Transdiagnostic factors differentially shape choices and reaction times in social evaluative learning

Aamir Sohail*[a,b,c], Aleksandra Magda^[a], Muna Said^[a], Laura Phillips^[a], Yifei Lu[a], Georgina Reynolds[a], Alice Comer-Fenn[a], Jessie Russell[a], Huafeng Lu[d], Tally Sims[a], Anne C. Saulin[a,b], Lei Zhang*[a,b,c,e]

[a]School of Psychology, University of Birmingham, Birmingham, B15 2TT, UK
[b]Centre for Human Brain Health, University of Birmingham, Birmingham, B15 2TT, UK
[c]Institute for Mental Health, University of Birmingham, Birmingham, B15 2TT, UK
[d]School of Architecture, Built Environment, Computing and Engineering, Birmingham City University, Birmingham, B4 7BD, UK
[e]Centre for Developmental Science, University of Birmingham, Birmingham, B15 2TT, UK

ORCID ID
Aamir Sohail: 0009-0000-6584-4579
Anne C. Saulin: 0000-0002-7200-468X
Lei Zhang: 0000-0002-9586-595X

*Correspondence should be addressed to:
Aamir Sohail: axs2210@bham.ac.uk
Lei Zhang: l.zhang.13@bham.ac.uk
Centre for Human Brain Health, University of Birmingham, Birmingham, B15 2TT, UK

^These authors contributed equally to this work.

Abstract

Background

Learning from feedback provided by others is important for navigating complex social environments. Aberrant processing of such social feedback is implicated in the psychopathology of several mental health disorders including social anxiety and depression. Recent psychiatric research increasingly adopts transdiagnostic approaches that examine mechanisms and constructs cutting across traditional diagnostic categories.

Methods

Here, we applied reinforcement learning models together with transdiagnostic measures of psychopathology to formally assess social evaluation learning. Human participants (n = 193) completed the Social Evaluation Learning Task, learning whether computer personas 'liked,' 'disliked,' or were 'neutral' toward them. Participants also completed six questionnaires assessing socially relevant psychiatric traits.

Results

Computational modeling revealed that a model with separate learning parameters for positive and negative feedback best accounted for the learning behavior. Exploratory factor analysis identified three transdiagnostic factors. A factor reflecting social avoidance predicted enhanced learning from negative feedback and a reduced positive learning bias. Socially avoidant individuals with heightened negative learning rates additionally showed faster reaction times specifically in positive social contexts. A second factor linking with emotional insensitivity for others predicted a lower choice accuracy in both positively and negatively valanced conditions and was associated with faster reaction times overall. No significant associations emerged for a third factor reflecting depressive and mood-related symptoms.

Conclusions

These results demonstrate how transdiagnostic trait dimensions shape social learning mechanisms through specific behavioral and computational processes.

Keywords

social evaluation; social learning; reinforcement learning; transdiagnostic; computational psychiatry

Introduction

Learning from social feedback is a vital cognitive process enabling individuals to adjust their behavior based on social cues, such as praise and criticism (Bauer, 1967). This mechanism is crucial for navigating complex social environments, maintaining relationships, and shaping one's self-concept (Boyd et al., 2011; Button et al., 2012; Elder et al., 2022; FeldmanHall & Chang, 2018; Kendal & Watson, 2023; Olsson et al., 2020; Peters et al., 2024; Schröder et al., 2025). In the form of social evaluative feedback, individuals update beliefs about the self and others (Behrens et al., 2008; Diaconescu et al., 2014) from external judgements or feedback which can differ in valence and expectancy (Peters et al., 2024). Appropriately learning from social feedback is integral to mental and physical well-being, with core self-appraisals being significantly influenced by external evaluations (Lundgren, 2004). Conversely, atypical social cognition is a feature of several mental health disorders (Gkika et al., 2018; Green & Leitman, 2008; Hoertnagl & Hofer, 2014; Patin & Hurlemann, 2015; Weightman et al., 2014) where individuals report biased perceptions of oneself which persist in response to social feedback (Gilboa-Schechtman et al., 2017; Hoffmann et al., 2024; Kirchner et al., 2025; Koban et al., 2017; Korn et al., 2012; Kube, 2023). Although previous studies have explored the computational basis of social evaluation learning and its relation to various mental health disorders, to date, none have specifically taken a transdiagnostic approach when doing so.

Social learning can be formally investigated using the framework of reinforcement learning (RL) (Sutton & Barto, 2018), a form of error-based learning driven by differences between actual and expected outcomes (i.e., prediction error). Specifically, individuals form a social prediction error (Joiner et al., 2017; Zhang et al., 2020; Zhang & Gläscher, 2020) - the difference between expected social feedback and the social feedback actually received. Application of the RL framework has revealed distinct biases in social feedback learning in several psychiatric disorders. For example, individuals with social anxiety and borderline personality disorder are biased towards learning from negative social feedback (Korn et al., 2016; Müller-Pinzler et al., 2019; Zabag et al., 2022, 2024) and demonstrate reduced positivity regarding one's own attributes (Hoffmann et al., 2024; Hopkins et al., 2021; Korn et al., 2012, 2016). In social anxiety, this is specifically linked to the maintenance of a negative self-view and low self-esteem (Koban et al., 2017, 2023). Similarly, in contrast to healthy participants, depressed individuals do not update their beliefs more strongly following positive compared to negative social feedback (Hobbs et al., 2022; Hoffmann et al., 2024). This blunted learning

appears to reflect the domain-general effect of anhedonia which extends to the context of social interactions (Barkus & Badcock, 2019).

Psychiatric disorders may present differently within categorical diagnoses (Feczko et al., 2019), reflecting a complex behavioral profile which cannot be appropriately captured by a single dimension or questionnaire. Indeed, diagnostic manuals, despite being standard tools, have been critiqued for their categorical approach (Adriaens & De Block, 2013; Kendler, 2009; Kendler et al., 2011), leading to the development and advocacy of alternative approaches with broader, multi-faceted definitions of mental health (Borsboom, 2017; Insel et al., 2010; Kotov et al., 2017). Compared to categorical assessments, a transdiagnostic approach to psychopathology (Gillan et al., 2016; Insel et al., 2010; Robbins et al., 2012; Tanaka, 2024) favours a broader definition of mental illness, spanning normal variations in psychopathology in a general population sample. Focusing on dimensional traits that cut across multiple disorders, latent psychological and cognitive constructs or factors of mental health can be extracted that more accurately map onto candidate cognitive processes (Gillan & Seow, 2020). This approach can further be combined with the theory-driven computational modeling of behavior such as RL models (Sohail & Zhang, 2024), to infer the construct-specific computational processes that characterize a given factor (Wise et al., 2023). Combining transdiagnostic assessments of psychopathology with computational models have identified the specific cognitive processes underlying altered behavior in mental health disorders, including model-based planning (Gillan et al., 2016), metacognition (Rouault et al., 2018), reward processing (Suzuki et al., 2021) and uncertainty (Norbury et al., 2018).

In the context of social feedback learning, existing research has primarily taken a categorical approach often focusing on a single disorder or trait. However, a transdiagnostic approach can uncover latent factors across trait dimensions, allowing for shared cognitive components to be extracted. In the current study we adopted such a perspective, using a questionnaire battery in a nonclinical sample assessing diverse domains of psychopathology known to affect social behavior including autism (Chevallier et al., 2012), social anxiety (Hofmann, 2007), depression (Rygula et al., 2015; Tse & Bond, 2004), apathy (Le Heron et al., 2019), narcissism (Morf & Rhodewalt, 2001), and borderline personality disorder (Herpertz & Bertsch, 2014). Specifically, we used exploratory factor analysis (EFA), a data-driven approach yielding composite trait factors that can be related to behavioral measures and computational parameters. For our behavioral task, we used the Social Evaluation Learning Task (SELT) (Button et al., 2012, 2015), a well-established paradigm requiring participants to learn whether

a fictional computer persona liked or disliked them across three feedback conditions differing by valence.

In line with previous studies extracting latent factors from similar trait measures (Gillan et al., 2016; Oka et al., 2025), we anticipated transdiagnostic factors reflecting social engagement and mood to be extracted from our questionnaire battery. We also hypothesised these factors to differentially moderate social learning effects. Specifically, we expected for social engagement to negatively correlate with a negative learning bias, and for a depressive-mood factor to positively correlate with reduced learning from positive feedback. These hypotheses stem from previous studies reporting these biases for categorical assessments of social anxiety and depression (Button et al., 2012; Hoffmann et al., 2024; Hopkins et al., 2021). Using a within-subject design across feedback conditions and continuous trait measures, this study ultimately provides a nuanced look at how overlapping psychiatric traits relate to the ability to learn from social feedback.

Methods

Participants

A priori sample size estimation using G*Power 3 software (Faul et al., 2007) suggested a sample of 191 would be needed to detect a small effect size ($f = 0.20$) with 80% power and $\alpha = .05$ for a repeated-measures ANOVA. Participants were eligible for the study if they were 18-to-25 years old, proficient in English, and had no history of brain trauma, cognitive impairment, neurological conditions or psychological disorders. Participants were recruited using the University of Birmingham Research Participation Scheme (Sona Systems) and flyers on campus and were tested online using the online experiment platform Gorilla (www.gorilla.sc) (Anwyl-Irvine et al., 2020). Participants received one academic credit point credit for completing the study. Data collection comprised 198 participants, which after excluding five participants who failed attention checks, resulted in a final sample (Table 1) of 193 participants (female = 173; age = 19.4 $\pm$ 2.0). All participants provided informed consent prior to participation, and the study was approved by the University of Birmingham Research Ethics Committee (ERN_20-1897AP15).

Table 1

Participant demographics for the study

Demographic characteristics of the study participants, including group sizes (n) and percentages (%) for sex assigned at birth, gender identity, and ethnic group.

	n	*%*
Sex Assigned at birth		
Male	20	10.4
Female	173	89.6
Gender		
Male	18	9.3
Female	173	89.6
Transgender	1	0.5
Gender Variant	1	0.5
Ethnic Group		
White	88	45.6
Asian or Asian British	59	30.6
Black, Black British, Caribbean or African	23	11.9
Mixed or Multiple	13	6.7
Other	10	5.2
Total	193	

Measures

Social Evaluation Learning Task

Participants completed the Social Evaluation Learning Task (Button et al., 2012, 2015), a probabilistic learning paradigm involving social feedback. In our experiment, we included three self-referential test blocks, each with a computer profile (i.e., persona) that participants had to learn either liked, disliked, or was neutral towards the participants (**Figure 1**). The order of the three blocks was randomized and counterbalanced across participants, with each block consisting of 32 trials. In each trial, participants were presented with the sentence “I think you are…” and a different positive/negative word pair randomly selected from a list of 64 pairs (e.g., nervous vs. calm). The placement (i.e., lower left vs. upper right) of the two words on the screen was counterbalanced across trials, to avoid motor habituation. Each persona followed a unique feedback contingency rule: the positive word in the word pair was the correct choice 80% of the time for the ‘Like’ persona, 20% for the ‘Dislike’ persona, and 50% for the ’Neutral’ persona, whereas the negative word in the word pair was the correct choice 20% of the time for the ‘Like’ persona, 80% for the ‘Dislike’ persona, and 50% for the ’Neutral’ persona. On each trial, participants were instructed to choose which word in each pair described the persona’s opinion toward them, to which they received feedback according to the

contingencies for each persona. The feedback contingencies were not communicated to the participants; instead, participants needed to work these out using trial and error. At the end of each block, participants were asked to rate how much the persona liked them on a scale from 0 ('Dislike') to 100 ('Like').

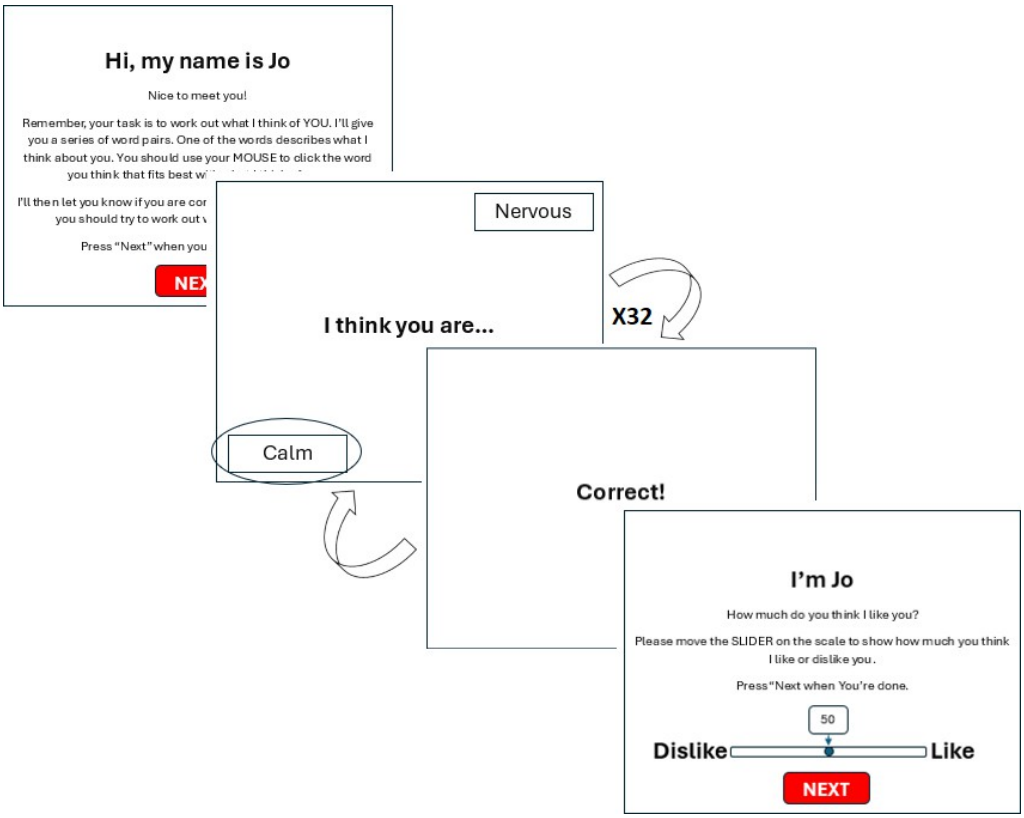


Fig 1. The Social Evaluation Learning Task. Participants were instructed to determine the computer persona's perspective of themselves. Participants played three blocks of 32 trials across three conditions, differing in the probability of positive/negative words being the correct choice. Feedback was provided after each trial. At the end of each block, participants were asked to provide a single rating on how much the persona disliked them using a Likert scale from 0 (Dislike) to 100 (Like).

Questionnaires

Immediately following the behavioral task, participants completed six self-report questionnaires to assess a wide range of psychopathology previously associated with deficits in social learning, namely the Autism Spectrum Quotient (AQ-10) (Allison et al., 2012), Social Phobia Inventory (SPIN) (Connor et al., 2000), Beck Depression Inventory (BDI-II) (Beck et al., 1996), Apathy Motivation Index (AMI) (Ang et al., 2017), Narcissistic Admiration and

Rivalry Questionnaire (NARQ) (Back et al., 2013), and Borderline Personality Disorder (BPD) checklist (First et al., 1997).

Data analysis approach

Data quality measures

An attention check was randomly assigned to two of the three blocks, where participants were asked to select a specific word: “Please select <word> to show you are paying attention...”. Participants who failed either of the two attention checks were removed from the analysis. Trials with a reaction time greater than 5000ms or less than 300ms were also excluded from all analyses.

Behavioral measures

Three primary dependent variables were selected as indicators of task behavior: choice accuracy, reaction time, and end-of-block persona rating. Choice accuracy was defined as selecting the word whose valence was associated with the feedback contingency; positive words for the ‘Like’ condition, negative words for the ‘Dislike’ condition. For the neutral condition, there is no actually correct response. We coded the 'Like' option as correct response in this condition ('Dislike' would reveal same results) to enable joint analysis of all conditions. Raw untransformed reaction times were used for repeated-measures ANOVA, as subject-level mean RTs were normally distributed across all conditions (Shapiro-Wilk tests: $W > 0.985$; skewness = 0.09-0.24). For trial-level mixed-effects models, we used log-transformed RT values. All mixed-effects models were fitted using the {lme4} package in R (Bates et al., 2015).

Computational modeling

A reinforcement learning framework was applied to quantify the computational processes underlying task behavior. Computational models included a simple Rescorla-Wagner model (Rescorla & Wagner, 1972), where for each trial (t), the value (V) of the chosen option was updated by the prediction error (PE) – the difference between the received outcome (O, coded as -1 and 1) and the value:

$$PE_t = O_t - V_t \quad (1)$$

Subsequently, the value of the next trial was adjusted by the prediction error, weighted by the learning rate (α, $0 < \alpha < 1$):

$$V_{t+1} = V_t + \alpha PE_t \quad (2)$$

Considering that participants may learn differently from positive versus negative social feedback (Zhang et al., 2020), a second RL model was tested with two distinct learning rates (α^+ and α^-) for positive and negative outcomes:

$$\begin{aligned} V_{t+1} &= V_t + \alpha^+ PE_t, \quad O_t > 0 \\ V_{t+1} &= V_t + \alpha^- PE_t, \quad O_t < 0 \end{aligned} \quad (3)$$

For both models, a softmax choice function was employed to compute the action probability based on the action values, where during each trial, the action probability of choosing the positive word over the negative word was defined as:

$$p(Positive)_t = \frac{1}{1 + e^{-\beta(V(Positive)_t - V(Negative)_t)}} \quad (4)$$

where the inverse temperature (β, $\beta > 0$) regulates the stochasticity of how value computation influences choice. Choices were coded according to word valence (positive word = 1, negative word = 2), with action values initialised at zero at the start of each condition block.

Model estimation was conducted under the hierarchical Bayesian framework using the {rstan} (Carpenter et al., 2017), and {hBayesDM} (Ahn et al., 2017) R packages. Posterior inference was conducted through Markov Chain Monte Carlo (MCMC) sampling, using four MCMC chains with 1000 post-warmup iterations per chain. R-hat values for all parameters were inspected to examine model convergence (Gelman & Rubin, 1992) and were below 1.02. Model comparison was conducted by computing the leave-one-out information criterion (LOOIC) using the {loo} package (Vehtari et al., 2017), a commonly applied criterion for Bayesian analysis balancing model fit and complexity. Lower LOOIC scores signified superior model performance, and the model with the lowest LOOIC was chosen for further analyses.

Exploratory factor analysis

An exploratory factor analysis was conducted to identify transdiagnostic factors across the six psychiatric questionnaires used. To account for mixed item types (binary and ordinal), polychoric correlations were computed using the {polycor} package. Factor analysis was performed using the 'fa' function from the {psych} package in R with maximum likelihood estimation and oblique (oblimin) rotation (Costello & Osborne, 2005). Eighty of the eighty-two

items (two removed due to zero-variance) across the six questionnaires were included in the analysis. A Cattell-Nelson-Gorsuch test was used through the {nFactors} package to determine the optimal number of factors. Factor loadings ≥ 0.3 were deemed interpretable. Factor scores were calculated using regression scoring weights and standardized for subsequent analyses. To uncover transdiagnostic influences on behavioral and computational outcomes, multiple regression analyses were conducted where transdiagnostic factor scores and task condition, and their interaction served as independent variables in models, to predict choice accuracy, reaction times, end-of-block ratings, and learning parameters as dependent variables.

Results

Behavioral

To test whether feedback valence affected learning accuracy, we examined choice accuracy across the three persona conditions. We initially tested whether accuracy across valanced conditions exceeded chance level (50%), where one-sample t-tests confirmed that accuracy exceeded chance (50%) in both valanced conditions (Like: $t_{(192)} = 28.94$, $p < 0.001$; Dislike: $t_{(192)} = 23.02$, $p < 0.001$) but not Neutral ($t_{(192)} = 1.24$, $p = 0.215$). (**Figure 2**). One-way repeated measures ANOVA further revealed a significant main effect of condition on choice accuracy ($F_{(2, 384)} = 208.41$, $p < .001$, $\eta^2p = 0.52$). Post-hoc comparisons with Tukey corrections showed that the accuracy was the highest in the 'Like' condition (M = 80.6 ± 14.7%), followed by those from the 'Dislike' (M = 75.9 ± 15.6%) and 'Neutral' (M = 51.6 ± 18.1%) conditions (**Figure 2A**; Like vs. Neutral: $t_{(192)} = 17.259$, $p < 0.001$, Dislike vs. Neutral: $t_{(192)} = 14.597$, $p < 0.001$, Like vs Dislike: $t_{(192)} = 3.982$, $p < 0.001$).

As a manipulation check, we examined end-of-block ratings of how much the feedback agent liked them across conditions. One-sample t-tests against the scale midpoint confirmed that Like ratings (M = 67.75 ± 18.37) were significantly above the midpoint ($t_{(192)} = 13.42$, $p < .001$), whilst Dislike (M = 39.83 ± 17.56; $t_{(192)} = –25.98$, $p < .001$), and Neutral ratings (M = 19.71 ± 16.20) were significantly below the midpoint ($t_{(192)} = –8.04$, $p < .001$). After applying a Greenhouse-Geisser correction ($\varepsilon = 0.924$) due to a sphericity violation ($W = 0.918$, $p < .001$), a one-way repeated measures ANOVA revealed a significant effect of condition on ratings ($F_{(1.85, 354.82)} = 331.93$, $p < .001$, $\eta^2p = 0.634$). Post-hoc Tukey-corrected pairwise comparisons demonstrated significant differences between all conditions (all $p < .001$) (**Figure 2B**).

Finally, to examine whether feedback valence influenced decision time, we tested for an effect of persona condition on reaction time (RT). A one-way repeated measures ANOVA showed a significant effect of condition on RT ($F_{(2,384)} = 4.18$, $p = 0.016$), with post-hoc Tukey-corrected comparisons showing participants responded significantly faster in the 'Dislike' (M =1604 ± 476ms) condition compared to 'Neutral' (M = 1698 ± 529ms) ($p = 0.009$), while other pairwise comparisons were not significant ('Like' condition: M = 1667 ± 466ms**, Figure 2C**).

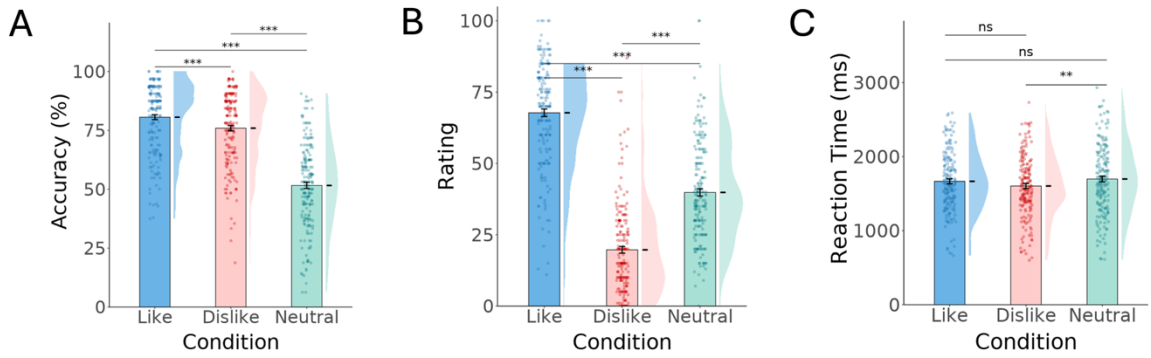


Fig 2. Behavioral results for the Social Evaluation Learning Task. Feedback valence demonstrated a significant effect upon all conditions with choice accuracy (**A**) and final ratings (**B**), and between Neutral and Dislike conditions for reaction time (**C**).

*** p < 0.01, *** p < 0.001*

Computational modeling

To determine whether participants exhibited asymmetric learning from positive versus negative feedback, we compared two computational models: a standard Rescorla-Wagner (RW) model with a single learning rate, and a dual-learning rate model with separate parameters for positive (α^+) and negative (α^-) feedback. Model comparison revealed the dual-learning model performed better than the simple RW model (ΔLOOIC = 271.09), confirming that participants processed positive and negative social feedback differently. Parameters from the winning model showed good-to-excellent recovery, with posterior predictive checks accurately replicating the key patterns in our behavioural data (see Supplementary Materials: Supplementary Figure 1 and Supplementary Figure 2).

After identifying the winning model, we tested whether learning rates varied as a function of feedback valence and persona condition. A learning rate valence (2: positive vs negative) by condition (3: Like, Dislike, Neutral) repeated measures ANOVA revealed a significant main effect for learning rate valence ($F_{(1,192)} = 545.63$, $p < 0.001$) and condition ($F_{(1.34, 256.33)} = 71.05$, $p < 0.001$) (**Figure 3A**). There was a significant interaction ($F_{(1.35, 260.07)} = 265.38$, $p < 0.001$),

indicating that the relationship between feedback valence and learning varied across conditions. Post-hoc pairwise comparisons using Tukey's adjustment showed the learning rate for positive feedback was significantly higher in the 'Like' than 'Neutral' ($t_{(192)} = 34.38$, $p < 0.001$) and 'Dislike' ($t_{(192)} = 24.35$, $p < 0.001$) conditions, and in 'Dislike' than 'Neutral' ($t_{(192)} = 25.29$, $p < 0.001$). For negative learning rates, 'Neutral' and 'Dislike' conditions did not differ ($t_{(192)} = 0.01$, $p = 0.999$), but both differed significantly from 'Like' (Neutral vs Like: $t_{(192)} = 5.74$, $p < 0.001$; Like vs Dislike: $t_{(192)} = 4.51$, $p < 0.001$), with the 'Like' condition showing higher negative learning rates.

To examine whether individual differences in learning rates predicted behavioral performance, we tested correlations between computational parameters and choice accuracy across conditions. Both the positive learning rate (α^+; Kendall's $\tau = 0.249$, $p < 0.001$) and the difference between the positive and negative learning rates - the positive learning bias - ($\alpha^+ - \alpha^-$; Kendall's $\tau = 0.150$, $p = 0.014$) were positively associated with accuracy in the 'Like' condition after Bonferroni correction (**Figure 3B-C**). However, no such relationship was observed for the negative learning rate (all $\tau = 0.249$, $p > 0.36$). These results highlight a direct link between the capacity to learn from positive social feedback and its recognition in social contexts.

As a follow-up analysis, we computed the mean trial-averaged learned value of the positive-word option from the winning model ($V^{positive}$) and participants' end-of-block ratings of how much the feedback agent liked them (0–100), for each feedback condition (See Supplementary Materials). This way, we can link the learning signal in the decisions task to the end-of-block valence rating. Results showed that mean $V^{positive}$ differed substantially across conditions in the expected direction: Like ($M = 0.741 \pm 0.184$), Neutral ($M = 0.184 \pm 0.337$), and Dislike ($M = -0.145 \pm 0.288$). Kendall's τ correlations between mean $V^{positive}$ and explicit ratings were significant in the Neutral ($\tau = 0.168$, $z = 3.27$, $p = .001$, $r = 0.26$) and Like ($\tau = 0.225$, $z = 4.38$, $p < .001$, $r = 0.34$) conditions, but not in the Dislike condition ($\tau = 0.064$, $z = 1.30$, $p = .194$, $r = 0.10$) **(Figure 3D)**. These results provide an important validation check for the winning model, highlighting a strong relationship between the model-derived learning signal and observed persona ratings.

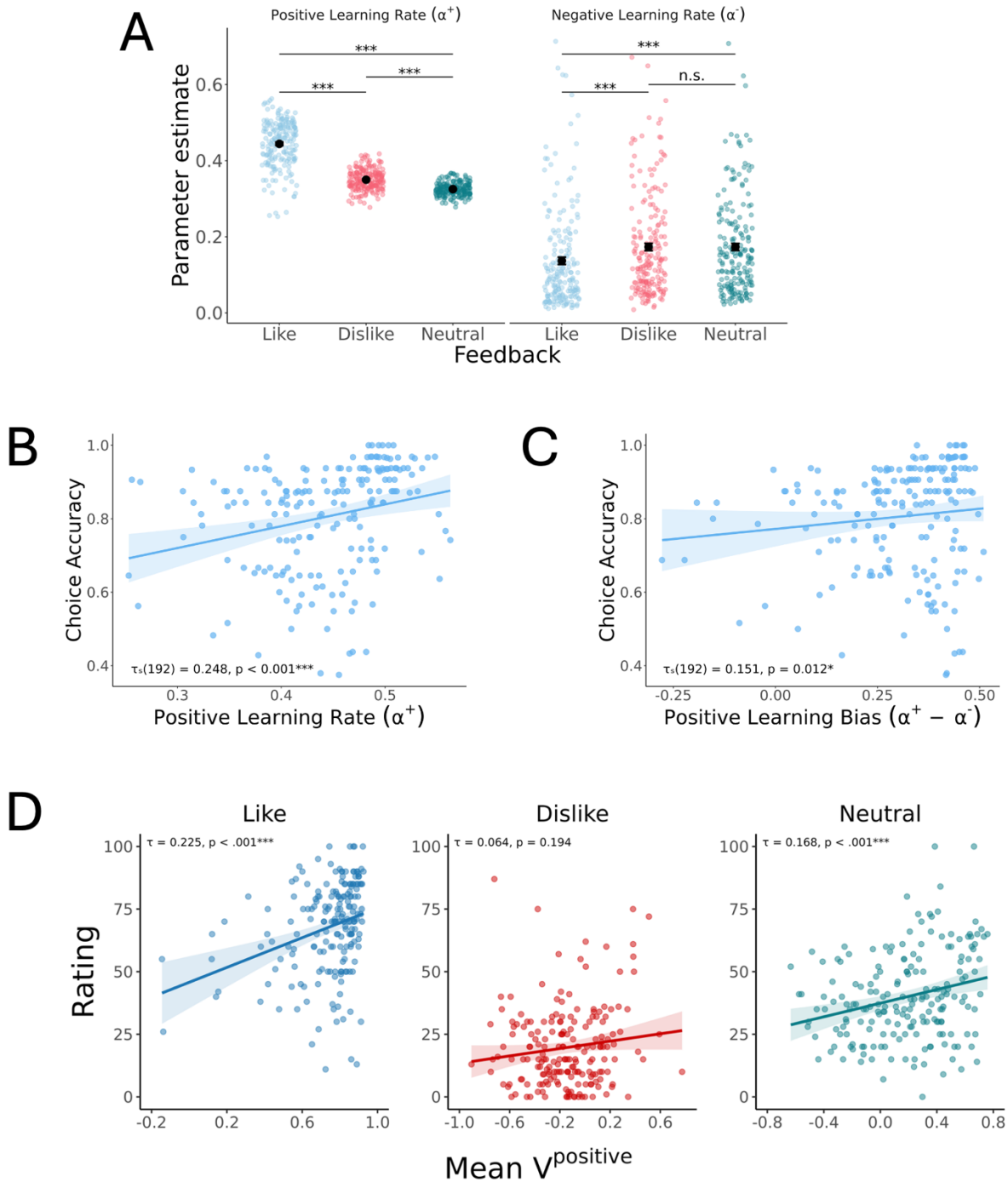


Fig 3. Parameter estimates from the winning model and associations with behavior. (A) Group-level parameter estimates for negative and positive learning rates from the winning model. Individual subject estimates shown as coloured points; black points indicate group means with error bars representing standard error of the mean. **(B)** Positive learning rate significantly correlated with choice accuracy in the Like condition (Kendall's $\tau = 0.248$, $p < 0.001$, Bonferroni-corrected across six planned comparisons). **(C)** Positive learning bias ($\alpha^+ - \alpha^-$) significantly correlated with choice accuracy in the Like condition (Kendall's $\tau = 0.151$, $p = 0.012$, Bonferroni-corrected). **(D)** Correlations between mean $V^{positive}$ and subject ratings for

each feedback condition. Each point represents a single participant. Shaded regions represent 95% confidence intervals.

Exploratory factor analysis and associations with social learning

To identify transdiagnostic dimensions of psychopathology that might relate to social feedback learning, we conducted an exploratory factor analysis (EFA) on the six questionnaires administered. Exploratory factor analysis using maximum likelihood estimation with oblique rotation, polychoric correlations to account for binary items, and the Cattell-Nelson-Gorsuch test revealed that a 3-factor solution best accounted for the questionnaire data, explaining 25.8% of the variance.

Based on the strongest individual item loadings (**Figure 4**), factors were labelled accordingly as ‘Social Avoidance’, with higher scores reflecting reduced social motivation and increased avoidance, ‘Emotional Insensitivity’, where higher scores indicate high apathy and difficulties evaluating and empathizing with others, and ‘Depressive Self-Identity’, where higher scores correspond to low mood, low self-esteem and negative self-identity. Specifically, for the Social Avoidance factor, the highest average loadings came from the SPIN questionnaire ($M = 0.57, \pm 0.11$), with significant contributions from AMI ($M = 0.19 \pm 0.23$) and AQ-10 ($M = 0.16 \pm 0.14$). The 'Emotional Insensitivity' factor showed only moderate loadings from AMI ($M = 0.20 \pm 0.26$) and AQ ($M = 0.17 \pm 0.22$), with no other questionnaires contributing meaningfully. Finally, Depressive Self-Identity was dominated by items from the BDI questionnaire ($M = 0.59 \pm 0.10$), followed by BPD ($M = 0.36 \pm 0.09$), with no other questionnaires reaching the 0.15 average loading threshold.

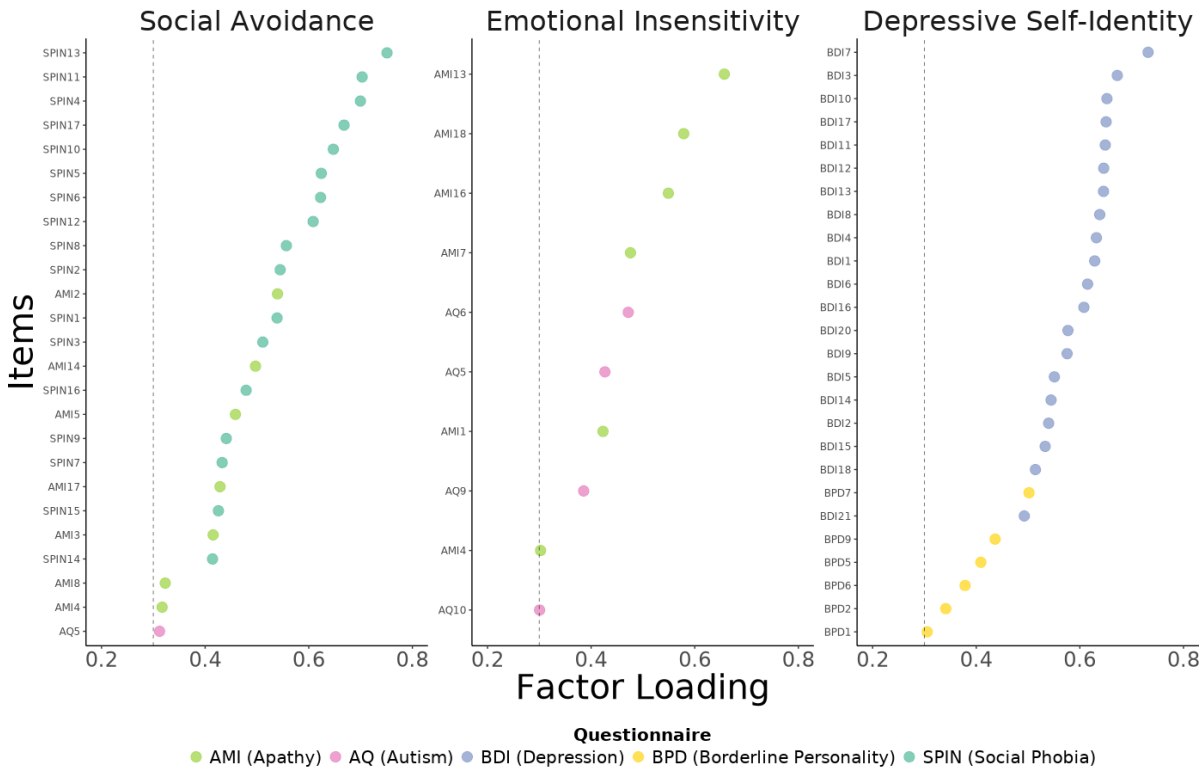


Fig 4. Factor scores (loadings > 0.3) for the three-factor solution. Factors were labelled 'Social Avoidance' featuring items measuring social phobia and motivation, 'Emotional Insensitivity' consisting of items reflecting apathy and autistic traits, and 'Depressive Self-Identity' primarily loading items assessing depression and personality disorder. No significant loadings (> 0.3) were observed for the Narcissistic Admiration and Rivalry Questionnaire (NARQ).

Transdiagnostic dimensions and condition-specific learning

To examine whether transdiagnostic factors predicted computational mechanisms underlying social feedback learning, we fit linear mixed-effects models with computational parameters from the winning model as dependent variables. Models included condition (Like/Dislike relative to Neutral), transdiagnostic factor scores (Social Avoidance, Emotional Insensitivity, Depressive Self-Identity), and their interactions as fixed factors, with random intercepts for subjects to account for the repeated measures structure (formula: DV ~ condition × (FA1 + FA2 + FA3) + (1|subID)) (**Figure 5**). Condition effects on learning parameters were robust, with the Like condition significantly increased positive learning rates ($\beta = 0.119$, $p < 0.001$) and decreased negative learning rates ($\beta = -0.036$, $p < 0.001$), while the Dislike condition significantly increased both positive learning rates ($\beta = 0.025$, $p < 0.001$) and positive learning bias ($\beta = 0.025$, $p < 0.001$). Additionally, the 'Social Avoidance' factor significantly predicted negative learning rates ($\beta = 0.025$, $p = 0.016$) and learning rate difference ($\beta = -0.024$, $p =$

0.029) in opposite directions. Specifically, individuals with higher scores on this 'Social Avoidance' factor learned more from negative social feedback, reflecting a negative learning bias. No other transdiagnostic factors significantly predicted computational learning parameters, suggesting that the negative learning bias observed is specific to traits related to social avoidance and low social motivation.

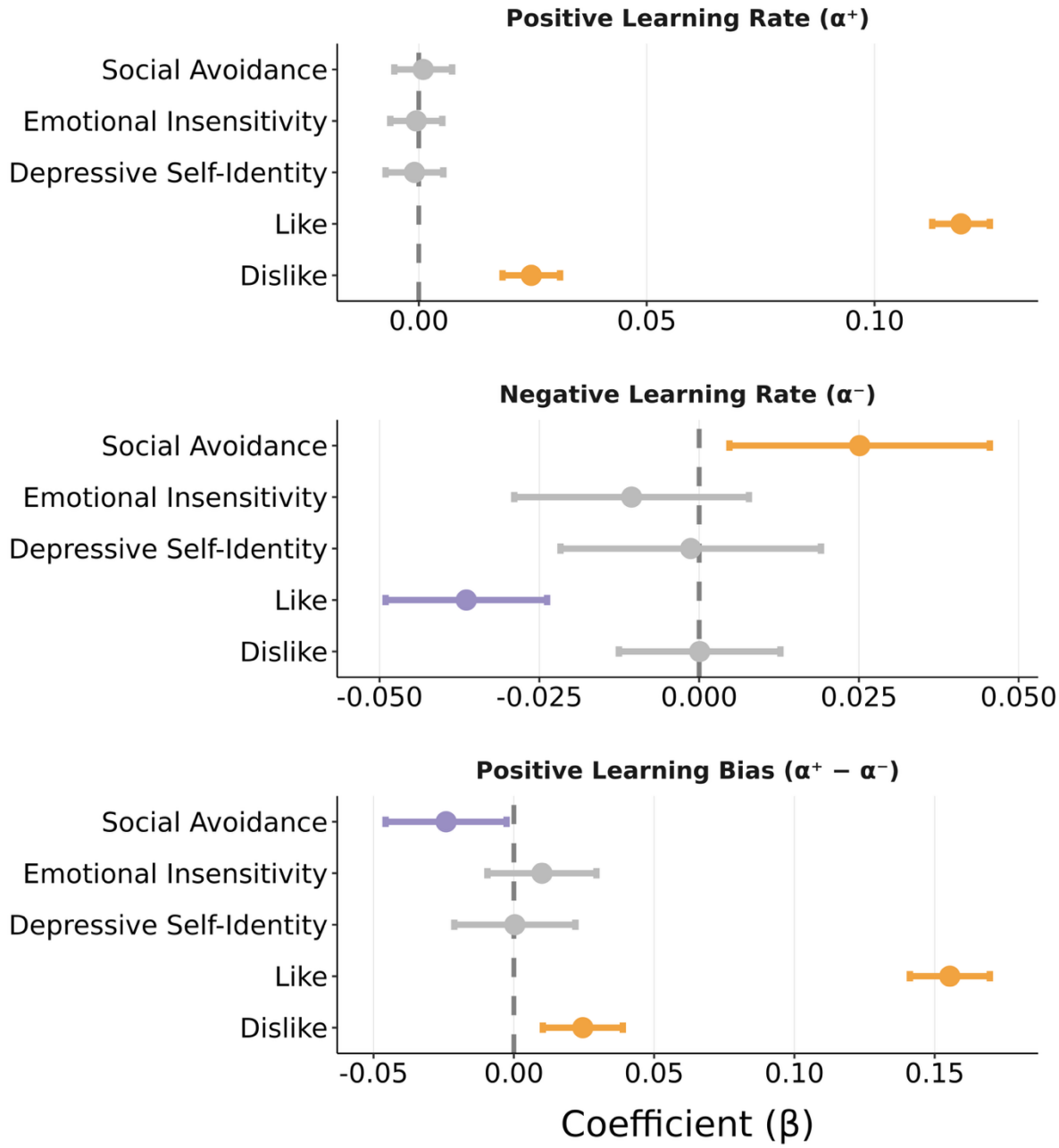


Fig 5. Main effect coefficients for linear regressions predicting computational parameters from the winning model. Top: Predictors of the positive learning rate (α^+) Middle: Predictors of the negative learning rate (α^-). Bottom: Predictors of the positive learning bias ($\alpha^+ - \alpha^-$). Yellow = significant positive effects ($p < 0.05$), Purple = significant negative effects ($p < 0.05$), grey = non-significant effects ($p \geq 0.05$). Error bars = 95% CI.

We next sought to understand whether the three transdiagnostic factors moderated the effect of feedback valence on behavioral performance. To test this, we fit a series of three-way mixed-effects regression models with transdiagnostic factors, learning rate parameters (α^+ and α^-), and valence conditions as predictors for task performance (**Figure 6**).

First, to examine effects on choice accuracy (**Figure 6A**), we fit a mixed-effects logistic regression model predicting choice accuracy, which revealed significant effects for both 'Like' ($\beta = 1.049$, $p < 0.001$) and 'Dislike' conditions ($\beta = 1.080$, $p < 0.001$), consistent with our repeated measures ANOVA analysis (cf. Figure 2A). Additionally, we observed a significant positive main effect of the positive learning rate parameter ($\beta = 4.935$, $p = 0.027$), indicating that individuals who learn more rapidly from positive feedback showed higher overall choice accuracy. A significant main effect of 'Emotional Insensitivity' was also observed ($\beta = 0.243$, $p = 0.041$), suggesting higher scores on this factor were associated with increased choice accuracy. However, significant two-way interactions between the 'Like' ($\beta = -0.480$, $p = 0.004$) and 'Dislike' ($\beta = -0.323$, $p = 0.020$) conditions with 'Emotional Insensitivity' revealed that higher scores on this factor attenuated the benefits in accuracy typically observed in value-congruent learning contexts. No other two-way interactions with condition or learning rate parameters were significant, and no three-way interactions reached significance (all $p > 0.05$).

Next, we fit linear mixed-effects models to determine whether transdiagnostic factors and learning rate parameters influenced decision time (**Figure 6B**). This revealed a significant main effect of the 'Dislike' condition ($\beta = -0.147$, $p = 0.014$), indicating faster responses. A significant main effect of the negative learning rate parameter was also observed ($\beta = 0.510$, $p = 0.006$), indicating that individuals who learned more rapidly from negative feedback showed slower overall response times. A significant two-way interaction between Social Avoidance and the negative learning rate parameter ($\beta = -0.436$, $p = 0.043$) linked higher Social Avoidance scores with a reduced relationship between negative learning rate and response time, whilst a significant two-way interaction between the 'Like' condition and 'Emotional Insensitivity' ($\beta = -0.166$, $p = 0.037$) indicated that higher scores on this factor were associated with faster responses specifically in the 'Like' condition. Furthermore, a significant three-way interaction between the 'Like' condition, Social Avoidance, and the negative learning rate parameter ($\beta = 0.567$, $p = 0.009$) revealed that individuals with higher Social Avoidance scores and higher negative learning rates exhibited slower response times in the 'Like' condition.

Similarly, linear mixed-effects models predicting participants' end-of-block ratings (**Figure 6C**) revealed significant main effects of task condition for both 'Like' ($\beta = 14.54$, $p < 0.001$) and 'Dislike' ($\beta = -9.01$, $p = 0.001$) personas, consistent with our manipulation of social feedback valence. No significant main effects of transdiagnostic factors or learning rate parameters were observed, and no two-way or three-way interactions between factors, learning rate parameters, and conditions reached significance (all $p > 0.05$).

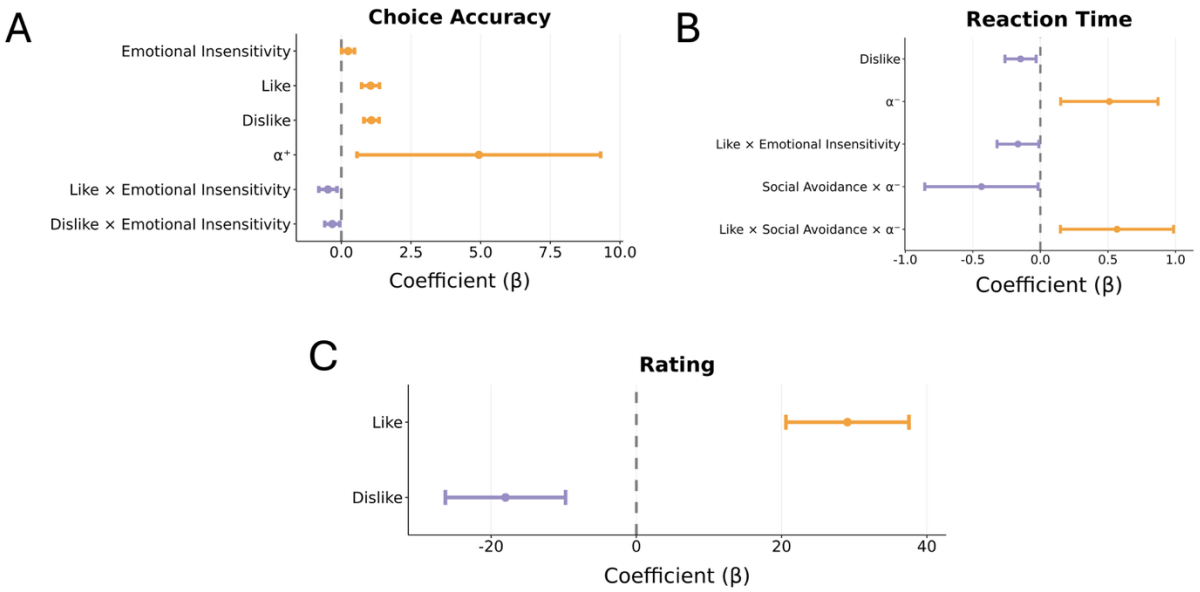


Fig 6. Significant three-way mixed-effects model coefficients. (A) Beta coefficients for choice accuracy logistic regression showing condition main effects and interactions. (B) Beta coefficients for reaction time showing condition main effects and interactions. (C) Beta coefficients for ratings showing only condition main effects. Yellow = significant positive effects ($p < 0.05$), Purple = significant negative effects ($p < 0.05$), grey = non-significant effects ($p \geq 0.05$). Error bars = 95% CI.

Discussion

The current study investigated how behavioral traits affect how individuals learn from social feedback utilizing a transdiagnostic approach and computational modeling under the reinforcement learning framework. Model comparison revealed that a modified Rescorla-Wagner model with separate learning rates for positive and negative feedback fit the data better than a standard model, indicating asymmetric learning processes. This aligns with previous findings that reinforcement learning involves differential updating from positive versus negative (social) outcomes (Behrens et al., 2008; Zhang et al., 2020). These learning

rates were found to vary by task condition, with the learning rate from positive feedback highest in the ‘Like’ condition. Individual differences in positive learning rates and positive learning bias were also both associated with higher choice accuracy in the 'Like' condition. This suggests that individuals who learn more strongly from positive feedback make more accurate evaluative judgements in socially positive contexts.

Our factor analysis extracted three robust transdiagnostic dimensions: ‘Social Avoidance’, ‘Emotional Insensitivity’ and ‘Depressive Self-Identity’. These dimensions partly overlap with previous studies extracting transdiagnostic factors of psychopathology from general population samples (Gillan et al., 2016; Hoven et al., 2023; Oka et al., 2025). On-the-other-hand, the presence of an ‘Emotional Insensitivity’ factor not reported previously reflects the current study’s specific focus towards assessing social evaluative learning. Indeed, this factor loaded items from the Autism Spectrum Quotient (AQ-10) and Apathy Motivation Index (AMI), questionnaires not included together in previous studies. The extraction of this factor highlights the ability to understand and interpret the emotions of others as an important transdiagnostic dimension of social learning.

Our regression analyses found significant associations between extracted transdiagnostic dimensions and task-related assessments of social evaluation learning. Specifically, we show that a latent Social Avoidance dimension, cutting across social anxiety and related traits, is selectively associated with a negative learning bias and slower response times towards positive social feedback. This finding extends previous research using categorical assessments of social anxiety and borderline personality disorder (BPD) demonstrating similar learning biases using reinforcement learning models (Hoffmann et al., 2024; Koban et al., 2017, 2023; Korn et al., 2016) and increased deliberation in response to positive social information (Hunter et al., 2022; Van Der Molen et al., 2014). The Social Avoidance factor strongly loaded items from the Social Phobia Inventory (SPIN) and those from the ‘Social Motivation’ subscale of the Apathy Motivation Index (AMI). Lowered social motivation - the interest and desire to engage in social interactions (Chevallier et al., 2012; Geen, 1991; Kanterman & Shamay-Tsoory, 2025) - is commonly reported among individuals with social anxiety (Abplanalp et al., 2022; Alden & Taylor, 2010; Blay et al., 2021; Geen, 1991; Yang et al., 2024), and co-morbidly presents in autism spectrum disorder (ASD) (Bagg et al., 2024; Briot et al., 2020; Spain et al., 2018), schizophrenia (Bershad & de Wit, 2023; Catalano & Green, 2023; Hajdúk et al., 2024) and attention deficit hyperactivity disorder (ADHD) (Martin et al., 2024). It is proposed that lowered social motivation distinctly arises from changes to

cognitive processes associated with effort-based decision-making, which lead to individuals underestimating the benefits and overestimating the effort costs associated with social interactions (Catalano & Green, 2023; Kanterman & Shamay-Tsoory, 2025). Recent accounts have identified the motivation to socially engage as a key transdiagnostic measure given its comorbid presentation across multiple disorders (Barkus & Badcock, 2019) that may have distinct clinical and neurobiological correlates and respond differently to treatments (Chetcuti et al., 2026). By extracting this specific measure and revealing an association with learning parameters, we highlight a potential target candidate for social motivation.

A second transdiagnostic factor, Emotional Insensitivity, reflecting lowered empathy and problems inferring the emotions of others, attenuated performance in both Like and Dislike conditions. As empathy requires accurate emotional identification (Bird & Viding, 2014; Shamay-Tsoory & Hertz, 2022), individuals with reduced empathy are hypothesized to be less capable of discerning changes in other's emotions (Coll et al., 2017). Indeed, reduced empathy and emotion recognition have been linked to poorer performance on complex social cognitive tasks through impaired interpretation of others' affective states (Gonzalez-Gadea et al., 2014). This is commonly observed among individuals with autism spectrum disorder (ASD) who show difficulties in social interactions (Bolis & Schilbach, 2017), due to challenges with understanding the intentions of others (Rosenthal et al., 2019). In the current study, measures from the Autism Quotient (AQ-10) and 'Emotional Sensitivity' subscale of the Apathy Motivation Index (AMI) loaded strongly onto the Emotional Insensitivity factor. Apathy is a multi-dimensional construct, consisting of executive, emotional, and initiation domains, each supported by distinct cognitive and neural mechanisms (Dickson & Husain, 2022). Most relevant to social behaviour is emotional apathy, characterised by emotional blunting, reduced empathy, and altered social interactions. Emotional apathy is robustly associated with prosocial behaviour across different contexts (Contreras-Huerta et al., 2022), with higher levels predicting a lowered tendency to help out others (Lockwood et al., 2017a). Yet, few studies have specifically investigated its role with social reward learning. In a study of dementia patients, those with high emotional apathy we found to demonstrate impaired learning towards social rewards (Wong et al., 2023). Our results further show that difficulties with interpreting the emotions of others directly impairs the ability to accurately evaluate positive and negative social feedback about the self. Therefore, emotional sensitivity is an important trait for correctly identifying and responding to the social preferences of others. Recent accounts also highlight individual differences in empathy and apathy, where affective empathy and emotional motivation are underpinned by the same latent factor (Lockwood et

al., 2017b). Our transdiagnostic approach was able to extract these socially-relevant measures from the AMI and AQ-10 that may be obscured in categorical assessments or questionnaires. We subsequently advocate for similar dimensional measures to be used when assessing social behaviour and cognition.

The present study has some limitations. Importantly, our sample is mainly comprised of females and consisted exclusively of university students with a mean age of approximately 19. This may explain the lack of any significant results with the Depressive Self-Identity factor, which strongly loaded items assessing depressive symptoms centred on guilt, lack of satisfaction and suicidal ideation. The reported Beck Depression Inventory (BDI) scores in our sample ($M = 12.52 \pm 9.34$) primarily reflect minimal (0-13) or mild (14-19) depression (Beck et al., 1996). However, depressive symptomatology may only emerge with more severe cases than those present in our non-clinical student population. This aligns with dimensional perspectives suggesting that certain cognitive-affective impairments may only manifest beyond a clinical threshold (Insel et al., 2010). Future studies should therefore aim to capture a wider range of depressive symptoms across a more diverse sample.

Taken together, combining a transdiagnostic measure of behavior with computational modelling, our study identifies key behavioral signatures of biased social evaluation learning. Computational modelling revealed that participants exhibited asymmetric learning from positive versus negative social feedback, with learning rates varying by feedback valence. A transdiagnostic factor reflecting social anxiety and motivational deficits predicted enhanced learning from negative feedback and a reduced positive learning bias, demonstrating biased social learning towards negatively salient information. An 'Emotional Insensitivity' factor capturing difficulties in empathy and mentalizing was associated with reduced choice accuracy in valanced conditions and faster reaction times in response to positive social feedback, suggesting reduced deliberative processing. Together, these findings demonstrate how transdiagnostic traits influence social evaluative learning, identifying candidate computational processes that can be targeted in future studies.

CRediT authorship contribution statement

A.S. – Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

A.M. – Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing

M.S. – Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing

L.P. – Methodology, Validation, Formal analysis, Investigation, Writing - Review & Editing

Y.L. – Validation, Formal analysis, Investigation, Writing - Review & Editing

G.R. – Validation, Formal analysis, Investigation, Writing - Review & Editing

A.C.-F. – Validation, Formal analysis, Investigation, Writing - Review & Editing

J.R. – Validation, Formal analysis, Investigation, Writing - Review & Editing

H.L. – Validation, Formal analysis, Investigation, Writing - Review & Editing

T.S. – Validation, Formal analysis, Investigation, Writing - Review & Editing

A.C.S. – Validation, Formal analysis, Writing - Review & Editing, Visualization, Supervision

L.Z. – Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data Curation, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition

Ethical standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Data availability statement

All code and data for data analysis and figure generation are publicly available at 10.5281/zenodo.19161237.

Acknowledgements

A.S was funded by an MRC AIM iCASE Studentship (MR/W007002/1). A.C.S. was funded by a Walter Benjamin Fellowship from the German Research Foundation (DFG/534697423). L.Z. was partially supported by the Wellcome Trust (228268/Z/23/Z) and the Royal Society (IES\R3\243253).

Competing interests

We declare no conflict of interest.

References

Abplanalp, S. J., Mote, J., Uhlman, A. C., Weizenbaum, E., Alvi, T., Tabak, B. A., & Fulford, D. (2022). Parsing social motivation: Development and validation of a self-report measure of social effort. *Journal of Mental Health*, *31*(3), 366–373. https://doi.org/10.1080/09638237.2021.1952948

Adriaens, P. R., & De Block, A. (2013). Why We Essentialize Mental Disorders. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *38*(2), 107–127. https://doi.org/10.1093/jmp/jht008

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 24–57. https://doi.org/10.1162/CPSY_a_00002

Alden, L. E., & Taylor, C. T. (2010). Interpersonal processes in social anxiety disorder. In *Interpersonal processes in the anxiety disorders: Implications for understanding psychopathology and treatment* (pp. 125–152). American Psychological Association. https://doi.org/10.1037/12084-005

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward Brief "Red Flags" for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

Ang, Y.-S., Lockwood, P., Apps, M. A. J., Muhammed, K., & Husain, M. (2017). Distinct Subtypes of Apathy Revealed by the Apathy Motivation Index. *PLOS ONE*, *12*(1), e0169938. https://doi.org/10.1371/journal.pone.0169938

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Back, M. D., Küfner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. A. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology*, *105*(6), 1013–1037. https://doi.org/10.1037/a0034431

Bagg, E., Pickard, H., Tan, M., Smith, T. J., Simonoff, E., Pickles, A., Carter Leno, V., & Bedford, R. (2024). Testing the social motivation theory of autism: The role of co-occurring anxiety. *Journal of Child Psychology and Psychiatry*, *65*(7), 899–909. https://doi.org/10.1111/jcpp.13925

Barkus, E., & Badcock, J. C. (2019). A Transdiagnostic Perspective on Social Anhedonia. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00216

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, R. A. (1967). Societal Feedback. *The ANNALS of the American Academy of Political and Social Science*, *373*(1), 180–192. https://doi.org/10.1177/000271626737300109

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. *Journal of Personality Assessment*, *67*(3), 588–597. https://doi.org/10.1207/s15327752jpa6703_13

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Bershad, A. K., & de Wit, H. (2023). Social Psychopharmacology: Novel Approaches to Treat Deficits in Social Motivation in Schizophrenia. *Schizophrenia Bulletin*, *49*(5), 1161–1173. https://doi.org/10.1093/schbul/sbad094

Bird, G., & Viding, E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience & Biobehavioral Reviews*, *47*, 520–532. https://doi.org/10.1016/j.neubiorev.2014.09.021

Blay, Y., Keshet, H., Friedman, L., & Gilboa-Schechtman, E. (2021). Interpersonal motivations in social anxiety: Weakened approach and intensified avoidance motivations for affiliation and social-rank. *Personality and Individual Differences*, *170*, 110449. https://doi.org/10.1016/j.paid.2020.110449

Bolis, D., & Schilbach, L. (2017). Beyond one Bayesian brain: Modeling intra- and inter-personal processes during social interaction: Commentary on “Mentalizing homeostasis: The social origins of interoceptive inference” by Fotopoulou & Tsakiris. *Neuropsychoanalysis*, *19*(1), 35–38. https://doi.org/10.1080/15294145.2017.1295215

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(supplement_2), 10918–10925. https://doi.org/10.1073/pnas.1100290108

Briot, K., Jean, F., Jouni, A., Geoffray, M.-M., Ly-Le Moal, M., Umbricht, D., Chatham, C., Murtagh, L., Delorme, R., Bouvard, M., Leboyer, M., & Amestoy, A. (2020). Social

Anxiety in Children and Adolescents With Autism Spectrum Disorders Contribute to Impairments in Social Communication and Social Motivation. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.00710

Button, K. S., Browning, M., Munafò, M. R., & Lewis, G. (2012). Social inference and social anxiety: Evidence of a fear-congruent self-referential learning bias. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(4), 1082–1087. https://doi.org/10.1016/j.jbtep.2012.05.004

Button, K. S., Kounali, D., Stapinski, L., Rapee, R. M., Lewis, G., & Munafò, M. R. (2015). Fear of Negative Evaluation Biases Social Evaluation Inference: Evidence from a Probabilistic Learning Task. *PLOS ONE*, *10*(4), e0119456. https://doi.org/10.1371/journal.pone.0119456

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1. https://doi.org/10.18637/jss.v076.i01

Catalano, L. T., & Green, M. F. (2023). Social Motivation in Schizophrenia: What's Effort Got to Do With It? *Schizophrenia Bulletin*, *49*(5), 1127–1137. https://doi.org/10.1093/schbul/sbad090

Chetcuti, L., Hardan, A. Y., Frazier, T. W., Loth, E., McPartland, J. C., Youngstrom, E. A., & Uljarevic, M. (2026). Advancing scientific understanding of the drive to socially engage: From broad constructs to transdiagnostic 'building blocks'. *Molecular Psychiatry*, *31*(1), 599–609. https://doi.org/10.1038/s41380-025-03185-9

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences*, *16*(4), 231–239. https://doi.org/10.1016/j.tics.2012.02.007

Coll, M.-P., Viding, E., Rütgen, M., Silani, G., Lamm, C., Catmur, C., & Bird, G. (2017). Are we really measuring empathy? Proposal for a new measurement framework. *Neuroscience & Biobehavioral Reviews*, *83*, 132–139. https://doi.org/10.1016/j.neubiorev.2017.10.009

Connor, K. M., Davidson, J. R. T., Churchill, L. E., Sherwood, A., Weisler, R. H., & Foa, E. (2000). Psychometric properties of the Social Phobia Inventory (SPIN): New self-rating scale. *British Journal of Psychiatry*, *176*(4), 379–386. https://doi.org/10.1192/bjp.176.4.379

Contreras-Huerta, L. S., Lockwood, P. L., Bird, G., Apps, M. A. J., & Crockett, M. J. (2022). Prosocial behavior is associated with transdiagnostic markers of affective sensitivity in multiple domains. *Emotion*, *22*(5), 820–835. https://doi.org/10.1037/emo0000813

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1). https://doi.org/10.7275/jyj1-4868

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014). Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLOS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Dickson, S. S., & Husain, M. (2022). Are there distinct dimensions of apathy? The argument for reappraisal. *Cortex*, *149*, 246–256. https://doi.org/10.1016/j.cortex.2022.01.001

Elder, J., Davis, T., & Hughes, B. L. (2022). Learning About the Self: Motives for Coherence and Positivity Constrain Learning From Self-Relevant Social Feedback. *Psychological Science*, *33*(4), 629–647. https://doi.org/10.1177/09567976211045934

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, *23*(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

FeldmanHall, O., & Chang, L. J. (2018). Social Learning. In *Goal-Directed Decision Making* (pp. 309–330). Elsevier. https://doi.org/10.1016/B978-0-12-812098-9.00014-0

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W., & Benjamin, L. S. (1997). *Structured Clinical Interview for DSM-IV Axis II Personality Disorders, (SCID-II)*. https://cir.nii.ac.jp/crid/1370004237531794560

Geen, R. G. (1991). Social motivation. *Annual Review of Psychology*, *42*, 377–399. https://doi.org/10.1146/annurev.ps.42.020191.002113

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4). https://doi.org/10.1214/ss/1177011136

Gilboa-Schechtman, E., Keshet, H., Livne, T., Berger, U., Zabag, R., Hermesh, H., & Marom, S. (2017). Explicit and implicit self-evaluations in social anxiety disorder. *Journal of Abnormal Psychology*, *126*(3), 285–290. https://doi.org/10.1037/abn0000261

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*, e11305. https://doi.org/10.7554/eLife.11305

Gillan, C. M., & Seow, T. X. F. (2020). Carving Out New Transdiagnostic Dimensions for Research in Mental Health. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(10), 932–934. https://doi.org/10.1016/j.bpsc.2020.04.013

Gkika, S., Wittkowski, A., & Wells, A. (2018). Social cognition and metacognition in social anxiety: A systematic review. *Clinical Psychology & Psychotherapy*, *25*(1), 10–30. https://doi.org/10.1002/cpp.2127

Gonzalez-Gadea, M. L., Herrera, E., Parra, M., Gomez Mendez, P., Baez, S., Manes, F., & Ibanez, A. (2014). Emotion recognition and cognitive empathy deficits in adolescent offenders revealed by context-sensitive tasks. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00850

Green, M. F., & Leitman, D. I. (2008). Social Cognition in Schizophrenia. *Schizophrenia Bulletin*, *34*(4), 670–672. https://doi.org/10.1093/schbul/sbn045

Hajdúk, M., Abplanalp, S. J., Jimenez, A. M., Fisher, M., Haut, K. M., Hooker, C. I., Lee, H., Ventura, J., Nahum, M., & Green, M. F. (2024). Linking social motivation, general motivation, and social cognition to interpersonal functioning in schizophrenia: Insights from exploratory graph analysis. *European Archives of Psychiatry and Clinical Neuroscience*, *274*(6), 1385–1393. https://doi.org/10.1007/s00406-023-01733-4

Herpertz, S. C., & Bertsch, K. (2014). The social-cognitive basis of personality disorders. *Current Opinion in Psychiatry*, *27*(1), 73. https://doi.org/10.1097/YCO.0000000000000026

Hobbs, C., Vozarova, P., Sabharwal, A., Shah, P., & Button, K. (2022). Is depression associated with reduced optimistic belief updating? *Royal Society Open Science*, *9*(2), 190814. https://doi.org/10.1098/rsos.190814

Hoertnagl, C. M., & Hofer, A. (2014). Social cognition in serious mental illness. *Current Opinion in Psychiatry*, *27*(3), 197. https://doi.org/10.1097/YCO.0000000000000055

Hoffmann, J. A., Hobbs, C., Moutoussis, M., & Button, K. S. (2024). Lack of optimistic bias during social evaluation learning reflects reduced positive self-beliefs in depression and social anxiety, but via distinct mechanisms. *Scientific Reports*, *14*(1), 22471. https://doi.org/10.1038/s41598-024-72749-6

Hofmann, S. G. (2007). Cognitive Factors that Maintain Social Anxiety Disorder: A Comprehensive Model and its Treatment Implications. *Cognitive Behaviour Therapy*, *36*(4), 193–209. https://doi.org/10.1080/16506070701421313

Hopkins, A. K., Dolan, R., Button, K. S., & Moutoussis, M. (2021). A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious

Individuals. *Computational Psychiatry (Cambridge, Mass.)*, *5*(1), 21–37. https://doi.org/10.5334/cpsy.57

Hoven, M., Luigjes, J., Denys, D., Rouault, M., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: A transdiagnostic approach. *Nature Mental Health*, *1*(5), 337–345. https://doi.org/10.1038/s44220-023-00062-8

Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M., & Daw, N. D. (2022). Increased and biased deliberation in social anxiety. *Nature Human Behaviour*, *6*(1), 146–154. https://doi.org/10.1038/s41562-021-01180-y

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2

Kanterman, A., & Shamay-Tsoory, S. (2025). From social effort to social behavior: An integrated neural model for social motivation. *Neuroscience & Biobehavioral Reviews*, *173*, 106170. https://doi.org/10.1016/j.neubiorev.2025.106170

Kendal, R. L., & Watson, R. (2023). Adaptive Social Learning: Social Learning Strategies and their Applications. In J. J. Tehrani, J. Kendal, & R. L. Kendal (Eds), *Oxford Handbook of Cultural Evolution* (1st edn, pp. 193–207). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198869252.013.14

Kendler, K. S. (2009). An historical framework for psychiatric nosology. *Psychological Medicine*, *39*(12), 1935–1941. https://doi.org/10.1017/S0033291709005753

Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(6), 1143–1150. https://doi.org/10.1017/S0033291710001844

Kirchner, L., Kube, T., Berg, M., Eckert, A.-L., Straube, B., Endres, D., & Rief, W. (2025). Social expectations in depression. *Nature Reviews Psychology*, *4*(1), 20–34. https://doi.org/10.1038/s44159-024-00386-x

Koban, L., Andrews-Hanna, J. R., Ives, L., Wager, T. D., & Arch, J. J. (2023). Brain mediators of biased social learning of self-perception in social anxiety disorder. *Translational Psychiatry*, *13*(1), Article 1. https://doi.org/10.1038/s41398-023-02587-z

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., Wager, T. D., & Arch, J. J. (2017). Social anxiety is characterized by biased learning

about performance and the self. *Emotion (Washington, D.C.)*, *17*(8), 1144–1155. https://doi.org/10.1037/emo0000296

Korn, C. W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine*, *46*(3), 575–587. https://doi.org/10.1017/S003329171500207X

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Kube, T. (2023). Biased belief updating in depression. *Clinical Psychology Review*, *103*, 102298. https://doi.org/10.1016/j.cpr.2023.102298

Le Heron, C., Holroyd, C. B., Salamone, J., & Husain, M. (2019). Brain mechanisms underlying apathy. *Journal of Neurology, Neurosurgery & Psychiatry*, *90*(3), 302–312. https://doi.org/10.1136/jnnp-2018-318265

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. J. (2017a). Prosocial apathy for helping others when effort is required. *Nature Human Behaviour*, *1*(7), 0131. https://doi.org/10.1038/s41562-017-0131

Lockwood, P. L., Ang, Y.-S., Husain, M., & Crockett, M. J. (2017b). Individual differences in empathy are associated with apathy-motivation. *Scientific Reports*, *7*(1), 17293. https://doi.org/10.1038/s41598-017-17415-w

Lundgren, D. C. (2004). Social Feedback and Self-Appraisals: Current Status of the Mead-Cooley Hypothesis. *Symbolic Interaction*, *27*(2), 267–286. https://doi.org/10.1525/si.2004.27.2.267

Martin, R., McKay, E., & Kirk, H. (2024). Lowered social motivation is associated with adolescent attention deficit hyperactivity disorder and social anxiety symptoms. *Clinical Child Psychology and Psychiatry*, *29*(1), 338–352. https://doi.org/10.1177/13591045231218475

Morf, C. C., & Rhodewalt, F. (2001). Unraveling the Paradoxes of Narcissism: A Dynamic Self-Regulatory Processing Model. *Psychological Inquiry*, *12*(4), 177–196. https://doi.org/10.1207/S15327965PLI1204_1

Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, *9*(1), 14416. https://doi.org/10.1038/s41598-019-50821-w

Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *eLife*, *7*, e34779. https://doi.org/10.7554/eLife.34779

Oka, T., Sasaki, A., & Kobayashi, N. (2025). A transdiagnostic dimensional approach to behavioral dysregulation: Examining self-reported reward and punishment sensitivity across psychopathology. *Journal of Affective Disorders*, *387*, 119493. https://doi.org/10.1016/j.jad.2025.119493

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), Article 4. https://doi.org/10.1038/s41583-020-0276-4

Patin, A., & Hurlemann, R. (2015). Social Cognition. In K. M. Kantak & J. G. Wettstein (Eds), *Cognitive Enhancement* (pp. 271–303). Springer International Publishing. https://doi.org/10.1007/978-3-319-16522-6_10

Peters, A., Helming, H., Bruchmann, M., Wiegandt, A., Straube, T., & Schindler, S. (2024). How and when social evaluative feedback is processed in the brain: A systematic review on ERP studies. *Cortex*, *173*, 187–207. https://doi.org/10.1016/j.cortex.2024.02.003

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning, Current Research and Theory*, *2*, 64–69.

Robbins, T. W., Gillan, C. M., Smith, D. G., Wit, S. de, & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences*, *16*(1), 81–91. https://doi.org/10.1016/j.tics.2011.11.009

Rosenthal, I. A., Hutcherson, C. A., Adolphs, R., & Stanley, D. A. (2019). Deconstructing Theory-of-Mind Impairment in High-Functioning Adults with Autism. *Current Biology*, *29*(3), 513-519.e6. https://doi.org/10.1016/j.cub.2018.12.039

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry, Translating Biology to Treatment in Schizophrenia*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Rygula, R., Clarke, H. F., Cardinal, R. N., Cockcroft, G. J., Xia, J., Dalley, J. W., Robbins, T. W., & Roberts, A. C. (2015). Role of Central Serotonin in Anticipation of Rewarding

and Punishing Outcomes: Effects of Selective Amygdala or Orbitofrontal 5-HT Depletion. *Cerebral Cortex*, *25*(9), 3064–3076. https://doi.org/10.1093/cercor/bhu102

Schröder, A., Czekalla, N., Mayer, A. V., Zhang, L., Stolz, D. S., Korn, C. W., Diekelmann, S., Luebber, F., Paulus, F. M., Müller-Pinzler, L., & Krach, S. (2025). Initial Expectations and Confidence Affect the Formation of Novel Self-Beliefs and Their Revision. *Open Mind*, *9*, 1576–1596. https://doi.org/10.1162/OPMI.a.36

Shamay-Tsoory, S. G., & Hertz, U. (2022). Adaptive Empathy: A Model for Learning Empathic Responses in Response to Feedback. *Perspectives on Psychological Science*, *17*(4), 1008–1023. https://doi.org/10.1177/17456916211031926

Sohail, A., & Zhang, L. (2024). Informing the treatment of social anxiety disorder with computational and neuroimaging data. *Psychoradiology*, *4*, kkae010. https://doi.org/10.1093/psyrad/kkae010

Spain, D., Sin, J., Linder, K. B., McMahon, J., & Happé, F. (2018). Social anxiety in autism spectrum disorder: A systematic review. *Research in Autism Spectrum Disorders*, *52*, 51–68. https://doi.org/10.1016/j.rasd.2018.04.007

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, 2nd edition*. The MIT Press.

Suzuki, S., Yamashita, Y., & Katahira, K. (2021). Psychiatric symptoms influence reward-seeking and loss-avoidance decision-making through common and distinct computational processes. *Psychiatry and Clinical Neurosciences*, *75*(9), 277–285. https://doi.org/10.1111/pcn.13279

Tanaka, M. (2024). Beyond the boundaries: Transitioning from categorical to dimensional paradigms in mental health diagnostics. *Advances in Clinical and Experimental Medicine*, *33*(12), 1295–1301. https://doi.org/10.17219/acem/197425

Tse, W. S., & Bond, A. J. (2004). The Impact of Depression on Social Skills: A Review. *The Journal of Nervous and Mental Disease*, *192*(4), 260. https://doi.org/10.1097/01.nmd.0000120884.60002.2b

Van Der Molen, M. J. W., Poppelaars, E. S., Van Hartingsveldt, C. T. A., Harrewijn, A., Gunther Moor, B., & Westenberg, P. M. (2014). Fear of negative evaluation modulates electrocortical and behavioral responses when anticipating social evaluative feedback. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00936

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Weightman, M. J., Air, T. M., & Baune, B. T. (2014). A Review of the Role of Social Cognition in Major Depressive Disorder. *Frontiers in Psychiatry*, *5*. https://doi.org/10.3389/fpsyt.2014.00179

Wise, T., Robinson, O. J., & Gillan, C. M. (2023). Identifying Transdiagnostic Mechanisms in Mental Health Using Computational Factor Modeling. *Biological Psychiatry, Emerging Topics in Computational Psychiatric Research*, *93*(8), 690–703. https://doi.org/10.1016/j.biopsych.2022.09.034

Wong, S., Wei, G., Husain, M., Hodges, J. R., Piguet, O., Irish, M., & Kumfor, F. (2023). Altered reward processing underpins emotional apathy in dementia. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(2), 354–370. https://doi.org/10.3758/s13415-022-01048-2

Yang, Y., Zhou, Y., Zhang, H., Kou, H., Zhao, J., Tian, J., & Guo, C. (2024). Social anxiety undermines prosocial behaviors when required effort. *International Journal of Clinical and Health Psychology*, *24*(4), 100533. https://doi.org/10.1016/j.ijchp.2024.100533

Zabag, R., Gilboa-Schechtman, E., & Levy-Gigi, E. (2022). Reacting to changing environment: Updating patterns in social anxiety. *Behaviour Research and Therapy*, *157*, 104159. https://doi.org/10.1016/j.brat.2022.104159

Zabag, R., Rinck, M., Becker, E., Gilboa-Schechtman, E., & Levy-Gigi, E. (2024). Although I know it: Social anxiety is associated with a deficit in positive updating even when the cost of avoidance is Obvious. *Journal of Psychiatric Research*, *169*, 279–283. https://doi.org/10.1016/j.jpsychires.2023.11.041

Zhang, L., & Gläscher, J. (2020). A brain network supporting social influences in human decision-making. *Science Advances*, *6*(34), eabb4159. https://doi.org/10.1126/sciadv.abb4159

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089

Supplementary material

Questionnaire battery

Autism Spectrum Quotient (AQ-10) (Allison et al., 2012), a 10-item screening instrument for autistic traits, where higher scores indicate greater social communication difficulties and rigidity. Items are rated as “Agree” or “Disagree,” and scored dichotomously (0 or 1), yielding a total score range of 0–10.

Social Phobia Inventory (SPIN) (Connor et al., 2000), a 17-item measure of social anxiety symptoms. Participants rate how much each statement (e.g., fear of embarrassment) applies to them, with higher scores indicating greater social anxiety. Each item is rated on a 5-point Likert scale (0 = “Not at all” to 4 = “Extremely”), producing a total score range of 0–68.

Beck Depression Inventory (BDI-II) (Beck et al., 1996), a 21-item inventory assessing depressive symptoms such as sadness, pessimism, and anhedonia. Higher scores reflect more severe depressive tendencies, and each item is rated from 0 to 3, leading to a total score range of 0–63.

Apathy Motivation Index (AMI) (Ang et al., 2017), an 18-item scale measuring apathy across behavioral and social domains. We used the total score as an index of general apathy (higher scores = greater apathy/lower motivation). Items are rated on a 5-point Likert scale (0 = “Strongly disagree” to 4 = “Strongly agree”), giving a total score range of 0–72.

Narcissistic Admiration and Rivalry Questionnaire (NARQ) (Back et al., 2013), a short 6-item version of a narcissism scale, with higher scores indicating stronger narcissistic tendencies. Items are rated on a 6-point scale (1 = “Disagree strongly” to 6 = “Agree strongly”), so total scores range from 6 to 36.

Borderline Personality Disorder (BPD) checklist (First et al., 1997), a 10-item yes/no screening checklist for BPD features, where higher scores indicate greater severity. One item with no variance in our sample was removed, so scores were based on the remaining 9 items, with a total score range of 0–9.

Winning model specification

The winning model implemented a hierarchical Rescorla-Wagner framework with asymmetric learning rates for reward (α^+) and punishment (α^-), alongside an inverse temperature parameter (τ) governing choice stochasticity. Parameters were estimated using Bayesian hierarchical modeling with non-centered parameterization and effect coding for within-subject condition contrasts.

The model defines the choice process by:

$$Choice_{(s,t+1)} \sim categorical\_logit(\tau_{(s,c)} \, . \, V_{(s,t)}) \quad (1)$$

where $V_{(s,t)}$ are action values updated according to the outcome valence:

$$\begin{aligned} V_{(t+1)} &= V_{(t)} + \alpha^+_{(s,c)} \, . \, (R - V_{(t)}) \quad if \; R > 0 \\ V_{(t+1)} &= V_{(t)} + \alpha^-_{(s,c)} \, . \, (R - V_{(t)}) \quad if \; R > 0 \end{aligned} \quad (2)$$

with hierarchical priors:

$\mu_{\alpha+}$, $\mu_{\alpha-}$, $\mu_\tau \sim$ Normal(0, 1)
$\sigma_{\alpha+}$, $\sigma_{\alpha-}$, $\sigma_\tau \sim$ Exponential(2)
$\varepsilon_{\alpha+}$, $\varepsilon_{\alpha-}$, $\varepsilon_\tau \sim$ Normal(0, 1)

Parameter recovery

After model fitting, we confirmed the identifiability of parameters through parameter recovery. We repeated the following procedure for 20 iterations. Denoting ϕ as a generic parameter, we iterated the following steps:

(a) We randomly drew one joint sample of group-level parameters from the joint posterior group-level distribution of the winning model, a Rescorla-Wagner model with separate learning rates for positive and negative outcomes. That is, a group-level mean (μ_ϕ) and a group-level standard deviation (σ_ϕ) of the parameter ϕ. We repeated this procedure for all group-level parameters of the winning model:

$$\mu_\varphi, \sigma_\varphi \sim p\left(\mu_\varphi, \sigma_\varphi \middle| D\right) \quad (3)$$

(b) Next, we simulated 193 synthetic participants whose parameters were sampled using the hierarchical structure implemented in the Stan model. Individual-level parameters ϕ_i were generated using non-centered parameterization in probit-transformed space, then transformed to natural parameter bounds via effect coding and the cumulative normal distribution function, preserving the model's hierarchical structure and parameter constraints. We repeated this procedure for all individual-level parameters of the winning model.

(c) Then, we used the winning model as a generative tool to simulate behavioral data for our social evaluation learning task, namely, to simulate choices for 96 trials per participant (32 trials per condition). Individuals' choices (denoted as $\tilde{D}_i$) were sampled from the sampling distribution conditional on individual-level parameters (ϕ_i) from the previous step (i.e., likelihood function):

$$D_i \sim p(D_i, \sigma_i) \quad (4)$$

(d) We fit the model to the simulated data ($\tilde{D}_i$) in the same way as we did for the real data obtaining parameter estimates (i.e., posterior distributions) at both the group-level (e.g., $\hat{\mu}_\phi$, $\hat{\sigma}_\phi$) and the individual level (e.g., $\hat{\phi}_i$).

(e) Finally, we compared whether the posterior distributions recovered the true data-generating parameters. At the group level, we assessed whether the true parameter values fell within the 95% highest density interval (HDI) of the recovered posterior distributions for each iteration, and report coverage rates (percentage of iterations where true value falls within 95% HDI) across the 20 iterations (Supplementary Figure 1A). At the individual level, we computed Spearman's rank correlations (ρ) between the true and recovered parameters for each parameter and condition combination within each iteration, then aggregated these correlations across iterations using Fisher's Z transformation to obtain mean correlation coefficients (Supplementary Figure 1B).

The parameter recovery analyses initially revealed differential recoverability across individual-level parameters, with positive learning rates showing lower correlations compared to negative learning rates. Examining the fitted parameters revealed extremely low between-subject variability for the positive learning rate (CV = 0.05-0.14), whereas in contrast, the negative learning rate showed high variability (CV = 0.74-0.98) (Supplementary Table 1).

Condition	α^+ mean	α^+ CV	α^- mean	α^- CV
Neutral	0.32 (0.17)	0.05	0.11 (0.89)	0.74
Like	0.44 (0.43)	0.14	0.04 (0.99)	0.98
Dislike	0.34 (0.20)	0.08	0.08 (0.70)	0.77

Supplementary Table 1. Parameter estimates for the winning model. Mean values are accompanied by the standard deviation (SD) in brackets. Mean values are group-level posterior means in natural space [0,1]. SD values are group-level posterior standard deviations in probit space. Coefficients of variation (CV) are calculated from individual-level posterior means in natural space to reflect between-subject variability.

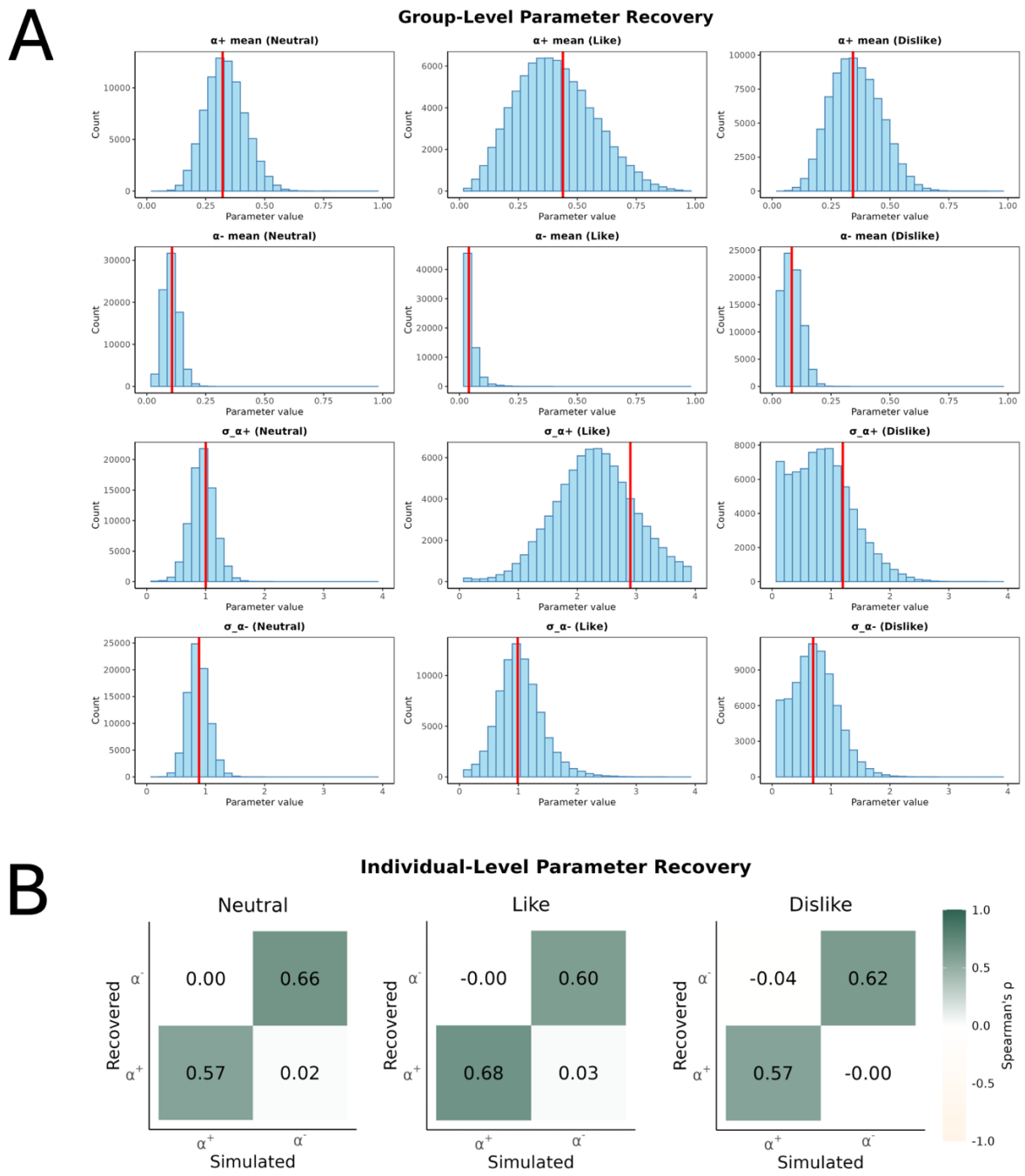
To test whether low variability of this parameter caused the observed differences with recovery, we re-ran parameter recovery after artificially increased the group-level standard deviation of α^+ while keeping all other parameters at empirical values. Specifically, we calibrated α^+ probit-space SDs to achieve CVs matching α^- levels (Neutral: SD = 1.00, CV = 0.74; Like: SD = 2.90, CV = 0.88; Dislike: SD = 1.20, CV = 0.78). Increasing α^+ variability subsequently improved its recovery across all three conditions (Neutral: r = 0.24 to r = 0.57; Like: r = 0.35 to r = 0.68; Dislike: r = 0.26 to r = 0.57). In contrast, α^- recovery remained stable (r = 0.60-0.66).

Posterior predictive checks

To validate that the winning model could generate behavioral patterns consistent with the observed data, we conducted posterior predictive checks using a generative simulation approach. We selected 4,000 posterior samples from the winning model's MCMC chains, where for each sample, we extracted the individual-level parameters (α^+, α^-, τ) for all 193 participants across all three conditions and used the model to simulate a complete synthetic dataset of 96 trials per participant (32 trials per condition), maintaining the same task structure and feedback contingencies as in the original experiment. We then analysed each simulated dataset using the same statistical procedures applied to the observed data. Specifically, we calculated mean choice accuracy for each condition and conducted repeated-measures ANOVA with post-hoc pairwise comparisons (Tukey-corrected). To evaluate model fit, we calculated Bayesian p-values as the proportion of predicted statistics more extreme than the corresponding observed statistic (two-tailed), with p-values < 0.05 indicating circumstances where the model fails to capture the observed data pattern.

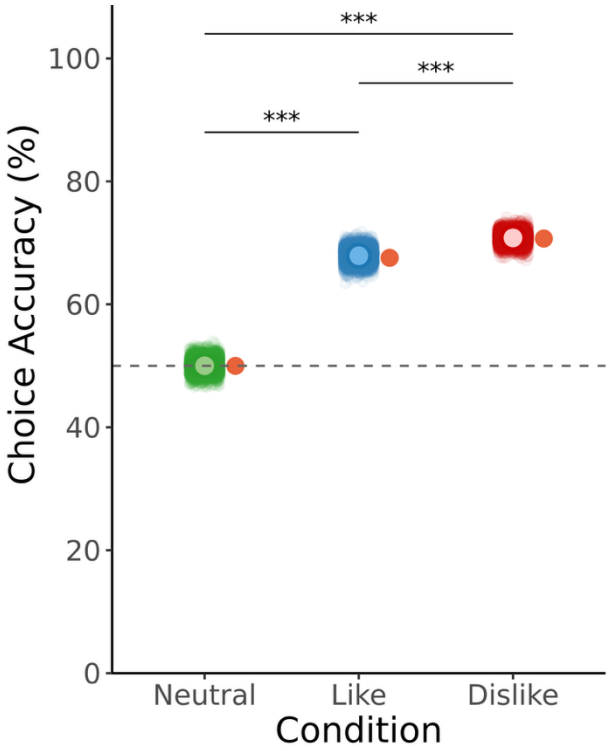
Posterior predictive checks (PPCs) on the empirical data showed deviations in the Like condition, where the model underpredicted accuracy. Model predictions accurately captured the Neutral (observed = 49.8%, predicted = 50.0%, 95% HDI: [48.8%, 51.3%]) and Dislike (observed = 76.2%, predicted = 76.2%, 95% HDI: [74.2%, 78.2%]), with observed values falling within the 95% prediction intervals. However, the model systematically underestimated accuracy in the Like condition (observed = 80.9%, predicted = 76.2%, 95% HDI: [73.6%, 78.6%]), with the observed value falling outside the prediction interval. Subsequently, a weaker predicted condition effect (Bayesian p = 0.005) and smaller predicted pairwise differences were observed for Neutral vs. Like (p = 0.001) and Like vs. Dislike (p = 0.003).

To test whether this also reflected low α^+ variance, we fit the model to simulated data with inflated α^+ variance as described above, and re-generated 4,000 posterior predictive datasets. Doing so significantly improved the similarity between the observed and predicted data for the Like condition (observed = 67.6%, predicted = 67.9%, 95% HDI: [65.7%, 70.0%], Bayesian p-value = 0.756). The remaining conditions were also well-recovered: Neutral (observed = 50.0%, predicted = 50.0%, 95% HDI: [47.9%, 52.2%], p = 0.982) and Dislike (observed = 70.7%, predicted = 70.8%, 95% HDI: [68.9%, 72.7%], p = 0.863), with all observed values falling within their respective 95% prediction intervals. The model also accurately captured the overall condition effect (observed F = 94.17, predicted mean F = 101.61, 95% HDI: [71.73, 136.80], p = 0.664) and all pairwise differences between conditions (all Bayesian p-values > 0.75; Supplementary Figure 2).



Supplementary Figure 1. Parameter recovery for the winning computational model. Group-level mean parameters were drawn from the empirical posterior of the winning model, with between-subject variance for α^+ inflated to match empirically plausible levels (CV ≈ 0.74–0.98). The data-generating model was then fitted to synthetic behavioral data from 193 simulated participants, and parameter estimates were compared to the true data-generating values. **(A)** At the group level, true mean and group-level standard deviation parameters fell within the 95% HDI of recovered posterior distributions across all conditions. Note that SD parameters are reported in probit space. **(B)** At the individual level, parameter recovery

showed good identifiability for both the negative learning rate (ρ = 0.60–0.66) and positive learning rate (ρ = 0.57–0.68)."



Supplementary Figure 2. Posterior predictive checks of the winning model. Each condition has the distribution of predicted mean choice accuracy across 4,000 posterior predictive datasets plotted per condition (Neutral = green, Like = blue, Dislike = red), within the mean predicted accuracy. The orange dot denotes the observed mean accuracy from the simulated dataset, representing the empirical estimate. Significance brackets reflect pairwise differences in the predicted data. *** $p < .001$

Subjective ratings and learned value

To verify that participants' explicit social appraisals tracked the feedback contingencies across conditions, we examined whether model-derived learned values predicted self-reported ratings of how much the feedback agent liked them. This served as a manipulation check, confirming that the computational learning signals captured by the model corresponded to participants' subjective experience of the social feedback.

For each participant and condition, we reconstructed the trial-by-trial trajectory of $V^{positive}$ - the learned value of the positive-word option - by replaying the Rescorla-Wagner update rule

forward using each participant's posterior mean α^+ and α^- estimates. Specifically, $V^{positive}$ was initialised to zero at the start of each block and updated on each valid trial according to:

$$V_t^{positive} = V_{t-1}^{positive} + \alpha^{\pm} \cdot (r_t - V_{t-1}^{positive}) \quad (5)$$

where α^+ was applied following positive outcomes and α^- following negative outcomes. The mean $V^{positive}$ across all valid trials within each block was taken as the learned social value for that condition. This quantity reflects the degree to which participants learned that the agent responded positively to their choices: values approaching +1 indicate strong learned association between positive choices and positive feedback, while negative values indicate the inverse.

Explicit ratings were collected via a continuous slider (0–100) at the end of each block, with participants indicating how much they felt the feedback agent liked them. We computed Kendall's τ correlations between mean $V^{positive}$ and ratings separately for each condition.

Supplementary tables

Predictor	Choice Accuracy	Reaction Time	Rating
Intercept	**0.290* (0.122)**	**7.423*** (0.066)**	**36.607*** (3.848)**
Like	**1.049*** (0.164)**	-0.105 (0.074)	**29.071*** (4.324)**
Dislike	**1.080*** (0.140)**	**-0.147* (0.059)**	**-18.017*** (4.221)**
α+	**4.935* (2.230)**	2.247 (1.260)	-64.323 (75.272)
α−	-0.117 (0.472)	**0.510** (0.184)**	-9.918 (10.725)
Depressive Self-Identity	0.008 (0.122)	-0.047 (0.078)	2.074 (4.604)
Social Avoidance	-0.022 (0.128)	0.032 (0.077)	-5.060 (4.606)
Emotional Insensitivity	**0.243* (0.119)**	0.042 (0.069)	-0.711 (4.056)
Like x α+	-0.697 (2.316)	-2.226 (1.203)	94.351 (78.096)
Dislike x α+	-1.421 (2.826)	-1.658 (1.144)	5.432 (89.455)
Like x Depressive Self-Identity	0.033 (0.177)	0.026 (0.087)	-1.935 (5.132)
Dislike x Depressive Self-Identity	0.045 (0.136)	0.058 (0.070)	0.042 (4.987)
Like x Social Avoidance	0.013 (0.183)	-0.011 (0.086)	2.736 (5.218)
Dislike x Social Avoidance	0.076 (0.144)	-0.027 (0.070)	3.409 (4.968)
Like x Emotional Insensitivity	**-0.480** (0.167)**	**-0.166* (0.079)**	-2.841 (4.615)
Dislike x Emotional Insensitivity	**-0.323* (0.139)**	-0.109 (0.062)	3.621 (4.440)
α+ x Depressive Self-Identity	-0.698 (2.082)	-0.159 (1.482)	24.555 (89.024)
α+ x Social Avoidance	1.576 (2.432)	0.451 (1.541)	-111.455 (94.878)
α+ x Emotional Insensitivity	3.414 (2.097)	2.336 (1.320)	-28.313 (78.797)

Like x α−	-0.046 (0.678)	-0.251 (0.189)	16.568 (14.527)
Dislike x α−	0.963 (0.653)	0.041 (0.183)	8.668 (14.948)
Depressive Self-Identity x α−	-0.492 (0.521)	0.378 (0.210)	-10.270 (12.059)
Social Avoidance x α−	0.897 (0.528)	**-0.436* (0.214)**	19.899 (12.408)
Emotional Insensitivity x α−	-0.328 (0.490)	0.179 (0.197)	6.092 (11.392)
Like x α+ x Depressive Self-Identity	1.259 (2.188)	0.302 (1.427)	-44.807 (92.670)
Dislike x α+ x Depressive Self-Identity	0.366 (2.445)	0.768 (1.338)	61.081 (104.559)
Like x α+ x Social Avoidance	-2.297 (2.622)	-0.939 (1.496)	153.429 (98.570)
Dislike x α+ x Social Avoidance	2.118 (3.014)	0.342 (1.438)	66.422 (111.565)
Like x α+ x Emotional Insensitivity	-3.126 (2.324)	-1.688 (1.264)	42.010 (82.483)
Dislike x α+ x Emotional Insensitivity	-2.417 (2.698)	-1.597 (1.189)	20.662 (92.636)
Like x Depressive Self-Identity x α−	0.709 (0.768)	-0.079 (0.217)	13.631 (16.758)
Dislike x Depressive Self-Identity x α−	-0.434 (0.723)	-0.102 (0.205)	17.190 (16.849)
Like x Social Avoidance x α−	-0.737 (0.738)	**0.567** (0.214)**	-20.751 (16.307)
Dislike x Social Avoidance x α−	-0.517 (0.767)	0.297 (0.215)	-23.714 (17.618)
Like x Emotional Insensitivity x α−	0.089 (0.712)	0.134 (0.202)	-16.903 (15.501)
Dislike x Emotional Insensitivity x α−	0.369 (0.710)	0.083 (0.200)	-4.951 (16.364)

Supplementary Table 2. Regression coefficients predicting behavioral task performance. Models consisted of a fixed-effects logistic regression for choice accuracy, linear mixed-effects model for reaction time, and standard linear regression for likeability ratings. All models included condition (Like, Dislike relative to Neutral), transdiagnostic factor scores and their interactions as predictors. Mixed-effects models included random intercepts for subjects. Model formula: DV ~ condition * alpha_pos * (fa1 + fa2 + fa3) + condition * alpha_neg * (fa1 + fa2 + fa3) + (1 + condition | Subject). Coefficients for factor scores represent semi-standardized estimates (standardized predictors, unstandardized outcomes) with standard errors in parentheses. ** p < 0.05, ** p < 0.01, *** p < 0.001*

Predictor	**α^{+}**	**α^{-}**	**α^{diff}**
Intercept	**0.325*** (0.003)**	**0.173*** (0.009)**	**0.152*** (0.010)**
Like	**0.119*** (0.003)**	**-0.036*** (0.006)**	**0.155*** (0.007)**
Dislike	**0.025*** (0.003)**	0.000 (0.006)	**0.025*** (0.007)**
Depressive Self-Identity	-0.001 (0.003)	-0.001 (0.010)	0.000 (0.011)
Social Avoidance	0.001 (0.003)	**0.025* (0.010)**	**-0.024* (0.011)**
Emotional Insensitivity	-0.001 (0.003)	-0.011 (0.009)	0.010 (0.010)
Like x Depressive Self-Identity	0.003 (0.004)	0.002 (0.007)	0.001 (0.008)
Dislike x Depressive Self-Identity	-0.001 (0.004)	-0.005 (0.007)	0.003 (0.008)
Like x Social Avoidance	0.002 (0.004)	-0.006 (0.007)	0.008 (0.008)
Dislike x Social Avoidance	-0.001 (0.004)	0.001 (0.007)	-0.002 (0.008)
Like x Emotional Insensitivity	-0.002 (0.003)	0.006 (0.006)	-0.008 (0.007)

Dislike x Emotional Insensitivity	0.000 (0.003)	-0.000 (0.006)	0.000 (0.007)

Supplementary Table 3. Regression coefficients predicting computational parameters from the winning model. The table shows results from three separate linear mixed-effects regression models with condition (Like/Dislike relative to Neutral), transdiagnostic factor scores (Depressive Self-Identity, Social Avoidance, Emotional Insensitivity), and their interactions predicting each computational parameter from the winning reinforcement learning model. Outcome variables across the four models were positive learning rate (α^+), negative learning rate (α^-) and positive learning bias ($\alpha^+ - \alpha^-$). Model formula: DV ~ condition × (fa1 + fa2 + fa3) + (1|subID). Values reflect semi-standardized regression coefficients (standardized predictors, unstandardized outcomes) with standard errors in parentheses. ** p < 0.05, ** p < 0.01, *** p < 0.001*

Item	Depressive Self-Identity	Social Avoidance	Emotional Insensitivity
NARQ1			
NARQ2			
NARQ3			
NARQ4			
NARQ5			
NARQ6			
SPIN1		0.539	
SPIN2		0.545	
SPIN3		0.511	
SPIN4		0.7	
SPIN5		0.624	
SPIN6		0.623	
SPIN7		0.433	
SPIN8		0.557	
SPIN9		0.441	
SPIN10		0.647	
SPIN11		0.703	
SPIN12		0.608	
SPIN13		0.751	
SPIN14		0.414	
SPIN15		0.426	
SPIN16		0.479	
SPIN17		0.668	
AMI1			0.423
AMI2		0.54	
AMI3		0.415	

AMI4		0.317	0.302
AMI5		0.458	
AMI6			
AMI7			0.476
AMI8		0.323	
AMI9			
AMI10			
AMI11			
AMI12			
AMI13			0.657
AMI14		0.497	
AMI15			
AMI16			0.549
AMI17		0.429	
AMI18			0.579
AQ1			
AQ2			
AQ3			
AQ4			
AQ5		0.312	0.427
AQ6			0.472
AQ7			
AQ8			
AQ9			0.386
AQ10			0.301
BDI1	0.629		
BDI2	0.54		
BDI3	0.672		
BDI4	0.632		
BDI5	0.551		
BDI6	0.615		
BDI7	0.732		
BDI8	0.638		
BDI9	0.575		
BDI10	0.652		
BDI11	0.649		
BDI12	0.646		
BDI13	0.646		
BDI14	0.544		
BDI15	0.533		
BDI16	0.608		
BDI17	0.651		
BDI18	0.514		
BDI19			

BDI20	0.577
BDI21	0.493
BPD1	0.305
BPD2	0.341
BPD3	
BPD5	0.409
BPD6	0.378
BPD7	0.502
BPD8	
BPD9	0.437

Supplementary Table 4. Factor loadings from the exploratory factor analysis three-factor solution. Only loadings ≥ 0.30 are displayed, consistent with standard reporting thresholds for exploratory factor analysis.