coordination tasks between individuals).(2) Measures bilateral social behaviours (focuses on real-time motor coordination dynamics).
<p>5. Schreiber et al. (Physiological coregulation in couples)</p> <ul style="list-style-type: none">(1) Human participants interacting with other humans (couples in conflict scenarios).(2) Measures bilateral social behaviours (physiological synchrony during real-time interactions).

6. Tschacher & Haken (Causation and chance in psychotherapy)

- (1) Human participants interacting with **other humans** (therapist-patient interactions).
- (2) Measures **bilateral social behaviours** (models stochastic and deterministic interaction processes over time).

7. Hale & Aarts (Hidden Markov models in psychotherapy)

- (1) Human participants interacting with **other humans** (therapist-patient interactions).
- (2) Measures **bilateral social behaviours** (interpersonal interaction dynamics are tracked across sessions).

8. Paz et al. (Vocal affect dynamics in psychotherapy)

- (1) Human participants interacting with **other humans** (therapist-patient vocal exchanges).
- (2) Measures **bilateral social behaviours** (vocal affect dynamics reflect real-time mutual influence).

REINFORCEMENT LEARNING

9. Will et al. (Neural/computational processes in self-esteem)

- (1) Interaction with **artificial agents** (tasks with simulated social feedback, e.g., hypothetical social evaluations).
- (2) Measures **unilateral social behaviours** (participants make choices or respond to static feedback without real-time interaction).

10. Will et al. (Aberrant social learning in low self-esteem)

- (1) Interaction with **artificial agents** (social learning tasks using pre-programmed cues or simulated outcomes).
 - (2) Measures **unilateral social behaviours** (learning from static social/non-social feedback, no dynamic interaction).
-

11. Contreras-Huerta et al. (Prosocial behavior and affective sensitivity)

- (1) Interaction with **artificial agents** (economic games with hypothetical partners).
- (2) Measures **unilateral social behaviours** (one-off prosocial decisions without dynamic feedback).

12. Lenow et al. (Trust learning in interpersonal violence victims)

- (1) Interaction with **artificial agents** (trust games with algorithmic partners).
- (2) Measures **unilateral social behaviours** (trust choices based on static outcomes, no real-time reciprocity).

13. Barnby et al. (Reinforcement learning, social inference, paranoia)

- (1) Interaction with **artificial agents** (social inference tasks using simulated scenarios).
- (2) Measures **unilateral social behaviours** (participants infer intentions from static cues).

14. Fineberg et al. (Differential valuation in BPD)

- (1) Interaction with **artificial agents** (social/nonsocial cues presented as stimuli, no live interaction).
- (2) Measures **unilateral social behaviours** (valuation of cues without dynamic feedback).

15. Shapiro-Thompson et al. (Trust modulation via imaginal exposure)

- (1) **Imagined interaction** (script-based imagination of betrayal scenarios).
 - (2) Measures **unilateral social behaviours** (self-reported trust changes post-imaginal exposure).
-

16. Story et al. (Self-other mergence in BPD)

- (1) Interaction with **artificial agents** (computational tasks involving judgments of self/others).
- (2) Measures **unilateral social behaviours** (ratings without real-time interaction).

17. Safra et al. (Depressive symptoms and social reward learning)

- (1) Interaction with **artificial agents** (social reward tasks with simulated feedback).
- (2) Measures **unilateral social behaviours** (learning from static social rewards).

18. Frey et al. (Social reinforcement learning in depression)

- (1) Interaction with **artificial agents** (social feedback tasks using pre-defined outcomes).
- (2) Measures **unilateral social behaviours** (learning from non-dynamic social cues).

BAYESIAN INFERENCE

19. Yoshida et al. (Cooperation in Autism)

- (1) Human participants interacting with **artificial agents** (tasks with simulated partners or hypothetical scenarios).
- (2) Measures **unilateral social behaviours** (cooperative choices in structured games without real-time interaction).

20. Forgeot d'Arc et al. (Social adaptation in Autism)

- (1) Human participants interacting with **artificial agents** (computational tasks involving simulated social scenarios).
 - (2) Measures **unilateral social behaviours** (adaptive choices based on static feedback).
-

21. Thomas et al. (Contagion of Temporal Discounting)

- (1) Interaction with **artificial agents** (observing/imitating choices from simulated others).
- (2) Measures **unilateral social behaviours** (individual discounting preferences influenced by static social cues).

22. Barnby et al. (Interpersonal similarity and intent attribution)

- (1) Interaction with **artificial agents** (hypothetical scenarios or simulated profiles).
- (2) Measures **unilateral social behaviours** (predictive judgments without dynamic interaction).

23. Na et al. (Illusion of control in delusional individuals)

- (1) Interaction with **artificial agents** (tasks with simulated control over outcomes).
- (2) Measures **unilateral social behaviours** (perceived control in non-interactive scenarios).

24. Barnby et al. (Social learning in paranoia)

- (1) Interaction with **artificial agents** (modified dictator game with algorithmic partners).
- (2) Measures **unilateral social behaviours** (static learning and policy uncertainty).

25. Siegel et al. (Disrupted moral inference in BPD)

- (1) Interaction with **artificial agents** (hypothetical moral dilemmas or simulated agents).
 - (2) Measures **unilateral social behaviours** (moral judgments without real-time feedback).
-

26. Story et al. (Social inference of idealization/devaluation)

- (1) Interaction with **artificial agents** (tasks with social evaluations).
 - (2) Measures **unilateral social behaviours** (static trait inferences).
-

27. Xiang et al. (Neural response to depth-of-thought)

- (1) Human participants interacting with **other humans** (two-person interactive tasks).
 - (2) Measures **bilateral social behaviours** (real-time neural coordination during interaction).
-

28. Hula et al. (Risk and mental state shifts)

- (1) Human participants interacting with **other humans** (social interaction tasks).
 - (2) Measures **bilateral social behaviours** (dynamic risk-taking and mental state changes).
-

29. Xiao et al. (Guilt-related dysfunction in OCPD)

- (1) Human participants interacting with **other humans** (two social interaction tasks).
 - (2) Measures **bilateral social behaviours** (modeling guilt dynamics in real interactions).
-

30. Henco et al. (Social learning in schizophrenia/BPD)

- (1) Interaction with **artificial agents** (computational tasks with simulated feedback).
 - (2) Measures **unilateral social behaviours** (learning static social/nonsocial cues).
-

31. Lamba et al. (Anxiety and social learning under uncertainty)

- (1) Interaction with **artificial agents** (pre-programmed social feedback).
 - (2) Measures **unilateral social behaviours** (learning under uncertainty).
-

Supplement III. Risk of Bias Assessment

The [NIH Quality Assessment Tools](#) were used to examine the risk of all theory-driven empirical studies (i.e., dynamical systems, reinforcement learning models, and Bayesian models). Dynamical systems studies were assessed using the risk assessment tool meant for observational cohort and cross-sectional studies, while reinforcement learning and Bayesian studies were assessed using the risk assessment tool meant for case-control studies.

For machine learning studies, the well-known [PROBAST tool](#) was used to examine risk of bias in both supervised and unsupervised studies (with some slight modifications being made for PROBAST to be able to examine unsupervised studies that only contained unlabelled data and focused on exploration rather than prediction) (see Table S7 or more information).

Table S3. NIH Quality assessment of dynamical systems studies.

Study	Objective	Population	Participation	Selection	Sample Size	Exposure Before Outcome	Timeframe	Exposure Levels	Exposure Validity	Repeat Exposure	Outcome Validity	Blinding	Loss to Follow-Up	Confounder Control	Bias
Baker et al.	Yes	Yes	Not reported	No	No	Yes	Yes	No	No	Yes	No	NA	NA	No	High
Diaz et al.	Yes	Yes	Not reported	No	No	Yes	Yes	No	No	Yes	No	NA	NA	No	High
Saint-Geor. et al.	Yes	Yes	Not reported	Yes	No	Yes	Yes	No	No	No	No	NA	NA	No	High
Varlet et al.	Yes	Yes	Not reported	Yes	No	Yes	Yes	No	Yes	Yes	No	NA	NA	No	Low
Schreiber et al.	Yes	Yes	Not reported	Yes	No	Yes	Yes	No	Yes	Yes	Yes	NA	NA	Yes	Low
Tschacher & Haken	Yes	Yes	Not reported	No	No	Yes	Yes	No	Yes	No	Yes	NA	NA	No	Low
Hale & Aarts	Yes	Yes	Not reported	Yes	No	Yes	Yes	No	Yes	Yes	Yes	NA	NA	No	Low
Paz et al.	Yes	Yes	Not reported	No	No	Yes	Yes	No	Yes	Yes	No	NA	NA	No	High

Table S4. NIH Quality assessment of reinforcement learning studies.

Study	Research question	Target population	Sample-size justification	Controls	Uniform Criteria	Clear Cases	Random selection	Concurrent controls	Exposure before diagnosis	Exposure validity	Exposure assessors blinded	Confounders	Bias
Will 2017	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Will 2020	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Contreras-Huerta 2022	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Lenow 2018	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Barnby 2022	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Fineberg 2018	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Shapiro-Thompson 2023	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Story 2023	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Safra 2019	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Frey 2021	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate

Table S5. NIH Quality assessment of Bayesian studies.

Study	Research question	Target population	Sample-size justification	Controls	Uniform criteria	Clear cases	Random selection	Concurrent controls	Exposure before diagnosis	Exposure validity	Exposure assessors blinded	Confounders measured & adjusted	Bias
Yoshida 2010	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
d’Arc 2020	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Thomas 2022	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Barnby 2022	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Na 2022	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Barnby 2020	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Siegel 2020	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Story 2023	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Xiang 2012	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Hula 2018	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	No	Moderate
Xiao 2023	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Henco 2020	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low
Lamba 2020	Yes	Yes	No	Yes	Yes	Yes	NA	Yes	No	Yes	NA	Yes	Low

Table S6. PROBAST assessment of supervised machine learning studies.

Study (year)	Participants	Predictors	Outcome	Analysis	Overall
Kang 2015	High	Low	High	High	High
Gómez-Zaragozá 2023	High	Low	High	High	High
Antonucci 2021	High	Low	Low	High	High
Doborjeh 2023	High	Low	Low	High	High
Haghish 2023	Low	Low	Low	High	High
Handing 2023	Low	Low	Low	High	High
Wang 2023	High	Low	Low	High	High
Hu 2023	High	Low	Low	High	High
Schorr 2021	High	Low	Low	High	High
Lu 2022	High	Low	Low	High	High

Note. All supervised ML studies were considered to have high risk of bias because they suffered major analytical flaws (e.g., every study relied solely on internal CV or a single random split; none reported external/temporal validation or calibration curves; and hyper-parameter tuning was often done inside the test fold).

Table S7. PROBAST-modified assessment of unsupervised machine learning studies.

Study	Participants	Predictors	Outcome	Analysis	Overall
Martinez 2019	Moderate	Moderate	NA	High	High
Atzil-Slonim 2021	Moderate	Moderate	NA	High	High
Tsakalidis 2021	Moderate	Moderate	NA	Moderate	Moderate
Goldberg 2020	Moderate	Moderate	NA	High	High
Xu 2021	Low	Moderate	NA	Moderate	Low
Tasca 2023	Moderate	Moderate	NA	High	High
Zavlis 2025	Low	Moderate	NA	Low	Low

Note. The standard PROBAST RoB tool was slightly modified to address the unsupervised natural language processing studies. The ‘Participants’ category remained the same; ‘Predictors’ mean all unstructured data; ‘Outcome’ was not applicable (because this is unstructured data), and analysis was applied specifically for the natural language processing methods.

Supplement IV. Validity Assessment

In this Supplement, we outline the complete results of our validity assessment. Specifically, in Tables S8, S9, and S10, we have outlined the values quantifying the empirical, theoretical, and generative validities of dynamical systems, reinforcement learning models, and Bayesian models, respectively. Overall, four classes of models were identified across these computational approaches: (1) models that scored highly only on empirical validity (because they were primarily data-driven), (2) models that scored highly only on theoretical validity (because they were based primarily on simulations), (3) models that scored highly on both theoretical and generative validity (because they were based both on simulations and previous empirical investigations) and (4) models that scored highly on all types of validity (because they were created in a theory-driven way and validated in a data-driven way).

Starting with dynamical systems models, we note that they clustered in three types (see Table S1). **First**, five models scored highly on only empirical validity, as they were only used empirically to examine interpersonal synchrony in various mental health difficulties.¹⁻⁵ **Second**, two models scored highly on only theoretical validity, because they were based on computer simulations (but nevertheless made falsifiable predictions).^{6,7} **Finally**, two notable models, namely, Liebovitch et al.^{8,9} and Tschacher et al.^{10,11}, scored highly on all measures of validity. These models were constructed in a theory-driven manner by primary investigators, and then applied empirically on time-series data to reveal positive and negative relational attractors,^{12,13} on the one hand, and interpersonal synchrony patterns,¹¹ on the other. Therefore, these models were both theoretically validated (covering theoretical validity) and empirically validated (covering empirical and generative validity).

Moving on to reinforcement learning models, we note that all of them scored highly on only empirical validity (see Table S2). The main reason behind this is that all studies employing them were purely data-driven. Although it could be argued that RL models have strong theoretical backgrounds (e.g., because they are inspired by behavioural psychology), it is worth noting that accumulating research suggests that Bayesian alternatives ‘fit the data better.’ As an example, the notion of ‘surprise’ can be replaced with a more sophisticated notion of Bayesian surprise, which takes into account the informational value of stimuli; that is, that not all unexpected stimuli are important and therefore ‘surprising’.¹⁴ In that sense, purely model-free RL models may be regarded as a rather ‘incorrect’ model, but a useful one nonetheless when applied in data-driven contexts.

Finally, Bayesian models exhibited the most heterogeneity in their validity by clustering into all four types. **First**, nine Bayesian models scored highly only on empirical validity (because they were primarily used empirically to examine belief-updating).¹⁵⁻²³ **Second**, three Free Energy models scored highly only on theoretical validity (because they were only based on simulations).²⁴⁻²⁶ **Third**, two Bayesian models scored highly on both theoretical validity and generative validity (because they were based on both simulations and past empirical investigations, making more informed theoretical predictions).^{27,28} **Finally**, four Bayesian models scored highly on all measures of validity (because they were created in a theory-driven manner, generated falsifiable predictions in simulations, and then examined those predictions in empirical investigations).²⁹⁻³²

Table S8. Validity assessment for dynamical systems.

	EMPIRICAL VALIDITY				THEORETICAL VALIDITY		GENERATIVE VALIDITY	
	Model aims to explain a real-world phenomenon	The target condition is identifiable within the model	The comparator is identifiable within the model	The model explains target condition vis-à-vis comparator	Model manipulation yields expected outcomes	Simulated interventions match real interventions	Data-generating process is specific and plausible	Model architecture is homologous to the target mechanism
Liebovitch et al. (2011)	1	0.8	0.6	0.2	1	1	0.4	0.4
Peluso et al. (2012)	1	0.8	0.6	0.2	1	1	0.4	0.4
Baker et al. (2021)	1	0.8	0.8	0.8	0	0	0.6	0.6
Diaz et al. (2023)	1	0.8	0.8	0.8	0	0	0.8	0.6
Schiepek et al. (2016)	0.6	0.6	0.4	0	0.8	0.8	0	0
Westermann & Banisch (2025)	0.6	0.4	0.4	0	0.8	0.8	0	0
Tschacher, Haken, and Kyselo (2015)	0.8	0.8	0.6	0.6	1	1	0	0
Saint-Georges et al. (2011)	0.8	0.8	0.5	0.6	0	0	0.2	0.2
Varlet et al. (2012)	1	0.8	0.8	0.6	0	0	0.2	0
Schreiber et al. (2020)	0.8	1	0.8	1	0	0	0.4	0.4
Tschacher & Haken (2019)	1	1	0.8	0.8	1	1	0.8	0.8
Hale and Aarts (2023)	0.6	0.8	0.8	0.6	0	0	0	0
Paz et al. (2021)	0.8	1	0.8	1	0	0	0.2	0

Note. Yellow denotes models that scored highly only on empirical validity (because they primarily data-driven). Blue denotes models that scored highly on mainly theoretical validity (because they were only based on simulations). Finally, green denotes models that scored relatively highly on all validity types (because they were created in a theory-driven manner, generated specific hypotheses based on simulations, and examined in empirical investigations).

Table S9. Validity assessment for reinforcement learning models.

	EMPIRICAL VALIDITY				THEORETICAL VALIDITY		GENERATIVE VALIDITY	
	Model aims to explain a real-world phenomenon	The target condition is identifiable within the model	The comparator is identifiable within the model	The model explains target condition vis-à-vis comparator	Model manipulation yields expected outcomes	Simulated interventions match real interventions	Data-generating process is specific and plausible	Model architecture is homologous to the target mechanism
Will et al. (2017)	1	0.8	0.6	0.2	0	0	0.4	0.2
Will et al. (2020)	1	0.8	0.6	0.2	0	0	0.4	0.2
Contreras-Huerta et al. (2022)	1	0.8	0.8	0.5	0	0	0.4	0.2
Lenow, Cisler, & Bush (2018)	1	0.8	0.8	0.5	0	0	0.2	0.2
Barnby et al. (2022)	0.8	0.8	0.8	0.8	0	0	0.8	0.8
Fineberg et al. (2018)	0.8	0.8	0.6	0.5	0	0	0.2	0.2
Shapiro-Thompson et al. (2023)	1	0.8	0.8	0.5	0	0	0.2	0.2
Story et al. (2023)	1	1	0.8	0.5	0	0	0.4	0.2
Safra, Chevallier, & Palminteri (2019)	0.8	1	0.8	0.5	0	0	0.4	0.2
Frey, Frank, and McCabe (2021)	1	0.8	0.8	0.5	0	0	0.4	0.2

Note. All reinforcement learning models were classified as scoring highly only on empirical validity, because they were purely employed in data-driven ways. (The study by Barnby et al. (2022) employed both RL and Bayesian inference, showing that the latter fits the data better than the former, implying that its RL model scores highly mainly on empirical validity whereas its Bayesian model scores highly on both empirical validity and generative validity.)

Table S10. Validity assessment for Bayesian models.

	EMPIRICAL VALIDITY				THEORETICAL VALIDITY		GENERATIVE VALIDITY	
	Model aims to explain a real-world phenomenon	The target condition is identifiable within the model	The comparator is identifiable within the model	The model explains target condition vis-à-vis comparator	Model manipulation yields expected outcomes	Simulated interventions match real interventions	Data-generating process is specific and plausible	Model architecture is homologous to the target mechanism
Yoshida et al. (2010)	1	0.8	0.6	0.8	1	0.8	0.8	0.8
d’Arc, Devaine, and Daunizeau (2020)	0.8	0.8	0.8	0.8	0	0	0.7	0.7
Thomas et al. (2022)	1	1	0.8	0.8	0	0	0.4	0.2
Barnby, Raihani, & Dayan (2022)	1	1	0.8	0.8	0	0	0.2	0.2
Na et al. (2022)	0.8	0.8	0.8	0.8	0	0	0.4	0.1
Barnby et al. (2020)	1	1	0.8	0.8	0	0	0.4	0.2
Alon et al. (2024)	0.8	0.8	0.2	0	0.8	0.8	0.8	0.8
Siegel et al. (2020)	0.8	0.8	0.8	1	0	0	0.2	0.2
Story et al. (2023)	1	1	0.8	1	1	1	0.8	0.8
Xiang et al. (2012)	1	1	0.8	0.8	0.6	0.6	1	0.8
Hula et al. (2018)	1	1	1	1	1	1	1	0.8
Xiao et al. (2023)	0.8	1	1	1	0	0	0.2	0.2
Zavlis et al. (2024)	0.8	0.8	0.4	0	0.8	0.8	0.8	0.8
Henco et al. (2020)	1	1	1	1	0	0	0.2	0.2
Lamba, Frank, and FeldmanHall (2020)	1	0.8	0.8	0.8	0	0	0.8	0.8
Prosser et al. (2018)	0.8	0.8	0.4	0	0.8	0.8	0.4	0.2
Cittern et al. (2018)	0.8	0.8	0.4	0	0.8	0.8	0.4	0.2
Constant et al. (2021)	0.8	0.8	0.4	0	0.8	0.8	0.4	0.2

Note. Yellow denotes models that scored highly only on empirical validity (because they primarily data-driven). Blue denotes models that scored highly on mainly theoretical validity (because they were only based on simulations). Turquoise denotes models that scored highly on both theoretical and generative validity (because they were based on simulations that made more specific data-generating predictions based on prior empirical investigations). Finally, green denotes models that scored relatively highly on all validity types (because they were created in a theory-driven manner, generated specific hypotheses based on simulations, and examined in empirical investigations).

References

1. 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
2. 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
3. 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
4. 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
5. 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
6. 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.
7. Westermann S, Banisch S. A Formal Model of Affiliative Interpersonality. *Clinical Psychological Science*. 2025;13(1):43-68. doi:10.1177/21677026241229663
8. 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
9. 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
10. 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
11. 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
12. 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

13. 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/fpsy.2023.980739
14. Nour MM, Dahoun T, Schwartenbeck P, et al. Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *Proc Natl Acad Sci USA*. 2018;115(43). doi:10.1073/pnas.1809298115
15. 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
16. 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
17. 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
18. 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
19. 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
20. 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
21. 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
22. Lamba A, Frank MJ, FeldmanHall O. Anxiety Impedes Adaptive Social Learning Under Uncertainty. *Psychol Sci*. 2020;31(5):592-603. doi:10.1177/0956797620910993
23. 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
24. 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

25. 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
26. 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
27. Alon N, Schulz L, Bell V, Moutoussis M, Dayan P, Barnby JM. (Mal)adaptive Mentalizing in the Cognitive Hierarchy, and Its Link to Paranoia. *Computational Psychiatry*. 2024;8(1):159-177. doi:10.5334/cpsy.117
28. Zavlis O, Fonagy P, Moutoussis M, Story GW. A generative model of personality disorder as a relational disorder. Published online 2024. Accessed December 29, 2024. <https://osf.io/wh6na/download>
29. 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
30. Story GW, Smith R, ..., Dolan RJ. A social inference model of idealization and devaluation. *Psychol Review*. Published online 2023. doi:10.1037/rev0000430
31. 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
32. 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