

# Learning About the Self: Motives for Coherence and Positivity Constrain Learning From Self-Relevant Social Feedback



Jacob Elder<sup>1</sup>, Tyler Davis<sup>2</sup>, and Brent L. Hughes<sup>1</sup>

<sup>1</sup>Department of Psychology, University of California, Riverside, and <sup>2</sup>Department of Psychological Sciences, Texas Tech University

Psychological Science  
2022, Vol. 33(4) 629–647  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/09567976211045934  
www.psychologicalscience.org/PS



## Abstract

People learn about themselves from social feedback, but desires for coherence and positivity constrain how feedback is incorporated into the self-concept. We developed a network-based model of the self-concept and embedded it in a reinforcement-learning framework to provide a computational account of how motivations shape self-learning from feedback. Participants ( $N = 46$  adult university students) received feedback while evaluating themselves on traits drawn from a causal network of trait semantics. Network-defined communities were assigned different likelihoods of positive feedback. Participants learned from positive feedback but dismissed negative feedback, as reflected by asymmetries in computational parameters that represent the incorporation of positive versus negative outcomes. Furthermore, participants were constrained in how they incorporated feedback: Self-evaluations changed less for traits that have more implications and are thus more important to the coherence of the network. We provide a computational explanation of how motives for coherence and positivity jointly constrain learning about the self from feedback, an explanation that makes testable predictions for future clinical research.

## Keywords

self-concept, reinforcement learning, network analysis, motivated cognition, positivity, coherence, open data, open materials

Received 1/28/21; Revision accepted 8/11/21

Our self-concepts are not static structures but rather fluctuate from social feedback throughout our lifetimes (Bem, 1972; Mead, 1934; Vygotsky, 1978). However, we are not mere reflections of feedback that we receive—we desire to view ourselves positively (Alicke & Sedikides, 2009; Taylor & Brown, 1988) and to maintain coherent and consistent views of ourselves (Markus & Wurf, 1987; Swann et al., 2003). How people acquire and maintain self-evaluations from social feedback may parallel how people more generally learn from feedback. Specifically, research demonstrates that positivity biases (Lefebvre et al., 2017), self-relevant biases (Koban et al., 2017; Müller-Pinzler et al., 2019), and associations between feedback (Vaidya & Badre, 2020) can shape how people modify their behavior from feedback. We unite these findings by considering how

feedback-based learning (i.e., reinforcement-learning) algorithms may be similarly employed to learn about the multifaceted self-concept (see Lockwood et al., 2020). We developed a computational framework to describe how the dual drives of positivity and coherence constrain learning about the self-concept from social feedback.

Although there is extensive evidence that people strive for self-concept positivity and coherence, research often pits these drives as antagonistic frameworks as opposed to dual constraints on self-representation. On

## Corresponding Author:

Brent L. Hughes, University of California, Riverside, Department of Psychology  
Email: bhughes@ucr.edu

the one hand, people frequently engage in self-enhancement processes in response to social feedback (Dunning, 1999; Hepper & Sedikides, 2012; Vandellen et al., 2010). People preferentially seek positive feedback over negative feedback (Pyszczynski et al., 1985), overestimate the positive feedback they receive (Hepper et al., 2011; Rodman et al., 2017; Somerville et al., 2010), and are more likely to update self-views from positive than from negative feedback (Korn et al., 2012). Likewise, people selectively forget or discard negative feedback (M. C. Anderson & Hanslmayr, 2014; Frey et al., 1986; Schröder-Abé et al., 2007) and overcompensate in response to negative feedback (Beer et al., 2013; Greenberg & Pyszczynski, 1985; Hughes & Beer, 2013). On the other hand, people are also motivated to maintain coherent self-concepts (Markus & Wurf, 1987), even at the expense of positivity (Swann et al., 2003). For example, individuals seek interpersonal feedback that is consistent with their self-views, even when it is unfavorable (Swann et al., 1992), and they resist inconsistent feedback (Swann & Hill, 1982). However, this area of research lacks a formal explanation of how motivations to maintain positive and coherent self-views dually constrain self-concept malleability.

We developed a computational framework for how people dynamically update self-views to maintain positivity and coherence. We formalized the self-concept as a causal network of traits embedded in a reinforcement-learning agent that learns asymmetrically from positive and negative outcomes. Using this framework, we were able to make predictions about how people update their beliefs about individual traits. From a causal-network perspective, traits that are critical for maintaining network coherence are those that are central in the network's structure (Sloman et al., 1998). Our causal trait network was developed by having an independent sample of participants provide semantic evaluations of the causal relationships between traits (Elder et al., 2022). For example, in this model, the trait "outgoing" causes the traits "sociable" and "fun," and "fun" in turn causes "witty." By extension, an individual's self-views on "outgoing" should thereby bear implications on their self-views of "sociable" and "fun." This network model describes how people represent traits and their relationships and how this structured representation shapes self-concept maintenance and malleability. The network model instantiates normative beliefs about the compatibility and causal dependence among traits, in contrast to how traits vary and cause each other in actual personalities. If people believe they are outgoing, they ought to endorse traits that being outgoing depends on, and if they believe they are not outgoing, they ought not to endorse such traits. Coherence is achieved by maintaining consistency between self-perceptions

### Statement of Relevance

People change how they see themselves from everyday social experiences, but this process is not without constraints. Rather, people may selectively incorporate feedback in ways that maintain self-concept positivity and coherence. We developed a trait network that describes perceived causal relationships between traits and how the updating of self-views may be constrained by beliefs about causal structure. Participants evaluated themselves on a network of traits while receiving social feedback from other people that varied in positivity. Participants learned how other people saw them and adjusted their self-views accordingly, but surprisingly, they disproportionately incorporated positive feedback more than negative feedback. This tendency was positively associated with individual differences in mental health and psychological adjustment. Crucially, people resisted changing their self-views on traits with more perceived implications, which may help to preserve coherence by avoiding changes that could contradict other downstream self-views. These findings provide insight into how mental health relates to asymmetrical self-learning and how people simultaneously balance motivations for coherence and positivity.

for traits in the network, which suggests that highly central traits, that many traits depend on, are particularly critical for maintaining coherence.

Our causal-network theory draws an important distinction between *out-degree* and *in-degree centrality*. Out-degree centrality describes the number of traits that depend on (or are outputs or children of) a given trait. A trait has more out-degree centrality if many traits semantically depend on it (e.g., "outgoing" causes "sociable"). In-degree centrality describes the number of traits that cause (or are inputs or parents of) a given trait (e.g., "witty" is caused by "fun") or the number of other traits that a trait depends on for its meaning. A trait that depends on many other traits will have more in-degree centrality. Both types of centrality can be important for self-concept coherence but likely relate to self-concept malleability in different ways. Traits with higher out-degree centrality are more important to self-concept coherence because changes to such traits would result in a cascade of changes to their dependent traits. Thus, traits with higher out-degree centrality should be more resistant to change. Traits with higher in-degree centrality, in contrast, receive more inputs from other

traits during feedback and may thus be more malleable because of their sensitivity to many influences.

To test predictions from this computational framework, we employed a task that assigns social feedback according to network-defined communities of traits. Participants evaluated themselves on all network traits and, after each trait rating, received social feedback on each trait (ostensibly based on an interview). Each trait community received different likelihoods of positive feedback, which determined the feedback assigned to each trait. One simple test of network structure is whether participants learn how they are perceived on particular trait communities by discriminating the different levels of feedback they received.

Our computational framework generated several predictions about how participants would maintain self-concept positivity and coherence. First, we expected that the motivation for positivity would manifest as a tendency to learn more from positive than negative feedback; specifically, network communities associated with higher probabilities of positive feedback would exhibit greater change in self-evaluations over the course of learning and from before to after feedback. Computationally, this observation would be reflected in asymmetrical learning rates (reinforcement-learning parameters that describe how much participants incorporate feedback). We predicted that learning rates would be higher for positive feedback than for negative feedback. The tendency to asymmetrically process social feedback when learning about the self-concept may help people to maintain positive self-views. Thus, we tested whether personality measures generally reflecting positive self-views and psychological adjustment are associated with learning rates. Second, we expected that a trait's out-degree and in-degree centrality would impact how much people update self-evaluations on the basis of feedback. Specifically, we predicted that self-evaluations for traits with higher out-degree centrality would change less from feedback, and self-evaluations for traits with higher in-degree centrality would change more from feedback.

## Method

### Participants

We recruited a convenience sample of 48 undergraduate students via the University of California, Riverside, credit pool in compliance with the protocols of the university's institutional review board. All participants provided informed consent. The sample was 67.3% female, between the ages of 18 and 24 years ( $M = 19.45$  years), and 32.7% Asian, 20.4% Hispanic, 16.3% mixed race, 12.2% African American, 10.2% Caucasian, and

8.2% other. Two participants were omitted because they did not believe the experimental manipulation, and two participants' questionnaire measures were missing because of protocol-administration errors. The target minimum sample size of 44 was determined by a prospective power analysis for our primary within-participants manipulation, which tested the average difference in self-evaluations between the five trait communities that received feedback. We assumed a moderately small effect ( $\eta_p^2$ ) of .03 (an effect size shown in related literature; e.g., Beer et al., 2013; Korn et al., 2012) and a correlation among repeated measures of .45 and power above 80%.<sup>1</sup> We collected data in weekly posted slots until the minimum was reached. After the minimum sample size was reached, data collection continued until the conclusion of the academic quarter in which it started.

We visually inspected the data for quality. We conducted further quality control by flagging any participants who had missing responses more than 15% of the time and who gave the same response more than 70% of the time at both initial self-evaluations and reevaluations. No participants met these criteria, and thus none were excluded.

### Trait-network development

One hundred fifty traits were selected from a list of 292 positive traits (N. Anderson, 1968; Hampson et al., 1987; Kirby & Gardner, 1972). The list was filtered to 148 by randomly drawing 150 traits across 10,000 simulations and identifying which traits achieved the greater consistency across five dimensions of normative data: desirability, prevalence, category breadth, observability, and interpersonalitv (for a list of all traits, see Table S1 in the Supplemental Material available online). Two traits were removed from the list of 150 for having a mean desirability below 4.0 (for more details, see Elder et al., 2022).

On each trial, participants were presented with one trait word as a target trait. Participants were then presented with a list of the remaining 147 positive trait words and asked, "Which traits does [target trait] depend upon?" Participants were able to nominate as many trait words as they believed were applicable. Participants completed a total of 10 trials with 10 randomly selected traits as the target trait. This procedure was executed for all 148 positive traits and 148 negative traits; at the conclusion, 144 positive traits were presented 12 times as the target word, whereas four positive traits were presented 13 times as the target trait across participants. An adjacency matrix of 148 rows by 148 columns for trait words was computed on the basis of the number of causal relationships nominated by participants. For

each nomination of dependency by a participant, the corresponding cell (i.e., edge) within the adjacency matrix increased by 1 (e.g., if a participant nominated “outgoing” as causing “nice,” then 1 was added to the row for “nice” and the column for “outgoing”). An a priori threshold of 25% was set for this adjacency matrix in order to remove edges with lower numbers of endorsements. Setting such a threshold should filter idiosyncrasies in dependency nominations that do not reflect consensus of trait relationships (for more details, see Elder et al., 2022).

Out-degree centrality was defined as the number of directed edges from a target word to other words (i.e., how many traits depend on a target trait). In-degree centrality was defined as the number of directed edges from other words to the target word (i.e., how many traits a target trait depends on). Additionally, we identified communities of traits in the network by using a “walktrap” community-detection algorithm (Pons & Latapy, 2005). This algorithm uses the length of random walks from a node in a network to other nodes in order to identify densely interconnected groups of traits. We defined the communities used to assign feedback from an initial procedure that detected five communities. However, later analyses uncovered a coding error that excluded one trait word that, when corrected, led the same algorithm to detect four total communities. Given the role of the original community construction in the experimental design, we retained the original solution for all analyses and modeling. All 148 traits from the positive trait network were incorporated into the subsequent experimental protocol (for a visualization of the network, see Fig. 1).

### **Experimental procedure**

Participants completed an initial lab visit in which they provided written consent and received the cover story for the experiment. They were informed that they would be undergoing an interview that would be recorded and shared with three to five members of the University of California, Riverside, undergraduate admissions committee. During the interview, participants were asked a range of questions about their personal characteristics, goals, and interests (for interview questions, see Table S2 in the Supplemental Material), and interviews lasted approximately 10 to 20 min. Following the interview, participants completed several questionnaires and were scheduled to return to the lab for the second appointment to complete the social-evaluative task (approximately 7 to 10 days later).

Participants returned to the lab for the second visit to complete a social-evaluative task on a computer. Participants were led to believe that in the time between

the first and second visits, three to five members of the University of California, Riverside, admissions committee had evaluated them on all 148 trait words on the basis of their video interviews. Participants were told that during the task, they would be asked to evaluate themselves on an array of traits. However, while evaluating themselves, they would also observe how the committee members had evaluated them. It was not made explicit whether they should take these evaluations into consideration when making their own self-judgments.

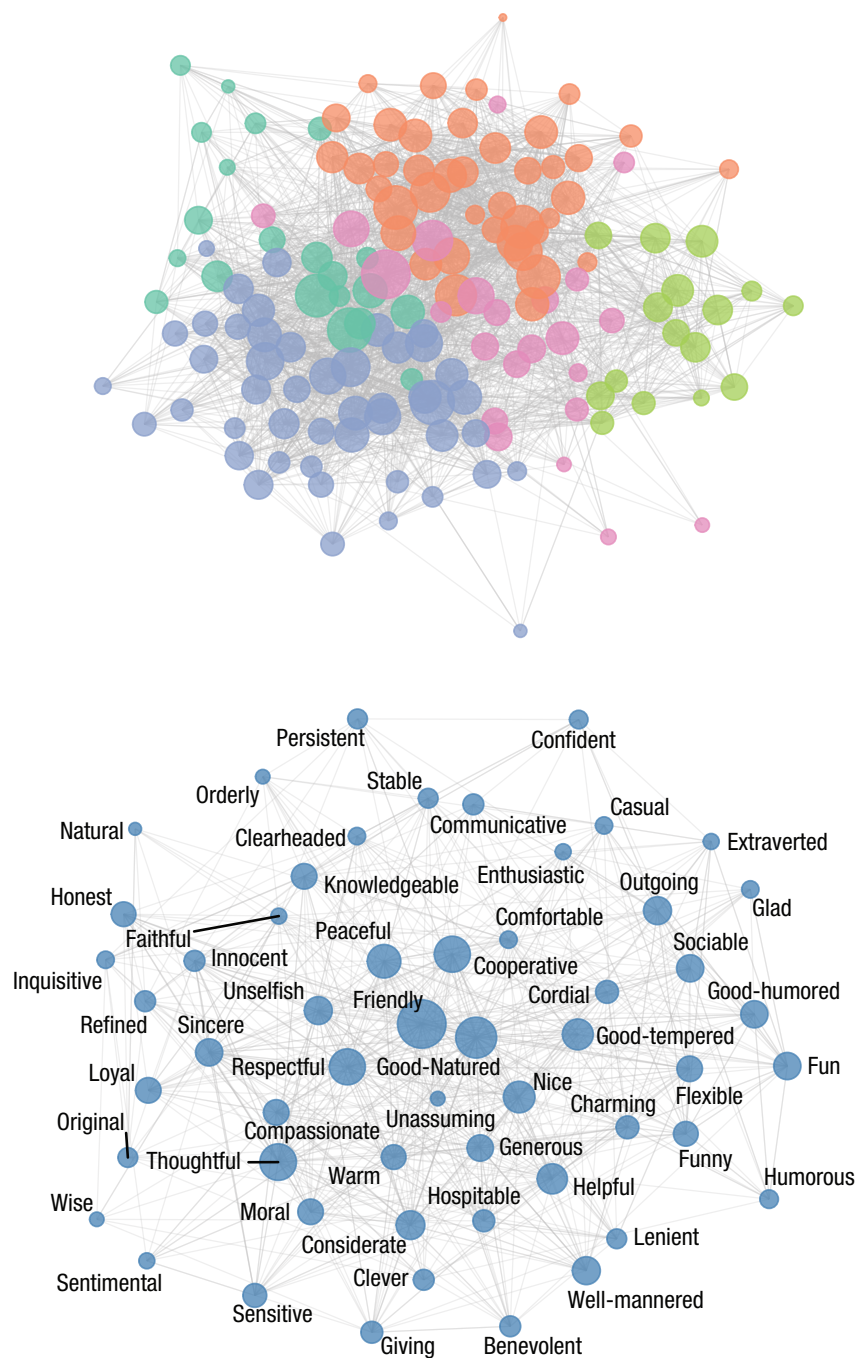
At the conclusion of the second visit, participants were debriefed and informed that the feedback was administered via a pseudorandom algorithm and that there were no committee members who had reviewed their interviews. They also were asked a series of offboarding questions to assess any confusion, engagement, and belief in the authenticity of the manipulation. Following this exit interview, they were awarded a credit for completion.

### **Social-evaluative-feedback task**

The experimental task was programmed in MATLAB (Version 2018b; The MathWorks, Natick, MA) and the Psychophysics Toolbox (Version 3.0.14; Brainard, 1997) and was similar to other social-evaluative-feedback tasks (Eisenberger et al., 2011; Hughes & Beer, 2013; Korn et al., 2012; Somerville et al., 2010; Will et al., 2017). Participants evaluated themselves on all 148 positive traits from the trait network on a scale ranging from 1 (*not at all*) to 7 (*very much*) in response to the prompt, “To what extent does the following trait describe you?” The number that participants selected as self-descriptive was framed in an orange square after the response was made. Participants were permitted to respond in as much time as they needed and could opt out of making a self-evaluation if they were unfamiliar with a given trait word. After each self-evaluation, the screen froze for a constant 0.20 s, and then participants were presented with the ostensible social-evaluative feedback. Feedback appeared with the prompt, “The reviewers see you as . . .,” with the trait at the center of the screen and a red square around the assigned feedback. The orange score denoting the participant self-evaluation remained on screen for the feedback phase, and a number in white denoted the difference between the participant’s rating and the reviewers’ rating (e.g., “+3” if reviewer feedback was 7 and participant self-evaluation was 4; for an illustration of the task, see Fig. 2).

Feedback was administered via a pseudorandom algorithm. Five different probabilities of positive feedback—90%, 70%, 50%, 30%, or 10%—were randomly assigned to each of the five trait-network communities for each participant. For instance, for a trait belonging

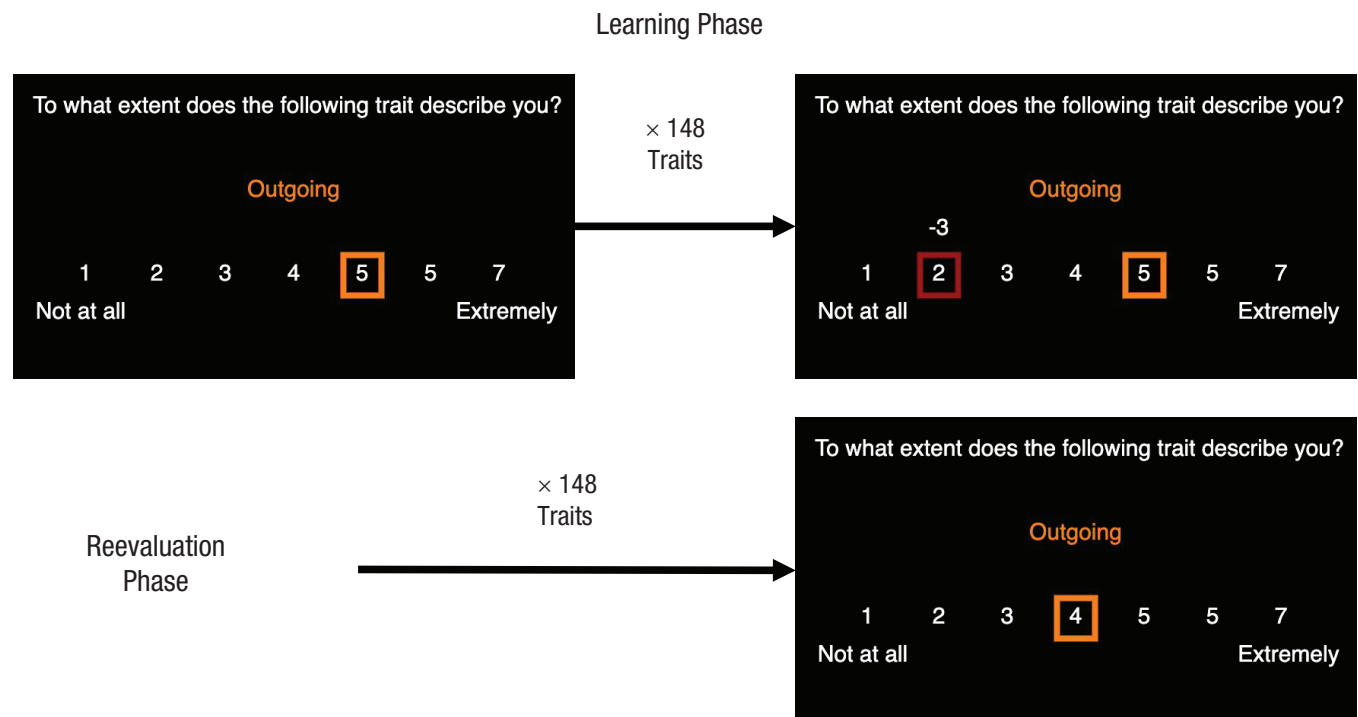




**Fig. 1.** Network visualization for the full network and a subset of the network. Larger nodes indicate greater out-degree centrality. The top panel shows a network visualization of all 148 traits, colored by community. The bottom panel shows out-degree connections for a trait with high out-degree centrality: “friendly.”

to the 70% community, a participant would have a 30% probability of receiving negative feedback (i.e., on average, feedback was lower than the self-evaluation) and a 70% probability of receiving positive feedback (i.e., on average, feedback was higher than the self-evaluation). Given that all traits used in the study were

positive, higher numerical feedback on the scale from 1 to 7 was considered more positive. After feedback was determined according to a given community’s probability, the feedback number on the ordinal scale from 1 to 7 was assigned according to criteria related to the participant’s response and the determined



**Fig. 2.** Illustration of the social-evaluative task. During the learning phase, participants evaluated themselves on each of 148 traits (e.g., "outgoing"). An orange square appeared around their selection. After each self-evaluation, they were then shown how an admissions committee ostensibly evaluated them on that same trait. A red square appeared around the committee's selection, and the discrepancy between the participant's self-evaluation and the committee's evaluation was displayed in white text. After completing self-evaluations and receiving feedback on all 148 traits, participants proceeded to the reevaluation phase. They then self-evaluated themselves again on each trait.

feedback valence (for feedback-assignment rules, see Table S3 in the Supplemental Material). For example, if the presented trait belonging to the 70% community was assigned negative feedback because of a 30% probability, and the self-evaluation was a 4, the feedback would be randomly assigned from 1 to 3. If the trait was assigned positive feedback because of the 70% probability, feedback would be randomly assigned a number from 4 to 7. Therefore, feedback was contingent on participant self-evaluations (for means and standard deviations of feedback and other variables, see Table S4 in the Supplemental Material).

### Self-report measures

**Rosenberg Self-Esteem Scale.** A 10-item questionnaire that assesses individual differences in global personal self-esteem was administered (Rosenberg, 1965). The scale demonstrated excellent reliability in the current sample ( $\omega = .93$ ).

**Self-Concept Clarity Scale.** A 12-item questionnaire was administered to assess individual differences in how clearly defined, internally consistent, and stable beliefs about the self are (Campbell et al., 1996). The

scale demonstrated good reliability in the current sample ( $\omega = .87$ ).

**Multidimensional Assessment of Interoceptive Awareness Scale, Version 2.** A 37-item questionnaire was administered to assess individual differences in conscious levels of interoception, "the process by which the nervous system senses, interprets, and integrates signals originating from within the body" (Mehling et al., 2018, abstract). The scale demonstrated excellent reliability in the current sample ( $\omega = .91$ ).

**Dialectical Self-Views Scale.** A 32-item questionnaire was administered to assess individual differences in tolerance for contradictory views about oneself, maintenance of both positive and negative self-views, the expectation for self-views to change, and cognitive holism (Spencer-Rodgers et al., 2015). The scale demonstrated good reliability in the current sample ( $\omega = .82$ ).

**Singelis Self-Construal Scale.** A 30-item scale was used to assess individual differences in independent (15-item subscale) and interdependent (15-item subscale) self-construal styles (Singelis, 1994). Both subscales exhibited good reliabilities, respectively ( $\omega = .83$  for both).

**Beck Depression Inventory.** A 21-item scale was used to assess individual differences in characteristic attitudes and symptoms of depression (Beck et al., 1996). The scale exhibited excellent reliability ( $\omega = .93$ ).

**Liebowitz Social Anxiety Scale.** A 24-item scale was administered to assess individual differences in social-anxiety-related symptomatology (Liebowitz, 1987). The scale exhibited excellent reliability ( $\omega = .96$ ).

**Self-Monitoring Scale.** An 18-item scale was administered to assess individual differences in concerns and ability to modify behaviors to meet the perceptions of other people (Snyder, 1974). The scale exhibited acceptable reliability ( $\alpha = .78$ ).

**Rejection Sensitivity Scale.** An 18-item scale was administered to assess individual differences in sensitivity to perceived or actual social rejection (Downey & Feldman, 1996). The scale exhibited good reliability ( $\omega = .83$ ).

**Big Five Inventory–2 Extra-Short Form.** A 15-item short scale of the Big Five Inventory (Soto & John, 2017) was administered to assess individual differences in the trait factors negative emotionality/neuroticism, openness, agreeableness, extraversion, and conscientiousness. Openness demonstrated poor reliability ( $\omega = .58$ ), neuroticism demonstrated relatively poor reliability ( $\omega = .61$ ), conscientiousness demonstrated poor reliability ( $\omega = .54$ ), extraversion demonstrated acceptable reliability ( $\omega = .75$ ), and agreeableness demonstrated very poor reliability ( $\omega = .32$ ). As noted by the creators, the low reliability is to be expected for brief scales that prioritize content validity over internal consistency.

### Computational model

To capture how individuals dynamically learn from social feedback and how the motivation for positivity shapes this learning, we employed a reinforcement-learning approach. We used a standard reinforcement-learning model (Rescorla & Wagner, 1972) in which the trait communities were treated as Pavlovian cues ( $C$ ), with social expectations ( $SE$ ) describing the model's estimate of expected rating for the current community. Weights ( $W_t$ ) were defined as the learned associations for the five trait communities. Social expectations were operationalized as estimates of other people's perceptions of them on a given community of related traits based on prior social feedback. The weight for the current trial's trait community is the current social expectation. Thus, the model describes how the expected rating ( $SE_{t+1}$ ) of a trait community ( $C_t$ ), as

determined by the trait present on a given trial, is updated from its existing social expectation ( $SE_t$ ), according to the error term ( $\delta_t$ ) and how much of the error term is incorporated by the learning rate ( $\alpha$ ). Weights ( $W_t$ ) were initialized at the midpoint 4 to represent participants' neutral expectations regarding other people's perceptions of them, consistent with the reinforcement-learning convention of initializing values between possible outcomes (Zhang et al., 2020). We tested a variety of different models (in the Supplemental Material, see Fig. S1 for model comparisons and Supplementary Text for model descriptions). Our final, best-performing model included four free parameters: a negative prediction-error learning rate ( $\alpha_n$ ; bounded from 0 to 1), a positive prediction-error learning rate ( $\alpha_p$ ; bounded from 0 to 1), an (inverse) feedback-sensitivity parameter ( $\beta$ ; bounded from 0.5 to 9), and a decay parameter ( $\phi$ ; bounded from 0 to 1).

At each trial, an individual's expectation ( $SE_t$ ) corresponded to the weight ( $W_t$ ) for the current trait's community ( $C_t$ ):

$$SE_t = W_t(C).$$

Feedback (in whole numbers ranging from 1 to 7) was presented after the self-evaluation. Feedback was scaled by a feedback-sensitivity parameter (inverse  $\beta$ ) that accounts for range attenuation of responses at the natural boundaries of the Likert-type scale:

$$F_t = \frac{1}{1 + e^{\beta \times -(F_t - 4)}}.$$

Feedback at the high (e.g., 6, 7) or low (e.g., 1, 2) ends of the scale had less range to modify participants' social expectations in positive or negative directions, respectively. A lower feedback sensitivity reflects that feedback outside the midpoint is enhanced and processed as more extreme than it is (e.g., 2 is processed as approximately 1), whereas a higher feedback sensitivity reflects that feedback outside the midpoint is blunted and processed as less extreme than it is (e.g., 1 is processed as approximately 2). In other words, individuals with a higher (inverse) feedback-sensitivity parameter are more insensitive to differences in feedback (feedback shrinks toward the midpoint), whereas individuals with a lower feedback sensitivity parameter are more sensitive to differences in feedback (feedback expands toward the extremes; for a visualization of feedback-sensitivity-scaling feedback, see Fig. S2 in the Supplemental Material).

Feedback is centered at 0 for the sigmoid function (i.e., feedback sensitivity) and then rescaled back to the Likert-type scale after transformation. It is not

possible for transformed feedback to be below 1 or above 7, and thus feedback and expectations will never exceed the boundaries of the Likert-type scale:

$$F_t = (F_t \times 6) + 1.$$

The error term consists of the difference between trial-level transformed feedback ( $F_t$ ) on a given trait community ( $C_t$ ) from the current expected rating ( $SE_t$ ) of the community. The error term, also known as *prediction error*, is computed as follows:

$$\delta_t = F_t - SE_t.$$

Because of separate learning rates for positive and negative prediction errors, reflecting asymmetrical learning, participants incorporated prediction errors differentially depending on whether feedback was positive or negative:

$$\begin{aligned} \text{if } \delta_t > 0: SE_{t+1} &= SE_t + \alpha_p \times \delta_t \\ \text{if } \delta_t \leq 0: SE_{t+1} &= SE_t + \alpha_n \times \delta_t. \end{aligned}$$

At the beginning of the next trial, the weights were allowed to decay toward the midpoint, representing the forgetting of associations across trials:

$$W_t = W_t \times \phi + 4 \times (1 - \phi).$$

On each trial for which a trait was presented, the social expectation for the community that it belongs to was updated. For example, if “sociable” was presented on a trial, its broader community of 47 traits was applied as  $C_t$ . If “sociable” received scaled feedback of 6 and the community ( $C_t$ ) had an expectation ( $SE_t$ ) of 5, this would result in an error ( $\delta_t$ ) of 1, while only .25 of which would be incorporated in updating the future expectation ( $SE_{t+1}$ ) for  $C_t$  because of the .25 positive learning rate ( $\alpha_p$ ). In summary, the model describes how people vary in sensitivity to feedback depending on its position on the scale and learn differently from positive and negative prediction errors while also forgetting learned outcomes over time.

### Model fitting and comparison

We fitted the reinforcement-learning model using a two-stage procedure that incorporated both individual and group information. First, model free parameters were fitted using ordinary least squares (OLS) to generate parameter estimates for each participant. Next, we pooled these parameters across participants and used their means as a prior to regularize participants'

individual parameters. Ridge-penalized OLS estimation was then used to adjust individual parameter estimates on the basis of the group priors. This regularization procedure helps to stabilize individual estimates, produces more accurate estimates of true parameters, and reduces the influence of outliers (Daw, 2011; Lockwood et al., 2016):

$$\begin{aligned} \sum_{i=1}^n (\text{evaluation} - \text{social expectation})^2 \\ + \lambda \times \sum_{k=1}^m z\text{-scored estimate}^2. \end{aligned}$$

Ridge-penalized OLS regularized estimates toward the estimate priors, adjusted by a constant penalty parameter ( $\lambda$ ). We arrived at an appropriate penalty ( $\lambda = .15$ ) by iterating through penalties from .05 to 1 at increments of .05 and determining which penalty produced parameter estimates that were most recoverable and predictive of behavior. Given that the parameters are on different scales, we  $z$ -scored the estimates on the basis of their priors (OLS-estimated means and standard deviations), such that regularization penalized the estimates toward the mean. Parameters were fitted to each participant's self-evaluations using the “L-BFGS-B” optimization algorithm from the *optimx* package (Version 2021-6.12; Nash & Varadhan, 2011), available in the R programming environment (Version 4.0.3; R Core Team, 2020). To compare models, we employed a formulation of the Akaike information criterion (AIC) for residual sums of squares:

$$AIC = 2k + n \times \ln\left(\frac{RSS}{n}\right),$$

where  $n$  is the number of trials for participant  $i$ , RSS is the residual sum of squares for participant  $i$ , and  $k$  represents the number of free parameters estimated in the model. A lower AIC value reflects a better performing model. AIC values were summed across participants to estimate model performance.

### Parameter recovery

To determine whether parameters are identifiable (Lockwood & Klein-Flügge, 2021; Wilson & Collins, 2019; Zhang et al., 2020), we tested whether parameters could be recovered from simulated data. We performed three different tests of parameter recovery: (a) randomly simulating parameters, generating behavior from simulated parameters, and testing whether the simulated parameters could be recovered during fitting to generated behavioral data; (b) generating behavioral data using the original fitted parameters from the 46



participants and testing whether the original participants' parameters could be recovered during fitting to generated behavioral data; and (c) testing whether behavior generated from the recovered parameters paralleled the observed experimental behavior.

For each parameter from the best-fitting model, we identified five equally spaced intervals between the 25th percentile and 75th percentile of the parameters' distributions. At each interval, we drew 50 values for a given parameter and added Gaussian noise equivalent to 1 standard deviation of the parameter to increase the range of possible parameters simulated for positive learning rate, negative learning rate, and decay. Because of the nonnormal distribution of feedback sensitivity, we generated noise from a uniform distribution centered at each interval ranging from its negative standard deviation to its positive standard deviation. If the noise caused a parameter to reach the boundary of the parameter space, we set the parameter to its boundary. To simulate 250 participants (e.g., Palminteri et al., 2017), we randomly sampled without replacement from each of the newly generated parameters to determine a parameter set for a given participant. Using these generated parameter sets, we simulated behavior by generating feedback under the same algorithm used for the original task and according to the trial-by-trial social expectations determined by the parameters. We rounded trial-by-trial simulated estimates to the nearest whole number to emulate the Likert-type behavioral responses of the participants. Then, as with the original behavioral data, we fitted parameters to the simulated behavioral data. We then correlated the fitted parameters with the "true" parameters generated from the simulations to estimate whether parameters were recoverable.

The second parameter-recovery simulation was aimed at identifying how recoverable parameters were while maintaining the observed covariance structure of the original fitted participant parameters. This is in contrast to the previous recovery test that used simulated parameters that were randomly generated at intervals with noise and were then randomly shuffled into parameter sets, resulting in simulated parameters that are independent and uncorrelated and thus do not share the covariance structure of the original participants' fitted parameters (Vaidya & Badre, 2020). To account for this, we used participant-fitted parameters to generate new behavioral data, and parameters were then fitted to behavioral data generated by the original participant parameters. We estimated correlations between the fitted parameters and the original participant parameters in order to determine whether the original fitted parameters were recoverable, while accounting for the original covariance structure of parameters.

As a final test, we assessed whether the recovered parameters paralleled the observed behavior in the task.

Behavioral data were generated from the recovered model parameters and compared with the original behavioral data to determine whether simulated behavior from the recovered parameters could reproduce key elements of the observed experimental behavioral data. Specifically, we conducted a statistical test on the simulated data that paralleled a test conducted on the original data and examined whether it produced similar outcomes.

## Analysis plan

Mixed models were implemented in R using the *lme4* package (Version 1.1-27.1; Bates et al., 2015), and Satterthwaite's approximation was used for determining *p* values in *lmerTest* (Version 3.1-3; Kuznetsova et al., 2017). Randomization tests were implemented using *multicon* (Version 1.6; Sherman & Serfass, 2015). Marginal and conditional  $R^2$ , and semipartial  $r^2$  for linear mixed models, were estimated using *r2glmm* (Version 0.1.2; Edwards et al., 2008; Jaeger et al., 2017). Likelihood-ratio tests were performed to determine which models were best supported by the data. All mixed models included crossed random effects (Baayen et al., 2008) and both traits and participants were included as random factors. Maximal random effects were tested and were removed if unsupported by the data (i.e., low variance estimates) or if the model failed to converge with the data.

**Feedback-based differences in self-evaluations.** We conducted preliminary analyses to examine aggregate changes in self-evaluations as a function of feedback. A paired-samples *t* test was conducted on change scores (the difference in reevaluations from initial self-evaluations for each trait) from negative and positive feedback. A  $5 \times 2$  repeated measures analysis of variance (ANOVA) was conducted on self-evaluations with feedback probability as a within-participants independent variable containing five levels (90%, 70%, 50%, 30%, 10%) and time as a within-participants independent variable containing two levels (initial self-evaluation and reevaluations). Effect sizes are reported for both *t* tests and ANOVAs. As a post hoc test for a linear effect of feedback probability on self-evaluations, separate mixed models were conducted for separate response variables containing initial self-evaluations and reevaluations; trials were nested within participants, and participants and traits were included as random factors. In the models, orthogonal polynomial contrasts were applied to the five-level categorical factor consisting of each feedback probability, and the effect was allowed to vary between participants as random slopes, whereas slopes were set as fixed for traits. First-order comparisons were examined to consider the linear relationship between the five feedback probabilities and whether self-evaluations increase incrementally as the probability of positive feedback increases.

### ***Descriptive statistics for the computational model.***

An advantage of reinforcement-learning models is that parameter fits can offer insights into individual differences in how participants learn from outcomes (Yechiam et al., 2005). We computed basic summary statistics (e.g., mean, median) to examine how participants varied on their parameter fits. A nonparametric paired-samples permutation  $t$  test was conducted (100,000 permutations), because of the nonnormal distributions of learning-rate parameters, to examine within-participants differences between positive and negative learning rates. The null hypothesis is that the distributions of learning rates among positive and negative prediction errors are identical, whereas the alternative hypothesis is that the distribution of learning rates among positive prediction errors is systematically lower or higher than among negative prediction errors. For the two-tailed test, the  $p$  value was estimated as the smaller proportion of permuted mean differences that are either lower or greater than the observed mean difference.

***Trial-by-trial learning.*** We tested whether the reinforcement-learning model can effectively predict participant self-evaluations from the prior feedback presented. To avoid overfitting, we employed leave-one-participant-out cross-validation: For participant  $i$  from sample  $N$ , participant  $i$ 's free parameters were omitted and the mean of free parameters, from 1 to  $N - i$ , was determined. Mean parameters determined by the leave-one-out procedure were included in the reinforcement-learning model for participant  $i$ ; any predictability produced from the reinforcement-learning model would not be a result of participant  $i$ 's data and overfitting but, rather, from robustness of the model itself. We tested a mixed model that contained trials nested within participants, initial self-evaluation as the response variable, random slopes for social expectation for participants, and fixed slopes for traits.

Additionally, we tested whether self-evaluations can be predicted by what has been learned about neighboring, connected traits. Given that the traits are learned within a network of interconnected traits, we used neighborhood link-based features to predict self-evaluations. For each trial, we computed values for the average prediction error for all neighbors of the current trial's trait. We then used traits' average neighboring prediction errors, by the trial the trait was observed, to predict self-evaluations. Therefore, this provided a dynamic model of how social feedback is learned by neighbors and how it is learned from within a network position. Social expectation was retained in the model to determine whether neighboring traits' prediction errors explain self-evaluations above and beyond the social expectations of trait communities. In this approach, only traits' neighbors that had been observed prior to the current trait were included in the average. Thus,

the trials prior to when a trait had been observed were treated as missing data. The effects of neighbors' average prediction errors were estimated as random slopes for participants and fixed slopes for traits.

***Analysis of self-evaluation change.*** A residualized-change approach (predicting reevaluations while controlling for initial self-evaluations) was employed to test for changes in self-views. To incorporate social expectations into the change analysis, we applied the social expectation for the last trial of each of the five communities during the feedback phase as the final learned social expectations to predict reevaluations (i.e., five final learned social expectations per participant). We tested a mixed model with initial self-evaluations, out-degree centrality, and in-degree centrality entered as fixed slopes. Prediction errors and social expectations were entered as random slopes for participants and fixed slopes for traits. Both out-degree and in-degree centrality were tested as interactions with both prediction errors and social expectations. Model comparison revealed that out-degree centrality was supported as an interaction with prediction error, whereas in-degree centrality was supported as an interaction with social expectation (for model comparisons, see Table S9 in the Supplemental Material).

***Simulating replications with cross-validation.*** Although our study was well powered for the focal within-participants analyses, the participant sample was limited and therefore placed constraints on generalizability. We thus employed cross-validation to simulate replications and strengthen generalizability claims (Koul et al., 2018; Yarkoni & Westfall, 2017). Data were randomly split into 10 folds. For each of the 10 folds, a fold was held out, and the linear mixed model was fitted to the data. Using the model trained on the nine folds, we conducted out-of-sample prediction on the remaining fold. Subsequently, cross-validation was repeated with different folds across 40 iterations. We determined out-of-sample predictive accuracy by extracting root mean squared error, mean absolute error, marginal  $R^2$ , and conditional  $R^2$  for each model. As a further test of predictive accuracy, we also fitted OLS regression models predicting target from prediction. Box-and-whisker plots were qualitatively examined for dispersion of model-fit metrics across folds, whereas means were compared between models. Additionally, the mean marginal and conditional  $R^2$ s for all folds for the final model were compared with the in-sample marginal and conditional  $R^2$  to determine whether coefficients tested on out-of-sample data performed comparably with the original model.

***Individual differences and computational parameters.*** To examine how personality variables may relate to how an individual incorporates social feedback, we

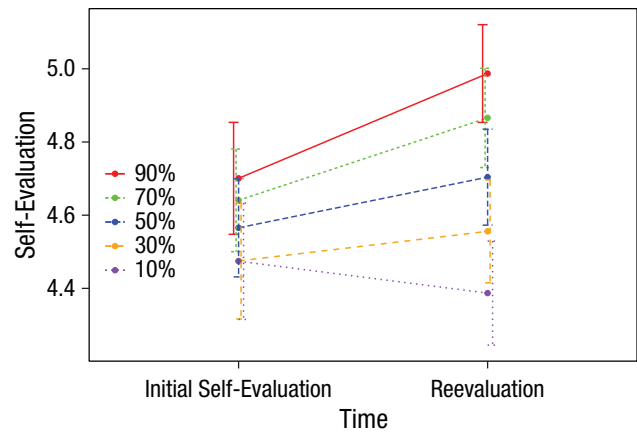
correlated all individual-differences measures with the positive learning rate and negative learning rate. Because of the large number of individual differences and parameters to test, we employed randomization tests and adjusted correlational  $p$  values using the Benjamini and Hochberg (1995) false discovery rate (FDR) procedure to control inflation of the false positive rate resulting from multiple comparisons. Because of the limited sample size and demographic range, the generalizability of between-participants inferences should be approached with caution.

## Results

### *Positive self-evaluations in response to aggregate feedback*

We first examined whether participants differed in self-evaluations because of broad feedback categories such as the valence of feedback and the probability of positive feedback by trait community. Preliminary analyses revealed that participants' self-evaluations varied as a function of feedback. A paired-samples  $t$  test on valence of feedback (positive vs. negative) revealed significant differences in change scores between the two types of feedback; from initial self-evaluations to reevaluations, participants changed their self-evaluations more after positive ( $M = 0.31$ ,  $SD = 0.39$ ) than negative ( $M = -0.07$ ,  $SD = 0.49$ ) feedback,  $t(45) = 6.60$ ,  $p < .001$ ,  $d = 0.973$ . This finding broadly supports the hypothesis that people are positively biased in how they incorporate social feedback.

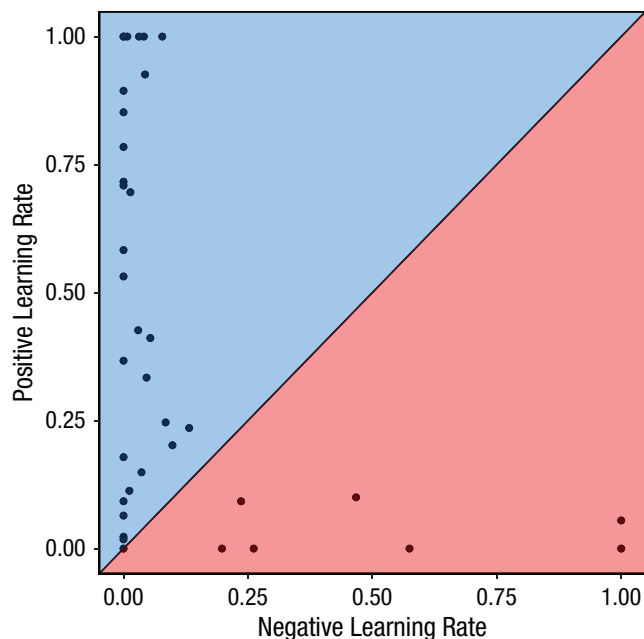
The  $5 \times 2$  repeated measures ANOVA tested for feedback probability-based differences in self-evaluations as well as changes in self-evaluations across time. The analysis revealed a main effect of feedback probability,  $F(4, 180) = 2.69$ ,  $p = .033$ ,  $\eta_p^2 = .06$ ; participants differentiated their self-evaluations according to the feedback probabilities of the trait community. Additionally, there was a main effect of time,  $F(1, 45) = 4.86$ ,  $p = .033$ ,  $\eta_p^2 = .10$ , reflecting that self-evaluations changed from initial self-evaluations to reevaluations. Finally, we also identified an interaction between time and feedback probability,  $F(4, 180) = 8.22$ ,  $p < .001$ ,  $\eta_p^2 = .16$ , suggesting that self-evaluations were differentiated across time more for particular probabilities than others. Consistent with this, results showed that the 90% probability community was the most differentiated of all probability groups, which suggests that participants may have been most sensitive to feedback when it was largely positive. Meanwhile, ratings from initial self-evaluation to reevaluation decreased only for the 10% probability group, which suggests that participants were resistant to incorporating negative feedback on traits unless it was nearly entirely



**Fig. 3.** Mean self-evaluation as a function of time and probability of positive feedback. Error bars represent standard errors. Figure S3 in the Supplemental Material available online shows the same analysis using raw data.

negative (see Fig. 3; for visualizations with raw data, see Figs. S3 and S13a, respectively, in the Supplemental Material).

We conducted further tests to examine how these feedback probabilities were differentiated, as the means between feedback probabilities revealed stepwise increases with probability order. Indeed, a post hoc mixed model employing first-order orthogonal polynomial comparisons demonstrated a linear effect; initial self-evaluations were higher for traits assigned to higher probability groups ( $\beta = 0.12$ ,  $SE = 0.06$ , 95% confidence interval [CI] = [0.007, 0.230]),  $t(45) = 2.14$ ,  $p = .038$ , semipartial  $r^2 = .003$ . Whereas participants experienced the social feedback, they appeared to be sensitive to the different feedback probabilities and began to align their self-evaluations in an order consistent with the feedback probabilities. This linear effect was stronger at reevaluations, after all feedback had been learned. The post hoc mixed model testing for first-order orthogonal polynomial comparisons revealed a strong linear effect between probabilities at reevaluations ( $\beta = 0.29$ ,  $SE = 0.06$ , 95% CI = [0.165, 0.407]),  $t(44) = 4.75$ ,  $p < .001$ , semipartial  $r^2 = .02$ . Tests revealed that self-evaluations were becoming differentiated by probability as early as initial self-evaluations and that the differentiation became greater by the reevaluation phase (for summary statistics and for the ANOVA table, see Tables S4 and S5, respectively, in the Supplemental Material). The interaction between time and feedback probability demonstrates that participants expressed a positivity bias in both their receptivity to positive feedback and their nearly entire dismissal of negative feedback, but this interaction does not allow inferences regarding the underlying processes.



**Fig. 4.** Scatterplot showing the relation between the positive and negative learning rate for each participant. The blue triangle highlights participants with greater positive than negative learning rates (i.e., positively biased), whereas the red triangle highlights participants with greater negative than positive learning rates (i.e., negatively biased).

### ***Asymmetries in self-related learning***

To assess how people learn from feedback, we fitted reinforcement-learning models to each participant's responses. The best-performing model contained separate learning rates for trials in which the prediction error was either positive or negative. A permutation-based paired-samples *t* test shows that fitted positive ( $M = 0.50$ ,  $Mdn = 0.42$ ,  $SD = 0.41$ ) and negative ( $M = 0.12$ ,  $Mdn = 0$ ,  $SD = 0.26$ ) learning rates per participant were significantly different from one another (observed difference = .38, mean permuted difference = .0003,  $p < .001$ ). The parameter fits reflect that people learn incrementally about positive outcomes related to the self but neglect negative outcomes related to the self. This serves as further evidence of the prevalent motivation for positivity (for a scatterplot showing the relation between positive and negative learning rates, see Fig. 4; for a correlation matrix and histograms of parameters, see Figs. S4a and S5, respectively, in the Supplemental Material).

### ***Social expectations and neighboring prediction errors predict self-evaluations***

As a proof of concept of the validity of the model, we evaluated how the model accounts for trial-by-trial

self-evaluations and tested whether trial-by-trial community-based expectations (i.e., social expectations) predicted self-evaluations, reflecting a global approach (i.e., groups of traits) to characterizing self-evaluations. Secondly, we leveraged the network structure to test a more local approach (i.e., neighbors of traits) by interpolating the average prediction errors of network neighbors. In the test of the model, community-based social expectations significantly predicted trial-by-trial self-evaluations ( $\beta = 0.09$ ,  $SE = 0.02$ , 95% CI = [0.046, 0.130]),  $t(48) = 4.21$ ,  $p < .001$ , semipartial  $r^2 = .009$ . Given that there is substantial variability (participant:  $\sigma^2 = .17$ , trait:  $\sigma^2 = .10$ ) in trait self-evaluations and that this effect can be observed only insofar as participants can infer the trait structural relationships and the associated feedback, this effect size is fairly substantial. Findings suggest that as participants learned about how they were perceived on different trait communities from feedback, they began to dynamically shift their self-evaluations to match their learned expectations. However, they incorporated prior feedback into expectations in an asymmetrical manner, resulting in generally favorable expectations and positively skewed updating of self-evaluations across learning (for model parameters, for visualizations of predicted effects with raw data, see Table S6 and Figs. S6 and S13b, respectively, in the Supplemental Material).

In addition to the more global analysis examining how community social expectations predict participant self-evaluations, we explored how prediction errors among local network connections (i.e., neighbors) may uncover additional information about participants' self-evaluations. We found that prediction errors among neighbors ( $\beta = 0.09$ ,  $SE = 0.02$ , 95% CI = [0.057, 0.118]),  $t(44) = 5.78$ ,  $p < .001$ , semipartial  $r^2 = .009$ , strongly predicted self-evaluations. Furthermore, after these neighboring features were incorporated, the trial-by-trial effect of community social expectation remained a strong predictor of self-evaluations ( $\beta = 0.07$ ,  $SE = 0.02$ , 95% CI = [0.029, 0.111]),  $t(48) = 3.48$ ,  $p = .001$ , semipartial  $r^2 = .006$ , suggesting that both neighbor- and community-level learned information contributes to learning and shape self-evaluations. Both the global, conventional approach and the local, averaged prediction errors of neighbors' approach offer useful tests of the reinforcement-learning model reflecting how social feedback may be learned at different levels of information processing. Results indicate that participants used not only community-level information about a trait's expected feedback (i.e., social expectation) in forming self-evaluations but also local information about how feedback from neighboring traits in the network deviates from expectation (i.e., prediction error). Self-evaluations may be updated to maintain consistency between

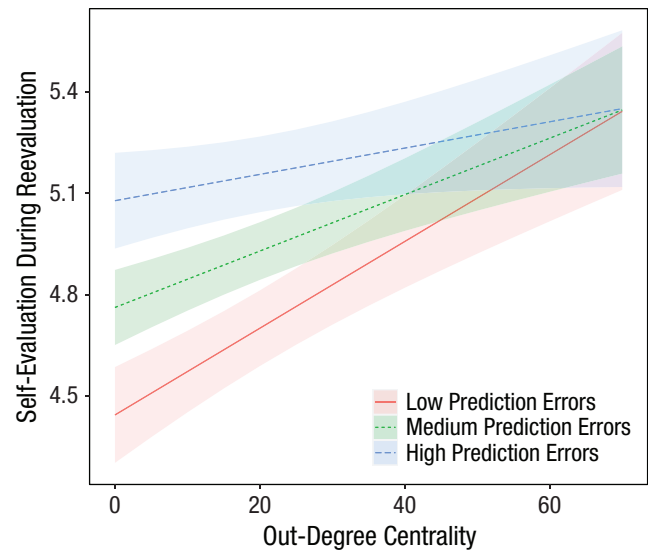


neighbors and coherence among the broader network of interconnected traits. Cross-validation performed on the trial-by-trial learning mixed models provides evidence of out-of-sample generalizability and indicates that overfitting was avoided (for cross-validation details, see Table S7, Fig. S7, and Supplementary Text in the Supplemental Material).

We additionally tested whether the feedback previously experienced and expectations formed during learning predicted later self-evaluations after learning. Prediction errors ( $\beta = 0.13$ ,  $SE = 0.02$ , 95% CI = [0.088, 0.177]),  $t(28) = 6.01$ ,  $p < .001$ , semipartial  $r^2 = .03$ , and learned social expectations ( $\beta = 0.21$ ,  $SE = 0.03$ , 95% CI = [0.148, 0.271]),  $t(30) = 6.90$ ,  $p < .001$ , semipartial  $r^2 = .06$ , predicted participant reevaluations in models controlling for initial self-evaluations. Therefore, participants used both the previously observed trait-specific prediction errors and their learned community-based social expectations to inform their updated self-evaluations. Consistent with the analysis of trial-by-trial learning, results showed that changes in self-evaluations after learning were also influenced by expectations at the local (i.e., trait-level prediction errors) and global (i.e., community-level social expectations) levels of the network (for main effects of residualized-change-model parameters, see Table S8 in the Supplemental Material).

### Self-evaluation updates depend on network features

Next, we tested how network structure imposes constraints on how people learn about themselves from social feedback and incorporate prediction errors and social expectations. To this end, we examined how out-degree and in-degree centrality impacted the malleability of traits from their initial to final evaluations. The effect of prediction errors on trait reevaluations interacted with out-degree centrality ( $\beta = -0.03$ ,  $SE = 0.01$ , 95% CI = [-0.051, -0.016]),  $t(6393) = -3.73$ ,  $p < .001$ , semipartial  $r^2 = .002$ ; self-evaluations for traits with greater out-degree centrality changed less from prediction errors after participants learned from social feedback (for a visualization of the out-degree centrality interaction, see Fig. 5; for visualizations with raw data, see Figs. S8 and S13c, respectively, in the Supplemental Material). To the extent that a trait has more dependencies, participants updated self-evaluations less on the basis of prediction errors, which reflects one way that people maintain coherence in the self-concept. The degree to which a trait is essential to self-concept coherence (on the basis of its causal relationships) may make it less susceptible to change. Further, the effect



**Fig. 5.** Self-evaluations during reevaluation as a function of out-degree centrality and prediction error. Predicted values (i.e., estimated marginal effects) were obtained from residualized-change-reevaluation models, which held covariates constant. Error bands show confidence intervals of  $\pm 1.96$ . Values reported on the y-axis were obtained from models controlling for initial self-evaluations. Figure S8 in the Supplemental Material available online shows the same analysis using raw data.

of social expectation on trait reevaluations interacted with in-degree centrality ( $\beta = 0.01$ ,  $SE = 0.01$ , 95% CI = [0.001, 0.037]),  $t(6388) = 2.12$ ,  $p = .03$ , semipartial  $r^2 = .001$ ; the reinforcement-learning model was better at predicting self-evaluations for higher in-degree traits than lower in-degree traits. This suggests that individuals may maintain coherence by modifying self-views on traits that are most susceptible to influence and most contingent on other self-views (for a visualization of the in-degree centrality interaction and for visualizations with raw data, see Figs. S9, S10, and S13d, respectively, in the Supplemental Material). Furthermore, it supports the causal, directed structure of the network because in-degree and out-degree centrality demonstrate different interaction patterns (for centrality-interaction model parameters and model comparisons, see Table S9). We again simulated replications with cross-validation, which supported the out-of-sample performance of the model and suggests that inadvertent overfitting was avoided (for cross-validation details, see Table S10 and Fig. S11 in the Supplemental Material).

### Parameter recovery

“True” parameters randomly generated from simulations and parameters fitted to the simulated behavioral data were correlated with each other to estimate the extent

to which parameters were recoverable. High correlations should reflect that behavioral data generated by a particular set of parameters will also produce parameter fits consistent with the behavioral data. Parameters, particularly the learning-rate parameters, were determined to be recoverable because fitted parameters were strongly correlated with simulated parameters:  $\alpha_p = .95$ ,  $\alpha_n = .91$ ,  $\phi = .61$ , and  $\beta = 0.74$  (for the correlation matrix, see Fig. S4b).

However, because these data were randomly generated, the previous test of parameter recovery assumed independence between parameters that was not observed in the original parameters. To more effectively address the recoverability of parameters that share the original covariance structure, we used the original participant parameters to simulate new behavioral data to which we fitted the model parameters. We correlated the original parameters with the fitted (simulated) parameters to test recoverability. Again, we found that parameters were determined to be recoverable:  $\alpha_p = .95$ ,  $\alpha_n = .97$ ,  $\phi = .91$ , and  $\beta = 0.62$  (for the correlation matrix, see Fig. S4c). Decay appears to benefit the most in terms of recoverability when we consider the covariance structure and nonindependence of parameters. Feedback sensitivity exhibits the lowest recoverability and may be a noisier parameter to estimate in general.

Lastly, we generated behavioral data from the recovered participant parameters in order to test the extent to which these parameters reproduce key elements of the experimental behavioral data. We conducted a time-by-feedback-probability within-participants ANOVA, identical to the ANOVA described above, on the simulated data. We identified a main effect of feedback probability,  $F(4, 180) = 28.59$ ,  $p < .001$ ,  $\eta_p^2 = .98$ , supporting the prediction that simulated behavior discriminates between feedback probabilities as observed behavior does. We also identified a main effect of time,  $F(1, 45) = 12.45$ ,  $p < .001$ ,  $\eta_p^2 = .21$ , which suggests that simulated behavior produces differences from initial self-evaluations to reevaluations, paralleling the observed data. Unlike in the original behavioral data, we did not observe a time-by-feedback-probability interaction,  $F(4, 180) = 1.83$ ,  $p = .12$ ,  $\eta_p^2 = .04$ , which suggests that the recovered parameters may not fully capture how responses change more for particular feedback probabilities than others between learning and reevaluation. However, this is unsurprising because the model stops learning when it stops receiving feedback and thus cannot, in its present form, account for any changes in self-evaluations prior to reevaluation that do not occur in response to immediate feedback.

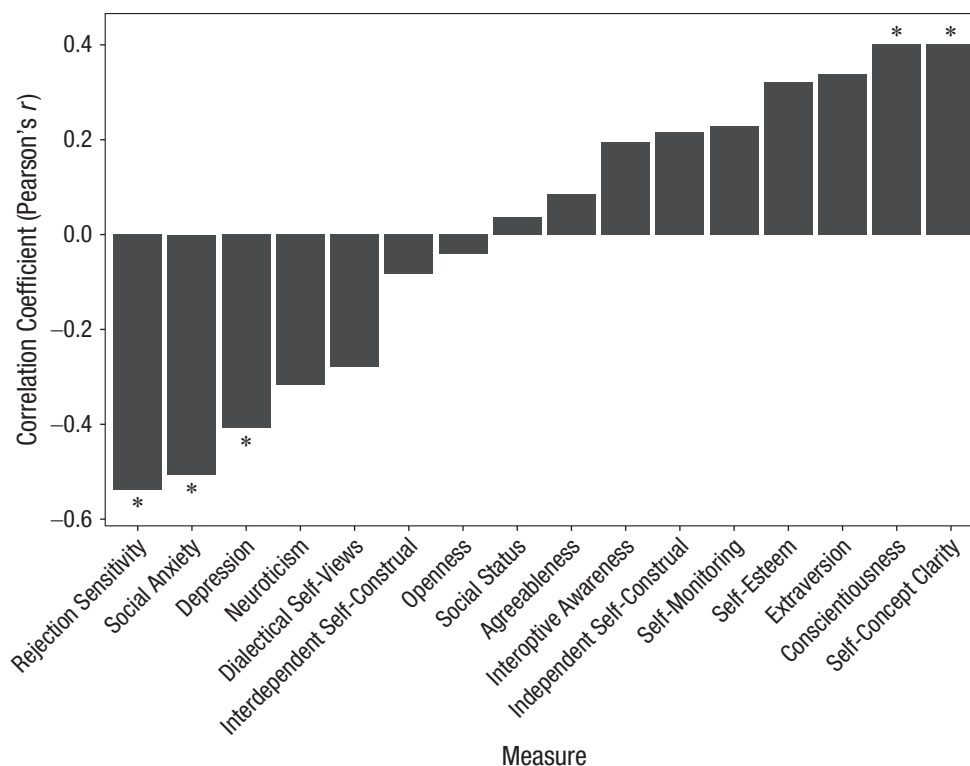
Altogether, model parameters are recoverable, and behavior produced by recovered parameters parallels observed participant behavior. It is important to note

that our fitted and simulated reinforcement-learning agents represented trait values only at the community level and assumed that these are uniform across the community (see the Discussion section), and thus our simulated data do not contain subcommunity information (immediate neighbors, trait centrality) from the independently constructed network portion of the model.

### ***Individual differences in learning from positive feedback***

To test the extent to which asymmetrical learning of social feedback reflects individual differences, we correlated learning-rate parameters estimated during model fitting with all individual differences collected through self-report. Randomization-tested correlations with positive learning rates revealed an average observed correlational effect size ( $r = .28$ ) greater than the expected average absolute correlational effect size under the null hypothesis ( $r = .13$ ) and a quantity of significant associations (eight) exceeding the expected number significant under the null hypothesis (0.80). Negative learning rates revealed an average observed correlational effect size ( $r = .15$ ) slightly greater than the expected average absolute correlational effect size under the null hypothesis ( $r = .13$ ) and a quantity of significant associations (one) slightly exceeding the expected number significant under the null hypothesis (0.78). Correlations between decay and feedback-sensitivity parameters were likely to be spurious and were omitted from the primary analyses (for all correlations, see Table S11 in the Supplemental Material).

After FDR correction, associations were significant for self-concept clarity ( $r = .40$ ,  $p_{\text{FDR}} = .022$ ,  $p = .007$ ), conscientiousness ( $r = .40$ ,  $p_{\text{FDR}} = .022$ ,  $p = .007$ ), social anxiety ( $r = -.51$ ,  $p_{\text{FDR}} < .001$ ,  $p < .001$ ), depressive symptoms ( $r = -.41$ ,  $p_{\text{FDR}} = .022$ ,  $p = .006$ ), and rejection sensitivity ( $r = -.54$ ,  $p_{\text{FDR}} < .001$ ,  $p < .001$ ). Extraversion ( $r = .34$ ,  $p_{\text{FDR}} = .065$ ,  $p = .024$ ), self-esteem ( $r = .32$ ,  $p_{\text{FDR}} = .034$ ,  $p = .033$ ), and neuroticism ( $r = -.32$ ,  $p_{\text{FDR}} = .073$ ,  $p = .036$ ) revealed uncorrected associations ( $p < .05$ ; positive-learning-rate correlations are displayed in Fig. 6). After FDR correction, no significant associations emerged for negative learning rate, but uncorrected, depressive symptoms were significantly associated with negative learning rate ( $r = .36$ ,  $p_{\text{FDR}} = .247$ ,  $p = .015$ ). Thus, individual differences reflective of psychological maladjustment and mental health dysfunction were negatively correlated with the positive learning rate (i.e., modifying self-views in response to positive prediction errors). The negative-learning-rate correlation was the converse of this, where depressive symptoms were positively correlated with negative learning rate



**Fig. 6.** Correlations between positive learning rates and individual-differences measures. Correlations are displayed in ascending order. Asterisks represent significant differences from zero ( $p < .05$ , false-discovery-rate corrected). Figure S12 in the Supplemental Material available online shows scatterplots depicting associations with  $p < .10$ , uncorrected. Table S11 in the Supplemental Material shows all correlations.

(i.e., modifying self-views in response to negative prediction errors).

## Discussion

We tested a network model of the self-concept embedded within a reinforcement-learning framework to examine how people update self-representations on the basis of social feedback and how this updating process is constrained by drives for self-concept positivity and coherence. Consistent with the drive to achieve positive self-views, results showed that participants learned more from positive feedback than negative feedback, which was largely neglected. Consistent with the drive to maintain coherence, results showed that prediction errors were less likely to cause updates for traits with higher out-degree centrality. These results reveal how people simultaneously maintain positivity and coherence of self-representations and shed light on the specific computational mechanisms underlying these processes. Together, the integration of network-analysis and reinforcement-learning approaches conferred combined insights into how motivations shape learning about the self.

We provide a mechanistic account of how people accomplish self-concept positivity. Past work has shown that people are positively biased in their self-evaluations (Hughes & Zaki, 2015; Sharot & Garrett, 2016), particularly in response to social feedback (Alicke & Sedikides, 2009; Hughes & Beer, 2013; Korn et al., 2012; Somerville et al., 2010). Here, we showed that people accomplish this by learning asymmetrically from positive over negative prediction errors. This mirrors asymmetrical learning observed across other contexts (Dorfman et al., 2019; Gershman, 2015; Niv et al., 2012), including positivity biases in learning about future outcomes (Lefebvre et al., 2017). Conversely, other distinct but related research on self-learning suggests that people are negatively biased when learning about their own task performance (Müller-Pinzler et al., 2019). It may be adaptive to be positively biased in learning about specific trait self-beliefs (i.e., Who am I?), while also adaptive to be negatively biased in learning about task performance that can improve (i.e., How did I perform?). Consistent with a potential adaptive function underlying positively biased learning about trait self-beliefs, our results showed a multitude of individual

differences implicated in positive self-views and adjustment that are associated with positive learning rate, such as self-esteem, rejection sensitivity, depressive symptomatology, and social anxiety. This aligns with ongoing work characterizing the computational underpinnings of aberrant self-relevant learning associated with clinical symptomatology (Will et al., 2017, 2020). For example, depressed individuals learn more from negative feedback relative to healthy controls (Garrett et al., 2014), and socially anxious individuals exhibit greater sensitivity to negative feedback (Hopkins et al., 2021).

Structural features of the self-concept and the motivation for coherence impose additional constraints on its malleability and how people update self-views. The prediction errors of neighboring traits in the network predict self-evaluations for a given trait, providing evidence that structural influences at different levels of organization shape self-concept learning and coherence maintenance. Moreover, self-views for more out-degree central traits were less likely to change on the basis of prediction errors. If out-degree central traits were modified, they may contradict the other, many self-views that depend on them. Thus, traits with higher out-degree centrality may require more or stronger evidence in order to destabilize traits that are important to coherence. It may not simply be undirected similarity to other traits that is integral to how people update self-views but, instead, coherence-preserving relationships that depend on causal structure to constrain malleability.

Although prior research has used reinforcement learning to understand social feedback learning (Hackel & Amodio, 2018; Lockwood & Klein-Flügge, 2021), this was the first application of reinforcement learning to examine how individuals learn from latent self-related feedback, on the basis of semantic associations between traits. Prior studies have used explicitly labeled cues to observe how people update behavior according to reward payout. For example, recent work examined how participants update self-esteem ratings as a function of approval from social groups, where each group's approval tendency was denoted by different colors (Will et al., 2017). In contrast, participants in our study learned other people's perceptions of them on traits from the relationships between previously observed traits. For this to occur, people must be sensitive to the structure of the network. Thus, self-evaluations are influenced not only by the feedback received for a particular trait but also by updates for nearby traits in the same community and updates for neighboring, adjacent traits. This network framework thus enables new insights into social learning by examining how latent mental constructs (e.g., other people's perceptions on groups of interconnected traits) are learned from social

feedback within a structured concept space. For example, this approach could be extended to understand how people learn about social groups that emphasize different communities of traits.

Our novel combination of reinforcement-learning and network approaches presents a computational theory of self-concept malleability and coherence during social experience. We showed that learning occurs from local network structure (e.g., neighbors) above and beyond the global network structure (e.g., communities) used for our experimental manipulation. In future studies, researchers should incorporate finer grained structural relations in models when examining self-related learning about trait values and consider how feedback travels across a structure via causal associations (Vaidya & Badre, 2020; Wu et al., 2021). Additionally, researchers using larger and more representative samples may examine mental health and well-being outcomes related to self-relevant learning to develop strategies that increase adjustment. Finally, we acknowledge that the self-concept is composed of more than the trait words used in our model, and future work can expand the network's scope beyond traits to include other facets of the self-concept (e.g., social roles, group membership, specific behaviors). We believe that this approach may allow a deeper understanding of how motivations interact in shaping self-perceptions and the myriad conditions by which the self-concept may remain stable or change across time.

## Transparency

*Action Editor:* Leah Somerville

*Editor:* Patricia J. Bauer

### Author Contributions

All the authors contributed to the study concept and design. J. Elder programmed the experimental task, supervised data collection, cleaned and analyzed the data, and drafted the manuscript. B. L. Hughes and T. Davis provided critical revisions. All the authors approved the final manuscript for submission.

### Declaration of Conflicting Interests

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

### Open Practices


Deidentified data, analysis scripts, and materials have been made publicly available via OSF and can be accessed at <https://osf.io/j6bkc/>. The design and analysis plan for the study were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.





## ORCID iDs

Jacob Elder  <https://orcid.org/0000-0002-5305-7006>

Brent L. Hughes  <https://orcid.org/0000-0001-8732-8727>

## Acknowledgments

We thank Samuel Gershman (Harvard University) for his input on the computational modeling. We thank the research assistants, Leanne Esconde, Bryant Ma, James Sobrino, and Julia Hopkins, for their assistance with data collection. We thank Bernice Cheung (University of Oregon) for her contributions in pilot testing and developing the network.

## Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976211045934>

## Note

1. In the final data set, repeated measures correlations were .493 for initial self-evaluations and .411 for reevaluations.

## References

- Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, 20(1), 1–48. <https://doi.org/10.1080/10463280802613866>
- Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences*, 18(6), 279–292. <https://doi.org/10.1016/j.tics.2014.03.002>
- Anderson, N. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, 9(3), 272–279. <https://doi.org/10.1037/h0025907>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- 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](https://doi.org/10.1207/s15327752jpa6703_13)
- Beer, J. S., Chester, D. S., & Hughes, B. L. (2013). Social threat and cognitive load magnify self-enhancement and attenuate self-deprecation. *Journal of Experimental Social Psychology*, 49(4), 706–711. <https://doi.org/10.1016/j.jesp.2013.02.017>
- Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60024-6](https://doi.org/10.1016/S0065-2601(08)60024-6)
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436. <https://doi.org/10.1163/156856897X00357>
- Campbell, J. D., Trapnell, P. D., Heine, S. J., Katz, I. M., Lavallee, L. F., & Lehman, D. R. (1996). Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology*, 70(1), 141–156. <https://doi.org/10.1037/0022-3514.70.1.141>
- Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance* (Vol. 23, pp. 3–38). Oxford University Press.
- Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, 30(4), 516–525. <https://doi.org/10.1177/0956797619828724>
- Downey, G., & Feldman, S. I. (1996). Implications of rejection sensitivity for intimate relationships. *Journal of Personality and Social Psychology*, 70(6), 1327–1343. <https://doi.org/10.1037/0022-3514.70.6.1327>
- Dunning, D. (1999). A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry*, 10(1), 1–11. [https://doi.org/10.1207/s15327965pli1001\\_1](https://doi.org/10.1207/s15327965pli1001_1)
- Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An  $R^2$  statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, 27(29), 6137–6157. <https://doi.org/10.1002/sim.3429>
- Eisenberger, N. I., Inagaki, T. K., Muscatell, K. A., Byrne Haltom, K. E., & Leary, M. R. (2011). The neural sociometer: Brain mechanisms underlying state self-esteem. *Journal of Cognitive Neuroscience*, 23(11), 3448–3455. [https://doi.org/10.1162/jocn\\_a\\_00027](https://doi.org/10.1162/jocn_a_00027)
- Elder, J., Cheung, B., Davis, T., & Hughes, B. L. (2022). *Mapping the self: A network approach for understanding psychological and neural representations of self-concept structure*. PsyArXiv. <https://doi.org/10.31234/osf.io/hj87w>
- Frey, D., Stahlberg, D., & Fries, A. (1986). Information seeking of high- and low-anxiety subjects after receiving positive and negative self-relevant feedback. *Journal of Personality*, 54(4), 694–703. <https://doi.org/10.1111/j.1467-6494.1986.tb00420.x>
- Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Frontiers in Human Neuroscience*, 8, Article 639. <https://doi.org/10.3389/fnhum.2014.00639>
- Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, 22(5), 1320–1327. <https://doi.org/10.3758/s13423-014-0790-3>
- Greenberg, J., & Pyszczynski, T. (1985). Compensatory self-inflation: A response to the threat to self-regard of public failure. *Journal of Personality and Social Psychology*, 49(1), 273–280. <https://doi.org/10.1037/0022-3514.49.1.273>

- Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, 24, 92–97. <https://doi.org/10.1016/j.copsyc.2018.09.001>
- Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social-desirability values for 573 personality terms. *European Journal of Personality*, 1(4), 241–258. <https://doi.org/10.1002/per.2410010405>
- Hepper, E., Hart, C. M., Gregg, A. P., & Sedikides, C. (2011). Motivated expectations of positive feedback in social interactions. *The Journal of Social Psychology*, 151(4), 455–477. <https://doi.org/10.1080/00224545.2010.503722>
- Hepper, E. G., & Sedikides, C. (2012). Self-enhancing feedback. In R. M. Sutton, M. J. Hornsey, & K. M. Douglas (Eds.), *Feedback: The communication of praise, criticism, and advice* (pp. 43–56). Peter Lang.
- 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*, 5(1), 21–37. <https://doi.org/10.5334/cpsy.57>
- Hughes, B. L., & Beer, J. S. (2013). Protecting the self: The effect of social-evaluative threat on neural representations of self. *Journal of Cognitive Neuroscience*, 25(4), 613–622. [https://doi.org/10.1162/jocn\\_a\\_00343](https://doi.org/10.1162/jocn_a_00343)
- Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, 19(2), 62–64. <https://doi.org/10.1016/j.tics.2014.12.006>
- Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An  $R^2$  statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, 44(6), 1086–1105. <https://doi.org/10.1080/02664763.2016.1193725>
- Kirby, D. M., & Gardner, R. C. (1972). Ethnic stereotypes: Norms on 208 words typically used in their assessment. *Canadian Journal of Psychology*, 26(2), 140–154. <https://doi.org/10.1037/h0082423>
- 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*, 17(8), 1144–1155. <https://doi.org/10.1037/emo0000296>
- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience*, 32(47), 16832–16844. <https://doi.org/10.1523/JNEUROSCI.3016-12.2012>
- Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, 9, Article 1117. <https://doi.org/10.3389/fpsyg.2018.01117>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(1). <https://doi.org/10.18637/jss.v082.i13>
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4), Article 0067. <https://doi.org/10.1038/s41562-017-0067>
- Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry*, 22, 141–173. <https://doi.org/10.1159/000414022>
- Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is there a ‘social’ brain? Implementations and algorithms. *Trends in Cognitive Sciences*, 24(10), 802–813. <https://doi.org/10.1016/j.tics.2020.06.011>
- Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences, USA*, 113(35), 9763–9768.
- Lockwood, P. L., & Klein-Flügge, M. C. (2021). Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, 16(8), 761–771. <https://doi.org/10.1093/scan/nsaa040>
- Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology*, 38, 299–337. <https://doi.org/10.1146/annurev.ps.38.020187.001503>
- Mead, G. H. (1934). *Mind, self, and society from the standpoint of a social behaviorist*. University of Chicago Press.
- Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The Multidimensional Assessment of Interoceptive Awareness, Version 2 (MAIA-2). *PLOS ONE*, 13(12), Article e0208034. <https://doi.org/10.1371/journal.pone.0208034>
- 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), Article 14416. <https://doi.org/10.1038/s41598-019-50821-w>
- Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: Optimx for R. *Journal of Statistical Software*, 43(1). <https://doi.org/10.18637/jss.v043.i09>
- Niv, Y., Edlund, J. A., Dayan, P., & O’Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, 32(2), 551–562. <https://doi.org/10.1523/JNEUROSCI.5498-10.2012>
- Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, 13(8), Article e1005684. <https://doi.org/10.1371/journal.pcbi.1005684>
- Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In P. Yolum, T. Güngör, F. Gürgeç, & C. Özturan (Eds.), *Computer and Information Sciences - ISCIS 2005* (pp. 284–293). Springer. [https://doi.org/10.1007/11569596\\_31](https://doi.org/10.1007/11569596_31)
- Pyszczynski, T., Greenberg, J., & LaPrelle, J. (1985). Social comparison after success and failure: Biased search for information consistent with a self-serving conclusion. *Journal of Experimental Social Psychology*, 21(2), 195–211. [https://doi.org/10.1016/0022-1031\(85\)90015-0](https://doi.org/10.1016/0022-1031(85)90015-0)
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement

- and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64–99). Appleton-Century-Crofts.
- Rodman, A. M., Powers, K. E., & Somerville, L. H. (2017). Development of self-protective biases in response to social evaluative feedback. *Proceedings of the National Academy of Sciences, USA*, 114(50), 13158–13163. <https://doi.org/10.1073/pnas.1712398114>
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.
- Schröder-Abé, M., Rudolph, A., Wiesner, A., & Schütz, A. (2007). Self-esteem discrepancies and defensive reactions to social feedback. *International Journal of Psychology*, 42(3), 174–183. <https://doi.org/10.1080/00207590601068134>
- Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, 20(1), 25–33. <https://doi.org/10.1016/j.tics.2015.11.002>
- Sherman, R. A., & Serfass, D. G. (2015). The comprehensive approach to analyzing multivariate constructs. *Journal of Research in Personality*, 54, 40–50. <https://doi.org/10.1016/j.jrp.2014.05.002>
- Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. *Personality and Social Psychology Bulletin*, 20(5), 580–591. <https://doi.org/10.1177/0146167294205014>
- Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22(2), 189–228. [https://doi.org/10.1207/s15516709cog2202\\_2](https://doi.org/10.1207/s15516709cog2202_2)
- Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, 30(4), 526–537. <https://doi.org/10.1037/h0037039>
- Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2010). Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cerebral Cortex*, 20(12), 3005–3013. <https://doi.org/10.1093/cercor/bhq049>
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143. <https://doi.org/10.1037/pspp0000096>
- Spencer-Rodgers, J., Srivastava, S., Boucher, H. C., English, T., Paletz, S. B., & Peng, K. (2015). The dialectical self scale [Unpublished manuscript].
- Swann, W. B., & Hill, C. A. (1982). When our identities are mistaken: Reaffirming self-conceptions through social interaction. *Journal of Personality and Social Psychology*, 43(1), 59–66. <https://doi.org/10.1037/0022-3514.43.1.59>
- Swann, W. B., Rentfrow, P. J., & Guinn, J. S. (2003). Self-verification: The search for coherence. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 367–383). Guilford Press.
- Swann, W. B., Wenzlaff, R. M., & Tafarodi, R. W. (1992). Depression and the search for negative evaluations: More evidence of the role of self-verification strivings. *Journal of Abnormal Psychology*, 101(2), 314–317. <https://doi.org/10.1037/0021-843X.101.2.314>
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103(2), 193–210. <https://doi.org/10.1037/0033-2909.103.2.193>
- Vaidya, A. R., & Badre, D. (2020). Neural systems for memory-based value judgment and decision-making. *Journal of Cognitive Neuroscience*, 32(10), 1896–1923. [https://doi.org/10.1162/jocn\\_a\\_01595](https://doi.org/10.1162/jocn_a_01595)
- Vandellen, M. R., Campbell, W. K., Hoyle, R. H., & Bradfield, E. K. (2010). Compensating, resisting, and breaking: A meta-analytic examination of reactions to self-esteem threat. *Personality and Social Psychology Review*, 15(1), 51–74. <https://doi.org/10.1177/1088868310372950>
- Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>
- Will, G.-J., Moutoussis, M., Womack, P. M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., & Jones, P. B., NSPN Consortium, Rutledge, R. B., & Dolan, R. J. (2020). Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Translational Psychiatry*, 10(1), Article 96. <https://doi.org/10.1038/s41398-020-0702-4>
- Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *eLife*, 6, Article e28098. <https://doi.org/10.7554/eLife.28098>
- Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8, Article e49547. <https://doi.org/10.7554/eLife.49547>
- Wu, C. M., Schulz, E., & Gershman, S. J. (2021). Inference and search on graph-structured spaces. *Computational Brain & Behavior*, 4(4), 125–147. <https://doi.org/10.1007/s42113-020-00091-x>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science*, 16(12), 973–978. <https://doi.org/10.1111/j.1467-9280.2005.01646.x>
- 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>