Behavioral/Cognitive

A Fluid Self-Concept: How the Brain Maintains Coherence and Positivity across an Interconnected Self-Concept While Incorporating Social Feedback

Jacob J. Elder,[1] Tyler H. Davis,[2] and Brent L. Hughes[1]
[1]Department of Psychology, University of California, Riverside, Riverside, California 92521 and [2]Independent researcher, Fremont, California

People experience instances of social feedback as interdependent with potential implications for their entire self-concept. How do people maintain positivity and coherence across the self-concept while updating self-views from feedback? We present a network model describing how the brain represents the semantic dependency relations among traits and uses this information to avoid an overall loss of positivity and coherence. Both male and female human participants received social feedback during a self-evaluation task while undergoing functional magnetic resonance imaging. We modeled self-belief updating by incorporating a reinforcement learning model within the network structure. Participants learned more rapidly from positive than negative feedback and were less likely to change self-views for traits with more dependencies in the network. Further, participants back propagated feedback across network relations while retrieving prior feedback on the basis of network similarity to inform ongoing self-views. Activation in ventromedial prefrontal cortex (vmPFC) reflected the constrained updating process such that positive feedback led to higher activation and negative feedback to less activation for traits with more dependencies. Additionally, vmPFC was associated with the novelty of a trait relative to previously self-evaluated traits in the network, and angular gyrus was associated with greater certainty for self-beliefs given the relevance of prior feedback. We propose that neural computations that selectively enhance or attenuate social feedback and retrieve past relevant experiences to guide ongoing self-evaluations may support an overall positive and coherent self-concept.

Key words: network analysis; prefrontal cortex; reinforcement learning; self-concept; semantics; social feedback

Significance Statement

We humans experience social feedback throughout our lives, but we do not dispassionately incorporate feedback into our self-concept. The implications of feedback for our entire self-concept plays a role in how we either change or retain our prior self-beliefs. In a neuroimaging study, we find that people are less likely to change their beliefs from feedback when the feedback has broader implications for the self-concept. This resistance to change is reflected in processing in the ventromedial prefrontal cortex, a region that is central to self-referential and social cognition. These results are broadly applicable given the role that maintaining a positive and coherent self-concept plays in promoting mental health and development throughout the lifespan.

Introduction

People maintain complex and multifaceted self-views (Markus and Wurf, 1987) and dynamically learn about themselves from feedback they receive through interactions with others and their environment. However, people experience instances of feedback as interdependent with potential implications for their entire self-concept. How do people maintain positivity and coherence across the self-concept while updating self-views from social feedback? Past research has examined how the brain engages in self-referential cognition (Wagner et al., 2012) and processes self-relevant social feedback (Somerville et al., 2010; Eisenberger et al., 2011). This work finds that the medial prefrontal (mPFC), anterior cingulate (ACC), and posterior cingulate (PCC) cortices process self-relevant feedback (Hughes and Beer, 2013; Yang et al., 2016; Will et al., 2017) and respond stronger to positive than negative feedback (Somerville et al., 2010; Korn et al., 2012; Yoon et al., 2018). However, prior work has neglected how the interrelationships among self-views shape how people process feedback and update self-views. Here, we investigate whether the structural interrelationships between self-views are critical to the computations the brain makes when modifying self-views from feedback.

Received Oct. 12, 2022; revised Feb. 16, 2023; accepted Apr. 4, 2023.
Author contributions: J.J.E., T.D., and B.H. designed research; J.J.E. performed research; J.J.E. contributed unpublished reagents/analytic tools; J.J.E. analyzed data; and J.J.E., T.D., and B.H. wrote the paper.
The authors declare no competing financial interests.
Correspondence should be addressed to Brent Hughes at bhughes@ucr.edu.
https://doi.org/10.1523/JNEUROSCI.1951-22.2023

To preserve self-concept positivity and coherence, the brain may use beliefs about semantic interdependencies between traits within the self-concept when processing feedback. We developed a network model (Fig. 1) that explains how people use semantic dependencies between traits to maintain positivity and coherence in the self-concept (Elder et al., 2023). These dependency relationships describe beliefs that people may commit themselves to when self-evaluating and updating from feedback. For example, if people believe that being witty depends on being outgoing, when a person updates their beliefs about how witty they are, they may feel committed to also update their beliefs about how outgoing they are.

Our network model can be used to generate predictions for how psychological and neurobiological processes related to self-updating are affected by the number of dependencies of a trait. Previously, we found that people tend to rate traits with a larger number of dependencies as more self-descriptive when the traits are positive and less self-descriptive when they are negative while also rating such traits more consistently with immediately connected (i.e., neighboring) traits in the network (Elder et al., 2023). Further, the ventral mPFC (vmPFC) tracks trait dependencies during self-evaluation, suggesting it may have a role in maintaining self-concept coherence. We later added reinforcement learning mechanisms to our network model and found that people tend to update self-views for traits with higher numbers of dependencies less when receiving feedback (Elder et al., 2022). These findings are consistent with the hypothesis that people maintain self-concept coherence by resisting change to self-beliefs about higher-dependency traits (Chen et al., 2016).

Here, we test how the brain uses dependency information when updating self-views from social feedback. Our primary region of interest is the vmPFC, given our hypothesis that this region is involved in maintaining self-concept coherence, and findings that this region is involved in general value-based learning (Bartra et al., 2013). Specifically, we expect the response of the vmPFC to feedback will vary based on the dependencies of a trait, consistent with our past behavioral results that people update their self-views less for traits with higher numbers of dependencies. Another region that may be implicated in processing dependency relations during feedback is the dorsal mPFC (dmPFC), which tracks trait dependencies in prior work (Elder et al., 2023) and is involved in trait inferences during mentalizing and self-reflection more broadly (Wagner et al., 2012; Lieberman et al., 2019; Tan et al., 2022).

Our model also allows us to test how the brain guides generalization by retrieving relationally similar traits during self-evaluation and manages uncertainty when a trait is semantically similar to past traits that have received more variable feedback. We find that vmPFC signals the novelty of traits by tracking our model-based retrieval measure (Garrido et al., 2015; Cockburn et al., 2022) and that angular gyrus activity tracks traits with more certain feedback based on relational structure (Davis and Yee, 2019). Incorporating reinforcement learning (RL) into our network model allows us to measure these unexplored neural computations that allow people to dynamically update self-evaluations based on relational similarity to prior social experiences.

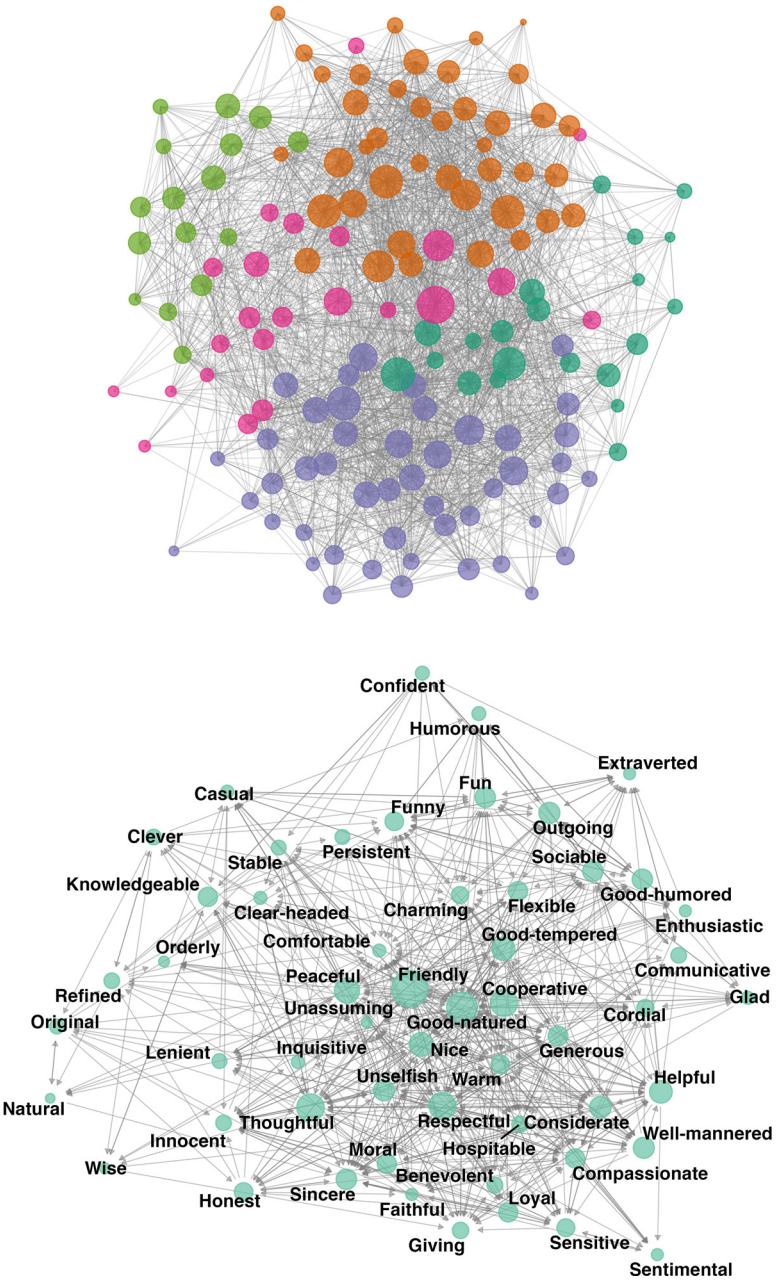

Figure 1. Network visualizations for full network and subset of network. Note: Nodes sized according to outdegree centrality. Top, Network visualization of 148 traits colored by community. Bottom, Outdegree connections for a high outdegree centrality trait Friendly.

Materials and Methods

Participants

Forty-six undergraduate students provided informed consent for participation and received course credit in compliance with approved University of California, Riverside (UCR) Institutional Review Board protocols. The sample was 63.0% female, age 18–31 years (M_{Age} = 19.89), 37% Asian, 19.6% Hispanic, 19.6% mixed, 15.2% White, 2.2% African American, and 3.0% other. Two subjects had incomplete self-report questionnaires because of protocol administration errors. The

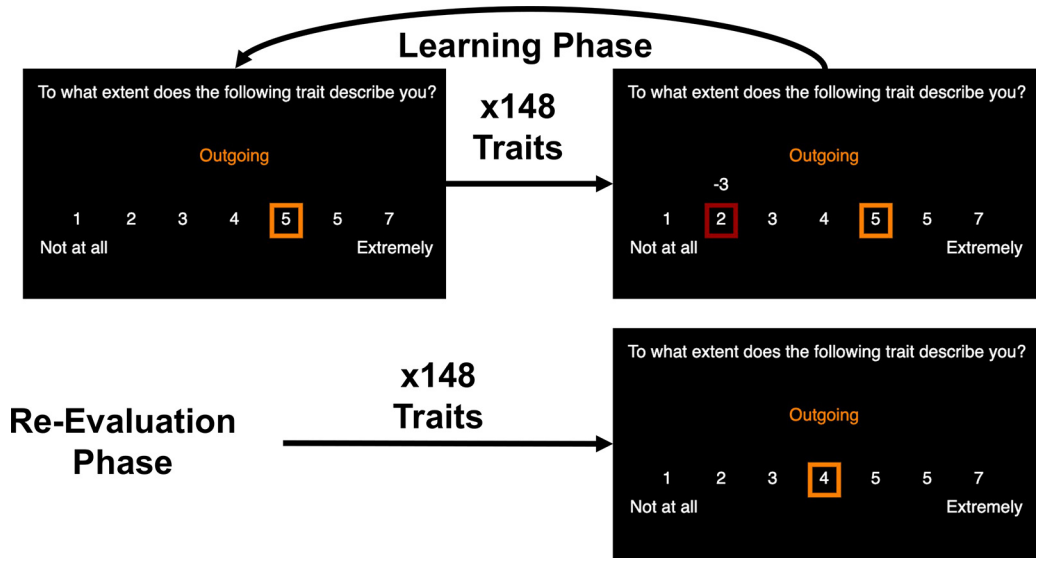

Figure 2. Illustration of the social-evaluative task. During the learning phase, participants evaluated themselves on each of 148 traits (e.g., Outgoing). An orange square appeared around their selection. After each self-evaluation, they were then shown how an admissions committee ostensibly evaluated them on that same trait. A red square appeared around the selection of the committee, and the discrepancy between the participant's self-evaluation and the evaluation of the committee was displayed in white. After completing self-evaluations and receiving feedback on all 148 traits, participants proceeded to the re-evaluation phase. They then self-evaluated themselves again on each trait. This figure is adapted from Elder et al. (2022) with permission.

target minimum sample size of 30 was determined by an a priori power analysis conducted using results from a previous behavioral study with the same design (Elder et al., 2022). We conducted Monte Carlo simulation-based power analyses (Green and MacLeod, 2016) on the mixed models and identified the minimum sample to achieve the previously observed main effect during learning ($N = 29$) and the interaction during re-evaluations ($N = 20$). We acquired a larger sample to improve power for fMRI analyses.

Experimental design

Procedure. When participants first entered the lab, they completed a consent form and received the cover story for the experiment. They were informed they would complete an interview that would be recorded and shared with three to five members of the UCR undergraduate admissions committee. During the interview, participants were asked a range of questions about their personal characteristics, goals, and interests. Interviews lasted ~10–20 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 (~7–10 d later).

Participants arrived at the UCR Center for Advanced Neuroimaging for the second visit to complete the social evaluative task while undergoing an fMRI scan. Participants were led to believe that in the time between the first and second visit, three to five members of the UCR admissions committee had reviewed the participant's video interview and evaluated the participant on all 148 trait words based on the interview.

Following the fMRI study, participants underwent a funnel offboarding interview and were asked a series of questions assessing the extent to which they were affected by feedback, were confused about the task, or how they felt about the experience. We then debriefed them and informed them the feedback was bogus, deception was involved, and asked them whether they believed the deception. Although some participants expressed skepticism after having been debriefed, all participants indicated having been affected by feedback before debriefing, so we included all participants in the sample.

Social evaluative feedback task

The experimental task was programmed in MATLAB Psychtoolbox (Kleiner et al., 2007) and was identical to a previous social evaluative feedback paradigm (Elder et al., 2022) but extended to fMRI. During the task, participants evaluated themselves on all 148 positive traits from the trait network on a 1 (not at all) to 7 (very much) scale 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 once the response was made. On each trial, participants were given 3 s to self-evaluate, after which there was a brief intertrial interval (ITI) in which the self-evaluation remained on the screen. All ITIs were drawn as random numbers from a truncated exponential distribution with a minimum of 2 s and a mean of 3 s. Feedback was then presented for 2 s with the prompt, "The reviewers see you as . . ." with the trait at the center of the screen and a red square framed around the assigned feedback. The orange score denoting the participant self-evaluation remained on the screen for the feedback phase (e.g., +3 if reviewer feedback was 7, and participant self-evaluation was 4). The task was presented using MATLAB Psychtoolbox and projected onto a screen that was viewed via a mirror mounted on the scanner (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 (see below, Trait network model for details on communities) for each participant. We used different probabilities of positive feedback as we wanted the different cues to have different probabilities of reward, akin to arms in a multiarmed bandit task commonly used in RL. Thus, we wanted to ensure some network communities, or groups of densely interconnected traits, had a higher probability of positive feedback (i.e., reward) than others to examine differences in learning. For instance, a trait belonging to the 70% community would have a 70% probability of receiving positive feedback (i.e., on average, feedback is higher than the self-evaluation) and a 30% probability of receiving negative feedback (i.e., on average, feedback is lower than the self-evaluation). Once feedback was determined according to the probability of a given community, the feedback number from one to seven was assigned according to criteria related to the participant's response and the determined feedback valence. We implemented this pseudorandom algorithm, whereby feedback was contingently positive or negative to ensure that similar groups of traits (i.e., communities) received similar feedback. Structuring the feedback along with communities ensured that participants could learn expected feedback for semantically related traits in the trait network so long as they represented the trait relationships described by our network. If the participant responded below the midpoint, positive feedback was two or more than participant responses, and negative feedback was

equivalent to or less than participant response. If the participant responded above the midpoint, positive feedback was equivalent to or more than participant response, and negative feedback was two or less than participant responses. If the participant's response was at the midpoint or the participant provided no response, positive feedback was above midpoint, and negative feedback was below the midpoint.

The first component of the task included four runs with 37 trials each of self-evaluations followed by feedback. Following 148 trials of self-evaluation followed by feedback, participants proceeded to a second component of the second task where they were asked to self-evaluate again on all of the traits they previously evaluated on, but they no longer received feedback. This component of this task included two runs, with 74 trials per run. This allowed us to measure the extent to which self-evaluations changed after feedback had been received. The scan took ~1 h total per participant to complete.

Imaging acquisition

Imaging data were acquired on a 3T MRI scanner (Prisma, Siemens Healthineers) at the University of Riverside Center for Advanced Neuroimaging using a 32-channel receive-only coil. Images from a T_1-weighted MP-RAGE sequence [echo time (TE)/repetition time (TR)/inversion time = 3.02 ms/2600 ms/800 ms, respectively; flip angle (FA) = 8°, voxel size = 0.8 × 0.8 × 0.8 mm^3] were used to position imaging volumes in functional scans in addition to use for registration from subject space to common space.

Functional data were collected with an T_2*-weighted gradient echo-planar imaging (EPI) sequence with the following scan parameters: TE/TR = 32 ms/1700 ms; slices = 72; FA = 75°, FOV = 220 mm 190 mm; matrix size =130′112; voxel size = 1.7 * 1.7 * 1.7 mm^3; GRAPPA = 2; multiband factor = 3; bandwidth =1540 Hz/pixel, phase encode = AP. A pair of spin echo EPI acquisitions with identical spatial parameters and bandwidth but opposite phase encoding directions (anterior to posterior (AP) and posterior to anterior (PA)) were collected to correct for susceptibility-related distortions.

Computational models

Trait network model. We present a model of self-concept updating that describes how people track relationships between traits to maintain positivity and coherence when updating self-beliefs based on social feedback (Elder et al., 2022, 2023). At the core of the model is a trait dependency network, constructed from an independent sample of participants. In applying the trait dependency model to how people maintain coherence and positivity of the self-concept, we assume that people are generally committed to maintaining coherence between traits that are generally believed to depend on another. That is, if people believe that being witty depends on being outgoing, then they will be committed to not contradicting this belief by claiming they are not outgoing when they believe they are witty. Explicitly modeling people's beliefs about traits separates our model from models of personality (such as the Big Five model), which are based on statistical associations among traits, independent of people's beliefs, and make no assumptions about how people maintain coherence among their different trait endorsements. For example, a personality model may predict that people who are witty are also more likely to be outgoing, but such models make no assumptions about whether people believe that these traits are dependent or are at all committed to maintaining coherence (noncontradiction) between their ratings of them when self-evaluating. Thus, our trait network is a model of what trait dependencies people believe exist and may commit themselves to, but not necessarily what statistical associations between traits actually exist (whether being witty and being outgoing are actually statistically associated or whether a person can actually be witty without being outgoing).

To construct the network, an independent sample of 178 Amazon Mechanical Turk participants was asked to nominate which of 147 positive trait words they believed depended on a target trait for semantic meaning (i.e., What trait does [TARGET TRAIT] depend on?). We arrived at our final set of 148 traits, by first starting with a list of 292 positive traits motivating by other literature (Anderson, 1968; Kirby and Gardner, 1972; Hampson et al., 1987). We had collected normative data on how interpersonal, desirable, prevalent, broad, and observable each trait was. We then filtered the traits down by determining which traits had the most reliable normative ratings across raters. This led us to a list of 150 traits. We then reduced this list further by removing the two positive traits in the list of 150 that had normative desirability less than 4.0 (the midpoint).

Each participant made dependency nominations for 10 traits with all other 147 traits available as dependency options. If more than 25% of participants agreed on a given dependency relationship, it was thresholded and included as a binary, directed relationship between from trait i to trait j ($i \rightarrow j$). We arrived at the 25% cutoff on the basis of simulations and reliability tests, whereby we thresholded the network at different cutoff points and determined the range at which the network metrics became relatively stable. We further verified the reliability of the network metrics at this threshold by bootstrapping the network and recomputing the network metrics (Elder et al., 2023, provides greater detail on network validation). From this procedure, we generated an adjacency matrix of 148 rows by 148 columns for trait words (Table 1 shows all traits.), computed based on the number of dependency relationships nominated by participants, such that a one in a cell reflected a trait in that column depending on the trait in that row (Elder et al., 2022, 2023).

Using this directed graph (Fig. 1), we generated a variety of measures. Outdegree centrality was defined as the number of traits that depend on a given trait (sum of the row of a given trait in the adjacency matrix; how many of columns j depend on row i). Indegree centrality was defined as the number of traits a given trait depends on (sum of a column of a given trait in the adjacency matrix; how many of rows i column j depends on). Pairwise similarity (i.e., dice similarity) between traits was calculated as two times the number of common neighbors between a pair of traits (i.e., traits both traits are immediately connected to), divided by the sum of their degrees (total number of connections), and ranges from zero to one. Similarity reflects the proportion of overlap between two traits in terms of shared trait neighbors. We identified groups of traits with dense interconnections, known as communities, by using a walktrap community detection algorithm (Pons and Latapy, 2005), and the number of communities extracted is based on the underlying structure of the data. The original procedure detected five communities. However, later analyses uncovered a coding error that excluded one trait word, which when corrected led the same algorithm to detect four total communities. The revised communities and original communities shared many overlapping traits and consisted of many neighboring traits regardless. The revised communities could not be used for feedback administration as the issue was encountered following the design of the study and conclusion of data collection. Table 2 contains a glossary of network and computational model terms.

To verify that the directedness of the network was critical for network structure, we calculated reciprocity, a measure of the likelihood that there is a reciprocal connection in the graph, given a known directed connection (varies from zero, where all connections in the network are unidirectional, to one, where all connections in the network are bidirectional). We found a reciprocity of 0.3516, suggesting that connections in the graph are more likely to be asymmetric than symmetric.

Learning model

Base model. As a basic test that people can learn about themselves from social feedback based on feedback that they have received, we implement an RL model (Rescorla and Wagner, 1972), where the model learns from the five communities (C) of traits based on trial-by-trial feedback. The model assumes an expectation associated with each of the five communities in the network, reflecting the learned expected feedback for each network community. We initialized the expectations at 4.0, which is consistent with conventions of using the midpoint between highest (1) and lowest feedback (7) to set initial values (Zhang et al., 2020). Each trait observed on each trial belongs to a specific community, which received different probabilities of positive feedback. $RLSE_{c\ i}$ is the expected social feedback for community c (that trait t belongs to) observed on trial i, and is thus the expected feedback for the community c that trait t on trial i belongs to. The model updates the expected of

Table 1. Traits contained within each network community

C1	C1	C2	C3	C4	C4	C5
Accurate	Practical	Well-organized	Charming	Benevolent	Lenient	Calm
Capable	Precise	Well-read	Communicative	Casual	Loyal	Clever
Clear-headed	Prompt	Wise	Enthusiastic	Charitable	Moral	Confident
Concise	Prudent	Clean	Extraverted	Comfortable	Nice	Courageous
Constructive	Punctual	Clean-cut	Fun	Compassionate	Peaceful	Fearless
Contemplative	Purposeful	Composed	Funny	Considerate	Positive	Frank
Dedicated	Rational	Conscientious	Good-humored	Cooperative	Respectful	Healthy
Deep	Realistic	Delicate	Humorous	Cordial	Romantic	Lucky
Deliberate	Scientific	Democratic	Lively	Ethical	Sensitive	Natural
Disciplined	Self-critical	Dependable	Outgoing	Fair	Sentimental	Normal
Eager	Self-sufficient	Dignified	Outspoken	Faithful	Sincere	Open-minded
Economical	Skillful	Elegant	Quick-witted	Flexible	Understanding	Optimistic
Experienced	Smart	Graceful	Sociable	Friendly	Unenvious	Original
Foresighted	Steady	Level-headed	Talkative	Generous	Unselfish	Passionate
Industrious	Straightforward	Mature	Verbal	Gentle	Warm	Prideful
Inquisitive	Studious	Neat	Witty	Giving		Respectable
Inventive	Thrifty	Polished		Glad		Thoughtful
Knowledgeable	Tough	Quiet		Good		Unafraid
Mathematical	Untiring	Refined		Good-natured		Unassuming
Modern		Reserved		Good-tempered		Unprejudiced
Orderly		Self-controlled		Helpful		Unpretentious
Perfectionistic		Sophisticated		Honest		Versatile
Persevering		Stable		Hospitable		Well-spoken
Persistent		Subtle		Humble		
Philosophical				Innocent		

Feedback was administered according to different probabilities of positive feedback to each community. 'C' denotes each of five communities.

Table 2. Glossary of terms

Term	Description
Outdegree centrality	The number of traits that people believe depend on a given trait.
Indegree centrality	The number of traits that people believe a given trait depends on.
Communities	Groups of densely connected traits with many shared neighboring traits.
Distance	The length of shortest paths from one trait to another. Reflects how far a trait is from another.
Similarity	The proportion of overlap between two traits, in terms of their shared number of connections out of total connections.
RLSE	The expected feedback based on the accumulation of prior feedback via error propagation.
SimSE	The expected feedback based on the similarity of a trait to feedback of prior trait.
SE	The expected feedback based on the mixture of RLSE and SimSE.
PE	The difference between the feedback received and the expectation of feedback (RLSE).
Familiarity	The overall summed similarity of a trait to all prior traits observed.
Uncertainty	An information theoretic measure of the variability of prior feedback received, based on similarity of prior traits to the current trait. It measures how predictable the feedback is for a trait.

social feedback, the Reinforcement Learning-based Social Expectation (RLSE), on each trial i based on the following rule:

$$RLSE_{ci+1} = RLSE_{ci} + \alpha * \delta_i, \quad (1)$$

where α is a free parameter representing a learning rate, and δ_i is the prediction error (PE) on trial i defined as the difference between feedback received and feedback expected (RLSE) as follows:

$$\delta_i = F_i - RLSE_{ci}. \quad (2)$$

The model predicts that people will self-evaluate in a manner that minimizes the difference between their self-evaluations and expected social feedback, RLSE. This model uses one free parameter.

Asymmetrical learning model. As an extension of the base model, whereby the model learns from communities with a single learning rate, we incorporate an additional free parameter to allow the model to learn differently from feedback better than expected (i.e., positive prediction errors) and feedback worse than expected (i.e., negative prediction errors). This model operates identically to the base model, except with different learning rates for different prediction errors. This model has been validated for the same design and framework in prior work (Elder et al., 2022). This model uses two free parameters.

Overall propagation model. The base model and asymmetrical learning model assume learning occurs homogeneously within a community. All traits in the community are updated the same when a trait in the community receives feedback, and all traits outside of the community are not updated at all. To create a more realistic propagation of error that is based on the full set of relations described by the network, we implement a model where feedback for a given trait can affect expectations of feedback for all traits in the network. In the current model, instead of associated expectations being linked with five communities, associated expectations are instead linked with all 148 traits.

Therefore, rather than there being five associated expectations learned by the model, there are 148 associated expectations learned by the model. Next, to incorporate a more holistic approach to feedback updating, the model assumes that when social feedback is observed for a given trait, this feedback causes updates to all traits connected to the focal trait. Updating of the focal trait t on the current trial, i, is the same above, but now all traits, j, also update based on their distance from focal trait t. How updating decays as a function of the distance of a trait from the focal trait t is given by the following:

$$RLSE_{ji+1} = RLSE_{ji} + \alpha * \frac{\delta_i}{1 + d_{tj}}, \quad (3)$$

where d_{tj} denotes trait j's distance from the focal trait t receiving feedback (i.e., length of shortest path from one node to all nodes) and $RSLE_{j\,i\,+1}$ denotes the expectation update for each trait j in the network on trial i. The impact of the prediction error update is greatest for the trait receiving feedback itself and equivalent to a standard RL update ($d_{tj} = 0$), is next greatest for traits immediately neighboring the trait receiving

feedback ($d_{tj} = 1$), and is weakest for traits at the opposite end of the network ($d_{tj} = 3$). This equation allows updating of weights based on feedback to be holistic and not based on the community of a trait. All traits connected to a given trait in the network are updated on every trial, with traits that are closer being updated more than traits that are farther. For example, on a trial in which outgoing is self-evaluated and receives feedback, outgoing will be updated the most, but traits that are immediately connected to outgoing, such as sociable, will be updated a little less, and traits two edges away, such as smart, will be updated still less, and so on. Therefore, in contrast to typical RL models that might update a single expectation at each trial, this model updates expectations of all traits simultaneously as a function of their distance from the affected trait. This model also incorporates the asymmetrical learning rates of the prior model and uses two free parameters.

Forward-propagation model. This model is identical to the previous overall propagation model, except that it restricts propagation to occur on downstream paths rather than any path (i.e., traits are updated as a function of depending on the trait receiving feedback, in terms of the length of shortest path of outdegree edges). The formula is the same, except that d_{tj} now represents the each downstream distance of each trait (i.e., outdegree edges) from the trait receiving feedback. This model uses two free parameters.

Back-propagation model. This model is identical to the previous overall propagation model, except that it restricts propagation to occur on upstream paths rather than any path (i.e., traits are updated that cause the trait receiving feedback, in terms of the length of shortest path of outdegree edges). The formula is the same, except that d_{tj} now represents the upstream distance of each trait (i.e., indegree edges) from the trait receiving feedback. This model uses two free parameters.

Mixture model. To test how people retrieve prior experiences for their ongoing self-evaluations, we test whether the relational similarity of a trait to previously evaluated traits will contribute to how people decide to evaluate.

On a trial-by-trial basis, people may retrieve prior trait feedback based on its relational similarity, which may influence their current decision. To reflect this process, we generated similarity-based social expectations (SimSE), which are estimated as the similarity-weighted mean of prior feedback. Specifically, the influence of prior feedback will be amplified or diminished in expectations based on the relational similarity of the current trait to prior traits as follows:

$$SimSE_{ti} = \frac{\sum_{j \in J} S_{tj} * F_j}{\sum_{j \in J} S_{tj}}, \tag{4}$$

where S_{tj} is the similarity of focal trait t to the previously presented trait j, and F_j is the feedback received for previous trait j of all prior traits J ($j \in J$).

The model assumed overall SE is a mixture of the two different types of expectation—expected social feedback based on accumulated social feedback via trial-and-error learning (RLSE) and expected social feedback based on the similarity to prior traits (SimSE). The mixture of these two forms of expectation is described by the following:

$$SE_{ti} = RLSE_{ti} * \varphi + SimSE_{ti} * (1 - \varphi), \tag{5}$$

where ϕ is a mixture parameter reflecting how much participants rely on trial-and-error feedback (closer to one means more reliance) or similarity-based retrieval (closer to zero means more reliance) to shape current expectations of social feedback. Therefore, expectations during learning from self-relevant social feedback consist of updates from error propagation (i.e., RLSE) as well as the retrieval of prior traits and relevant feedback (i.e., SimSE). This model has three free parameters.

Familiarity. Additional measures can be extracted from the similarity model to predict brain activation during the task. The overall amount of relational similarity (i.e., retrieval) on a given trial i reflects the familiarity of a trait (Gillund and Shiffrin, 1984; Nosofsky, 1988).

Familiarity is calculated during self-evaluation as the denominator of Equation 4, the summed similarity of the trait t in trial i to all prior traits j through J. Past neuroimaging studies have used this measure for perceptual categories (Davis et al., 2012b, 2014; Zeithamova et al., 2019), whereas we apply this measure to the self-concept in a relational category model. Some research has found that familiarity is inversely related to vmPFC activation (Garrido et al., 2015).

Uncertainty. Given a set of probabilities reflecting the likelihood of different feedback responses, we can estimate a measure of overall uncertainty using a standard entropy formulation (Shannon, 1948; Davis et al., 2012a, b). Uncertainty represents the likelihood of all feedback given prior traits observed, such that more uncertainty may be represented by equivalent likelihoods across all feedback categories as follows:

$$Entropy_{ti} = -\sum_{F}^{K} P_{Ft} * log_2 P_{Ft}, \tag{6}$$

where $P_F\,t$ denotes the probability of receiving feedback rating F for trait t, F is one of K feedback categories possible. Here, the probability of receiving a particular feedback rating is computed as the summed similarity of the current trait to all prior traits that received that feedback over the summed similarity of all prior traits regardless of feedback. Thus, the probability that trait t receives feedback F, $P_{F\,t}$, is defined as follows:

$$P_{Ft} = \frac{\sum_{f \in F} S_{tf}}{\sum_{K} \sum_{k \in K} S_{tk}}, \tag{7}$$

where S_{tf} represents the similarity of the trait t (on trial i) to trait f that received feedback rating F, and the index $f \in F$ indicates that the sum is over all traits f that received feedback rating F.

In this uncertainty formula (Hirsh et al., 2012; FeldmanHall and Shenhav, 2019), if all feedback was equally likely because of all prior feedback being for traits of equivalent similarity to the current trait, the current trait self-evaluation would have higher uncertainty. Conversely, if the current trait is most similar to traits that received feedback six and seven, but not similar to traits that received other types of feedback, the current trait would have lower uncertainty.

Model fitting. Parameters were fit to each subject's self-evaluations using the L-BFGS-B optimization algorithm from the optimx package, available in R software (Nash and Varadhan, 2011). Model free parameters were fit using a least-squares approach by squaring the difference between the SE of a trial and the self-evaluation of a trial and minimizing the sum of squared differences across trials as follows:

$$RSS = \sum_{i=1}^{I} (Evaluation_i - SE_i)^2. \tag{8}$$

Model comparison. To compare models, we used a formulation of Akaike information criterion (AIC) for residual sums of squares as follows:

$$AIC = 2\mathrm{p} + \mathrm{n} * ln\left(\frac{RSS}{n}\right), \tag{9}$$

The above formula describes n as the number of trials for a given participant and RSS as the residual sums of squares for the participant while penalizing p for the number of free parameters estimated in the model. We attempted Ridge regularized ordinary least squares (OLS) as in a prior work (Elder et al., 2022), but it did not contribute to improvements in model performance or recovery, so we retained the traditional OLS procedure. AIC values were summed across subjects to estimate model performance.

Table 3. Comparisons of each model depicting number of parameters, AIC, AIC difference, BIC, BIC difference, and the relative comparisons

No.	Model name	Param	Compare	AIC	ΔAIC	BIC	ΔBIC
1	Base	1		5711.318		5846.595	
2	Asymmetrical learning	2	2 vs 1	4938.931	−772.388	5209.485	−637.111
3	Overall propagation	2	3 vs 2	4389.841	−549.09	4660.396	−549.09
			3 vs 4		−19.466		−19.466
			3 vs 5		3.470		3.470
4	Forward propagation	2	4 vs 2	4409.306	−529.625	4679.860	−529.625
			4 vs 3		19.466		19.466
			4 vs 5		22.935		22.935
5	Backward propagation	2	5 vs. 2	4386.841	−552.560	4656.925	−552.560
			5 vs 3		−22.935		−22.935
			5 vs 4		−3.470		−3.470
6	Mixture with similarity	3	6 vs 5	4355.706	−30.665	4761.537	104.612

Param refers to number of free parameters. Compare refers to which model number is compared to which other model number. AIC refers to the Akaike information criteria for each model. BIC refers to the Bayesian information criteria for each model. Delta AIC refers to the difference in model AICs at each comparison, while Delta BIC refers to the difference in model BIC at each comparison.

We use AIC as the information criterion for model comparison decisions because Bayesian Information Criterion (BIC) imposes a largely penalty for additional parameters than AIC. Although we believe BIC risks underfitting and penalizing too strictly, we report BIC as well in model (Table 3).

Parameter recovery method. To determine whether the model parameters were identifiable at the individual participant level (Wilson and Collins, 2019; Lockwood and Klein-Flügge, 2020; Zhang et al., 2020), we tested whether they could be recovered from simulated data. We performed two different tests of parameter recovery, (1) randomly simulating parameters, generating behavior from these parameters and testing whether the parameters could be recovered during fitting to generated behavioral data, and (2) 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.

For each parameter from the best fitting model, we identified five equally spaced intervals between the 25th percentile and 75th of the distributions of the parameters. At each interval, we drew 100 values for a given parameter and added Gaussian noise equivalent to one-fourth SD of the original parameter distribution to increase the range of possible parameters simulated for positive learning rate, negative learning rate, and mixture. To simulate 500 participants (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 simulated parameter sets, we generated participant behavior and rounded trial-by-trial simulated estimates to the nearest whole number to emulate the Likert behavioral responses of the participants. Then, as with the original behavioral data, we fit 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 subject parameters (Vaidya and Badre, 2020). Participant fitted parameters were used to generate new behavioral data, and parameters were then fit to behavioral data generated by the original participant parameters. Correlations were estimated between the fitted parameters and the original participant parameters to estimate how recoverable the originally estimated parameters are.

Statistical analyses

Behavioral analysis. Multilevel models were implemented in R using lme4 (Bates et al., 2015), and Satterwaithe's approximation was used for determining p values in lmerTest (Kuznetsova et al., 2017). Semipartial R^2 (sr^2) estimates were computed for each fixed effects predictor using the standardized generalized variance approach using r2glmm (Edwards et al., 2008; Jaeger et al., 2017). Likelihood ratio tests were performed to determine models best supported by the data. Maximal random intercepts and slopes were tested and were removed as needed if unsupported by the data (i.e., low variance estimates) or if the model failed to converge (Barr et al., 2013). Moreover, models included crossed random factors (Baayen et al., 2008) with both traits and subjects modeled as random factors.

Trial-by-trial learning. We tested whether learned expectations generated from prior feedback could predict participants' trial-by-trial self-evaluations. To avoid overfitting, leave-one-participant-out cross-validation was used; for subject n from sample N, subject n's free parameters were omitted, and the summary statistics (mean for learning rates, median for mixture) of free parameters, from one to $N - n$, was determined. Parameters determined by the leave-one-out procedure were included in the computational model for subject n, such that any predictability produced from the computational model would not be a result of participant n's data and overfitting but rather from robustness of the model itself. The model contained trials nested within subjects, with subjects and traits set as random factors, initial self-evaluation as the response variable, and random slopes for SE for subjects and fixed slopes for traits.

Analysis of self-evaluation change

A residualized change approach (predicting re-evaluations while controlling for initial self-evaluations) was used to test for changes in self-views from initial self-evaluations to re-evaluations. To test whether the computational model can predict changes in self-evaluations, the last SE for the model during trial-by-trial learning was extracted for each trait within each participant. Then, for each participant and the 148 traits they observed and re-evaluated after learning, the final model SEs were used to reflect participants' social expectations for traits after learning had concluded. We tested a crossed random effects mixed model that included both subjects and traits set as random factors. Initial self-evaluations, outdegree centrality, and indegree centrality were entered as fixed slopes. PEs and SEs were entered as random slopes for both subjects and traits. Outdegree centrality was tested as an interaction with both PEs and SEs. The extent to which PE and SE predict re-evaluations while controlling for initial self-evaluations reflects change from initial evaluation (self-evaluation before receiving feedback), whereas the interaction terms with outdegree reflect the extent to which change from model PEs and SEs is conditional on outdegree centrality.

Neuroimaging preprocessing. Results included in this article come from preprocessing performed using fMRIPrep 1.4.0 (Esteban et al., 2019), which is based on Nipype 1.2.0 (Gorgolewski et al., 2011).

Anatomical data preprocessing. The T1-weighted (T1w) image was corrected for intensity nonuniformity with N4BiasFieldCorrection (Tustison et al., 2010), distributed with Advanced Normalization Tools (ANTs 2.2.0; Avants et al., 2008), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of CSF, white-matter, and gray-matter was performed on the brain-extracted T1w using fast (Zhang et al., 2001). Volume-based spatial normalization to one standard space (MNI152Nlin6Asym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: Functional MRI of the Brain (FMRIB) Software Library (FSL) MNI International Consortium for Brain Mapping 152 nonlinear sixth-generation Asymmetric Average Brain Stereotaxic Registration Model (Evans et al., 2012).

Functional data preprocessing. For each of the six BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A deformation field to correct for susceptibility distortions was estimated based on two EPI references with opposing phase-encoding directions, using 3dQwarp (Cox and Hyde, 1997). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate coregistration with the anatomic reference. The BOLD reference was then coregistered to the T1w reference using FLIRT

(FMRIB Linear Image Registration Tool; Jenkinson and Smith, 2001) with the boundary-based registration (Greve and Fischl, 2009) cost function. Coregistration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using MCFLIRT (FMRIB Linear Image Registration Tool with motion correction; Jenkinson et al., 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI (Analysis of Functional Neuro Images) 20160207 (Cox and Hyde, 1997). The BOLD time series were resampled onto their original native space by applying a single composite transform to correct for head motion and susceptibility distortions. These resampled BOLD time series are referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time series were resampled into standard space, generating a preprocessed BOLD run in MNI152Nlin6Asym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e., head-motion transform matrices, susceptibility distortion correction when available, and coregistrations to anatomic and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Nongridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Neuroimaging data analysis

fMRI statistical analyses were conducted using FEAT (FMRI Expert Analysis Tool) version 6.00 in FSL. Regressors and parameters were set at first-level model regressing voxelwise activity onto explanatory variables (EVs). Partial smoothing was applied using a three-dimensional 6 mm Filtered White Gaussian Noise kernel. The entire 4D dataset was grand-mean intensity normalized by a single multiplicative factor. High-pass temporal filtering was applied to remove low frequencies (128 s cut-off). For all models, nuisance regressors were included for motion (six head-motion parameters, three translation and three rotation, and their temporal derivatives), and volumes exceeding head motion of 0.9 mm framewise displacement were scrubbed (Siegel et al., 2014). Models included temporal filtering and temporal derivatives for each task variable. EVs were convolved with a double-gamma HRF. Continuous variables were scaled within subjects and centered within runs. Time series statistical analysis was conducted using FILM (FMRIB Improved Linear Model) with local autocorrelation correction (Woolrich et al., 2001). Statistical analyses were conducted using a standard level-level analysis in FEAT.

The second-level models, averaging contrast estimates within subjects, were tested using a fixed effects analysis. A third-level model, averaging contrast estimates between subjects, was tested using FLAME (FMRIB Local Analysis of Mixed Effects) stage 1, a mixed effects analysis that accounts for both within- and between-subject variances (Beckmann et al., 2003; Woolrich et al., 2004; Woolrich, 2008). Final statistical maps were corrected for multiple comparisons at $p < 0.05$ using permutation-based cluster mass thresholding, implemented in FSL Randomize. Whole-brain analyses used a primary cluster-forming threshold of $t = 3.28$ (critical value of t for df = 45 and $\alpha = 0.001$) and 6 mm variance smoothing. To generate EVs for fMRI analyses that involve RL parameters, the RL model was applied across all participants using the mean parameters for learning rates and median parameter for mixture (because of its skewness). It is conventional in RL applications in fMRI research to generate group-level parameters to stabilize noisy parameter estimates (Daw, 2011) and to provide an estimate of population parameters (Holmes and Friston, 1998).

Feedback models. Feedback models included a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor of PE at feedback onset (2 s), and a dummy indicator regressor indicating any missing responses at feedback onset (2 s), for a total of four EVs. We focused only on the contrast examining voxels where the average effect of PE is significantly different from zero.

However, given that feedback outcome and PE are highly correlated, they will often exhibit similar associations with neural response. We therefore employ an identify-and-justify approach (Zhang et al., 2020), first identifying regions associated with PE and justifying that these regions are indeed uniquely associated with PE, and not merely observed feedback. To implement this, we conducted an additional GLM consisting of PE components rather than PE. Specifically, we modeled a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor of observed feedback at feedback onset (2 s), a parametric regressor of RLSE at self-evaluation onset (3 s), a dummy indicator regressor indicating any missing responses at self-evaluation onset (3 s), and a dummy indicator regressor indicating any missing responses at feedback onset (2 s) for a total of six EVs. We focused only on the contrast examining the voxels where the average effect of feedback is greater than the average effect of expected feedback (i.e., Feedback > RLSE).

To justify the distinct contributions of PE and break up the cluster into smaller, more interpretable clusters, we perform a conjunction analysis by comparing the overlap in clusters between contrasts testing the effect of PE and PE components (i.e., PE ∩ Feedback > RLSE). To do so, we tested using a conjunction analysis, with the minimum statistic approach (Nichols et al., 2005), which identified voxels that were statistically significant in both the PE and Feedback > RLSE contrasts. We removed any clusters smaller than 50 voxels. This conjunction analysis aids in the interpretation of PE by justifying that activation is associated with PE and not only merely feedback.

As an additional analysis of PE, and to build on prior research on positively biased responses to self-relevant feedback (Korn et al., 2012; Hughes and Beer, 2013), we implemented a GLM with regressors for positive and negative PEs, along with the previously described constants and covariates. Specifically, this consisted of a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor of positive PE (feedback better than expected) at feedback onset (2 s), a parametric regressor of negative PE (feedback worse than expected) at feedback onset (2 s), and a dummy indicator regressor indicating any missing responses at feedback onset (2 s) for a total of five EVs. We focused on contrasts examining voxels where the average effect of positive PE is greater than the average effect of negative PE (positive PE > negative PE), and where the average effect of negative PE is greater than the average effect of positive PE (negative PE > positive PE).

Finally, we investigated whether neural processing of PEs depends on people's perceptions of trait dependencies (i.e., outdegree). Specifically, we modeled a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor of observed feedback onset (2 s), a parametric regressor of outdegree centrality at feedback onset (2 s), the interaction between feedback and outdegree centrality, and a dummy indicator regressor indicating any missing responses at feedback onset (2 s) for a total of six EVs. This interaction should provide insight into how the brain processes feedback and how initial processing of observed feedback manifests in differences in the computation of PEs. We conducted a whole-brain analysis but were primarily interested in the vmPFC. When precise localization for a small brain region is not the priority, concerns about false negatives can justify further data reduction techniques (e.g., ROI analysis) and more liberal thresholding (Carter et al., 2016). As such, to promote sensitivity for a potential vmPFC interaction effect, we constrained the thresholding space to an a priori vmPFC region identified as negatively associated with outdegree centrality in previous work (Elder et al., 2023), using a more liberal primary cluster-forming threshold of $t = 2.41$ (critical value of t for df = 45 and $\alpha = 0.01$).

Updating during learning. How the brain is involved during feedback processing may also reflect how people update and change their self-views. To test this, we compute a self-evaluation change score, that is, the difference between re-evaluation and initial self-evaluations. To explore asymmetries in processing positive change, negative change, and resistance to change, we split the change score into three components, positive

change (a continuous regressor depicting the amount the participant will change positively on the trait), negative change (a continuous regressor depicting the amount the participant will change negatively on the trait), and no change (a dummy indicator variable for traits the participant did not change self-evaluations on). The final model included a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor of positive change at feedback onset (2 s), a parametric regressor of negative change at feedback onset (2 s), a dummy indicator of no change at feedback onset (2 s), a parametric regressor of outdegree centrality at feedback onset (2 s), the interaction between no change and outdegree centrality, and a dummy indicator regressor indicating any missing responses at feedback onset (2 s) for a total of seven EVs. We were primarily interested in the voxels where the average effect was greater for positive change than negative change (Positive Change > Negative Change), where the average effect was greater for negative change than positive change (Negative Change > Positive Change), and where the effect for No Change trials depended on outdegree centrality (No Change * Outdegree). We conducted both a whole-brain analysis for the interaction, as well as constrained the analysis to the vmPFC mask previously described. This model provides insight into the regions that activate for positive relative to negative change, as well as well as the regions that associated with outdegree centrality when people do not change self-views.

Retrieval models. During self-evaluations, people retrieve past experiences and related feedback to determine how they see themselves. The familiarity of a trait given structural relatedness to prior self-evaluations informs one's current self-evaluations. Moreover, the feedback received for prior traits and how likely different types of feedback are given their structural relationships informs how certain or uncertain people are about their decisions. The following are modeled at self-evaluation onset.

Familiarity. We next examined the regions associated with the familiarity of a trait given the aggregate similarity to previous traits observed. Thus, we modeled a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor familiarity (i.e., summed similarity of current trait to prior observed traits) at self-evaluation onset (3 s), and a dummy indicator regressor indicating any missing responses at self-evaluation onset (3 s) for a total of four EVs. We specifically tested voxels where the average effect of familiarity was significantly different from zero.

Uncertainty. The likelihood of different types of feedback contributes to self-evaluative processes, as represented by uncertainty. Self-evaluations may be facilitated by greater certainty or stymied by greater uncertainty. To examine the decision processes underlying self-evaluations after having experienced feedback and retrieved prior experiences, we modeled a constant of stimulus presentation at self-evaluation onset (3 s), a constant of stimulus presentation at feedback onset (2 s), a parametric regressor decisional uncertainty (i.e., entropy as defined by the similarity of the current trait to traits that received different feedback) at self-evaluation onset (3 s), and a dummy indicator regressor indicating any missing responses at self-evaluation onset (3 s) for a total of four EVs. We specifically tested voxels where the average effect of uncertainty was significantly different from zero.

Data availability

Code and materials are openly available at GitHub via our Open Science Framework page at https://osf.io/2v7jc/?view_only=1ce6398515784671b2be7e25d39fc683. We generated a preregistration that can be found at https://aspredicted.org/rz5fb.pdf. Some of our fundamental RL-based predictions remain the same, but many of our predictions and analyses shifted from what was initially preregistered. We preregistered this project while still analyzing and writing our previous related projects applying network and RL techniques (Elder et al., 2022, 2023), and our thinking and expertise have evolved over the course of working with related data. We decided to further advance the computational modeling by incorporating the network structure into the RL model, which then allowed us to test additional behavioral and neural questions.

Results

Behavioral results

Computational model performance

We compared several models to identify a winning model. Our simple base model with a single learning rate (one free parameter; AIC = 5711.318) was outperformed by an asymmetrical learning model with two separate learning rates for positive and negative PEs (two free parameters; AIC = 4938.931), which both learned from broad network communities.

Next, we tested whether feedback only affects the local community of traits or spreads more holistically to other traits via interconnections described by the network (e.g., forward to children nodes or backward to parent nodes). To test this question, we compared the asymmetrical learning model (which learned only locally from communities) against models that propagated error based on the distance between traits (i.e., larger error-based updates for immediately connected nodes, less for more distant nodes). The propagation models incorporated the two learning rates for positive and negative prediction errors, such that the primary difference in the comparison was whether the model learned values for five communities or 148 traits simultaneously. The models that learn holistically and update all traits simultaneously vastly outperformed a model that learns only from communities. We further interrogated whether there are constraints to how prediction errors propagate. Indeed, the back-propagation model (two free parameters; AIC = 4386.371) that propagates prediction errors to the rest of the network based on neighbors that it depends on (indegree edges), outperformed a forward-propagation model (two free parameters; AIC = 4409.306) that propagates prediction errors to the rest of the network based on neighbors that depend on it (outdegree edges), and the propagation model that ignores directionality (two free parameters; AIC = 4389.841). The difference between the back-propagation and overall propagation AICs was small, so we interrogated individual AICs and found that 56.52% of participants had information criteria that were smaller for back-propagation than overall propagation. Furthermore, we applied the same modeling procedure to our previous dataset with an identical design (Elder et al., 2022) and found that the better fit for the back-propagation model replicated there.

Finally, we tested whether the back-propagation model could be further improved by incorporating similarity-based retrieval mechanisms at the cost of an additional free parameter for mixing trial-and-error-based expectations of social feedback and similarity-based expectations of social feedback. Indeed, the mixture back-propagation model was the best performing model (three free parameters; AIC = 4355.706). Table 3 shows model comparison statistics.

The propagation findings suggest that people do not receive feedback in isolation but rather use the feedback they experienced for a particular trait to inform their expectations for semantically related traits. Indeed, past work shows that people generalize errors during learning (Gershman and Niv, 2015; Jocham et al., 2016; Rudebeck et al., 2017; Baram et al., 2021). We extend this work by showing that people back-propagate errors to the parents of a trait (the traits it depends on) to resolve differences between feedback and expectations rather than propagating that error forward to the children of the trait. This is consistent with our overall theory that it is critical to maintain consistency in self-views between traits and those they depend on (Elder et al., 2023), as well as prior work on how people maintain coherence in beliefs more broadly (Thagard, 1989; Read and Marcus-Newhall, 1993;

Gershman, 2019). To be coherent and consistent, when one receives feedback about a proposition that differs from expectation, one ought to infer backward to correct the beliefs that led them to this error.

Computational model parameters and recovery

Consistent with our hypothesis that participants would learn more rapidly from feedback that was more positive than expected (positive PEs) than from feedback that was more negative than expected (negative PEs), a paired-samples permutation t test revealed that the learning rate for positive PEs is greater than negative PEs [positive learning rates, mean = 0.354, median = 0.136, SD = 0.403; negative learning rates, mean = 0.080 median = 0.039, SD = 0.102; observed difference = 0.274, $p < 0.001$). To further test whether learning is supported by the current model, we computed the absolute value for each PE, averaged all absolute PEs across all participants' trials, and estimated the Spearman's Rho correlation between trial number and average absolute PE. We used Spearman's Rho to estimate the correlation between trial number and absolute PE, as absolute PE averaged across subjects may decrease at a monotonic, but not necessarily linear, rate. There is a negative association between trial and average absolute PEs (Fig. 3), such that absolute PEs become smaller across time [$r(146) = -0.168$, $p = 0.042$].

As a test of the reliability of our model fits, we performed parameter recovery for our models by fitting the models to data generated with random parameter values and testing the associations between simulated and fitted parameter values. Our model parameters were recoverable both using randomly simulated parameters [$r(\alpha_p) = 0.81$; $r(\phi) = 0.72$; $r(\alpha_n) = 0.57$], and using the nonindependent and correlated parameters observed in our own participants' parameter fits [$r(\alpha_p) = 0.82$; $r(\alpha_n) = 0.76$; $r(\phi) = 0.74$]. Figure 3 for confusion matrices of correlations among observed and simulated parameters.

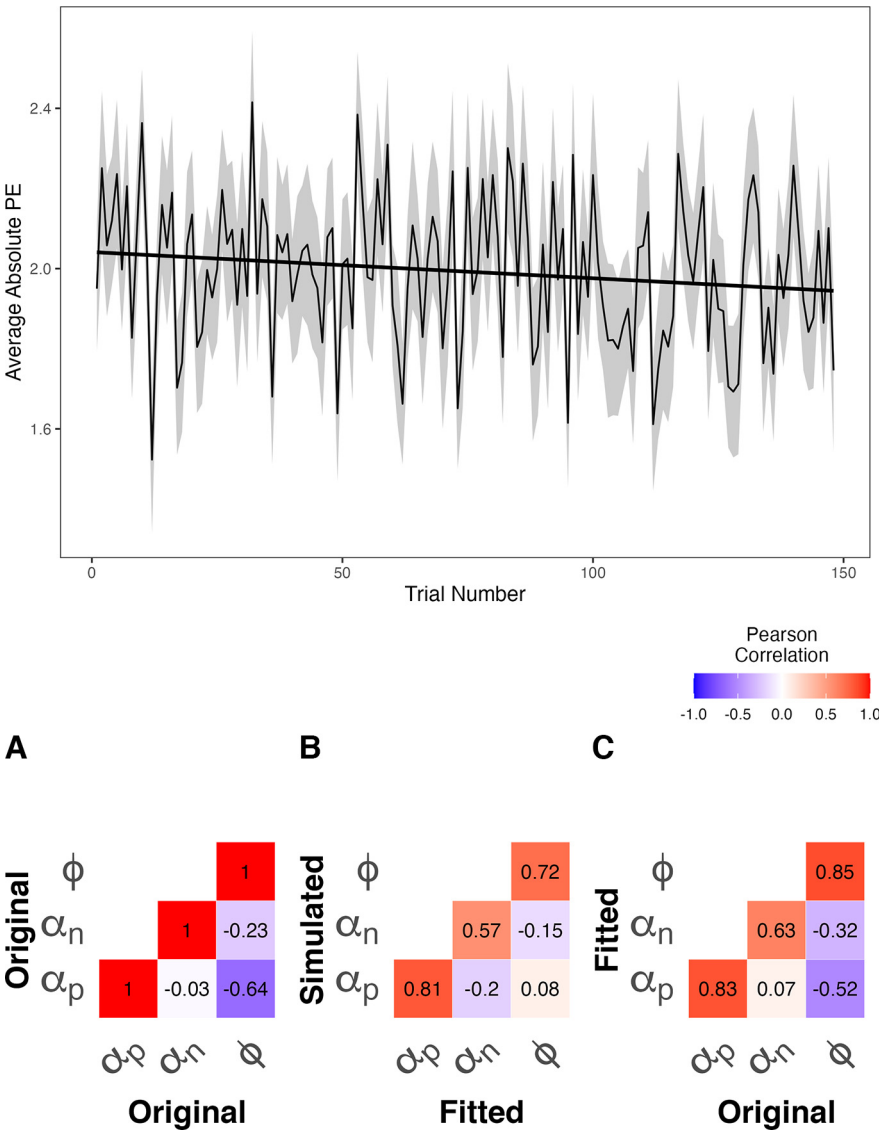

Figure 3. Top, Time series plot depicting the average absolute PE across trials. There is a negative Spearman's correlation between time and average absolute PEs such that absolute PEs become smaller across time [$r(146) = -0.168$, $p = 0.042$]. ***A–C***, Bottom, Confusion matrices depicting (***A***) the correlation among true parameters, (***B***) the correlation among Simulated and recovered (i.e., Fitted) parameters, and (***C***) the correlations among true and recovered (i.e., Fitted) parameters. Parameters appear generally recoverable.

Self-concept learning and change

The above analysis revealed that our final learning model was able to fit participants' behavior and support inferences about how their self-evaluations change from trial-by-trial feedback. As a test of the overall explanatory value and generalizability of our final model, we tested whether we could predict individual trial-by-trial responses using a leave-one-participant-out procedure, whereby each participant's computational model regressors were generated using the free parameter summary statistics (mean for learning rates, median for mixture) of all other participants. The model significantly predicted left-out participants' trial-by-trial self-evaluations ($\beta = 0.104$, SE = 0.021, $t_{(50)} = 5.015$, $p < 0.001$, $sr^2 = 0.013$), reflecting that the model effectively characterizes how people learn and self-evaluate in the present task.

Our next goal was to test whether the model predicts changes in self-evaluations between learning and re-evaluation and how outdegree centrality constrains participants' self-evaluation updates in response to feedback. To this end, we tested a residualized change model in which both expectations ($\beta = 0.227$, SE = 0.017, $t_{(51)} = 13.044$, $p < 0.001$, $sr^2 = 0.081$) and PE ($\beta = 0.108$, SE = 0.020, $t_{(45)} = 5.450$, $p < 0.001$, $sr^2 = 0.024$) predicted self-evaluations in the re-evaluation phase after all feedback had been received (controlling for initial self-evaluations during learning). Importantly, there was an interaction of PE with outdegree centrality ($\beta = -0.035$, SE = 0.009, $t_{(149)} = -3.830$, $p < 0.001$, $sr^2 = 0.003$), reflecting smaller changes in self-evaluations at higher levels of outdegree centrality. Consistent with our hypothesis and previous results (Elder et al., 2022) as well as related work (Chen et al., 2016), this suggests that outdegree centrality constrains the extent to which people update self-evaluations as a function of feedback (Fig. 4).

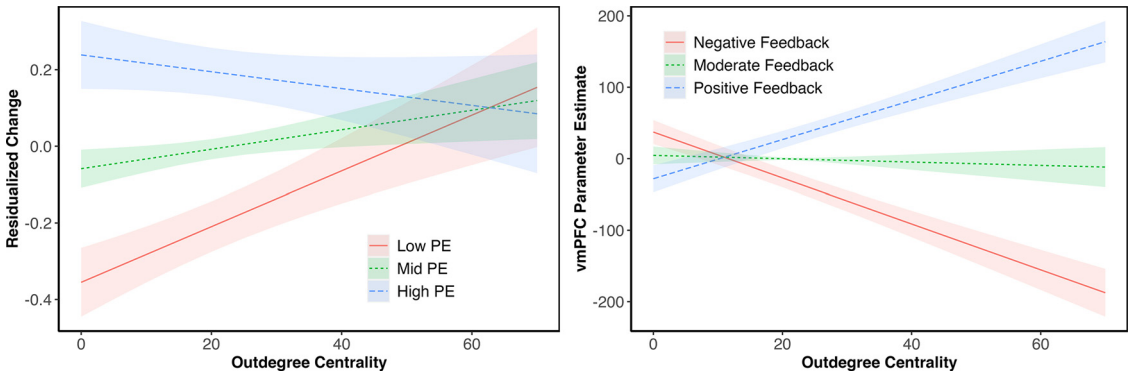

Figure 4. Predicted effects for outdegree-constrained effects of PE on changes in self-evaluations (left) and vmPFC subcluster activation (right). Note: Holding covariates constant, with 95% confidence intervals. Continuous variable of prediction error/feedback split into three levels for visualization. On the *x*-axis is outdegree centrality of trait. Left, Residualized change on *y*-axis is residuals from mixed model of re-evaluations predicted by initial self-evaluations, such that positive values are self-evaluations that increased relative to initial self-evaluations, and negative values are self-evaluations that decreased. Plots generated using ggeffects package in R (Lüdecke, 2018). Interaction illustrating that prediction errors contribute less to self-evaluations for traits higher in outdegree centrality. Right, Interaction illustrating feedback contributes to asymmetries in vmPFC response at higher outdegree. Predictions for the visualization were generated using the average parameter estimates across activated voxels in ventromedial prefrontal cortex.

Neuroimaging results

Feedback processes

Our primary neuroimaging questions surround how positivity and beliefs about dependency relations among traits constrain feedback processing in the brain. However, before testing our primary hypotheses, we first tested whether the vmPFC processes feedback and PEs in the present task, consistent with its role in more basic RL tasks.

Overall PE test and conjunction

We first estimated the effect of overall PE on brain activity. However, given that feedback outcome and PE are highly correlated, they will often exhibit similar associations with neural response. Therefore, we then employ an identify-and-justify approach (Zhang et al., 2020) by identifying regions associated with PE and justifying that these regions are indeed uniquely associated with PE and not merely with observed feedback.

First, we found that overall PE was associated with a large, undifferentiated, contiguous cluster centered in the occipital cortex (14, −78, 8; $k = 33\,203$, $t = 17.9$; $p < 0.001$) that stretched across cortical midline structures to vmpFC, as well as other clusters located in ventral striatum and extending to other subcortical regions (−14, 0, −14; $k = 1157$, $t = 7.32$, $p = 0.006$; Table 1). However, inferences about specific regions should not be made for contiguous, undifferentiated clusters (Woo et al., 2014a). Additionally, given that vmPFC responds to feedback more generally (Korn et al., 2012) and that the effect of PE can be difficult to disentangle from the effect of overall feedback (Zhang et al., 2020), we tested the extent to which the PE effect overlapped with the effect of the constituent parts of PE (Feedback – Expectation, represented in a contrast as Feedback > RLSE) using a conjunction analysis. This approach justifies the interpretation of activation as reflecting PE, rather than just feedback. Specifically, we first computed the contrast of Feedback > RLSE, and then examined the conjunction between the two contrasts (1) PE and (2) constituent components of PE (Feedback > RLSE) to test the regions of PE that are uniquely associated with the components of PE, and not merely overall feedback. This also served the purpose of breaking up the large undifferentiated cluster. This analysis revealed overlapping activation in regions such as vmPFC and posterior cingulate cortex (Table 4; Fig. 5). Consistent with prior work (Corlett et al., 2022), these results suggest that the vmPFC and other regions are involved in processing social feedback and further that it is sensitive to the amount that feedback deviates from expectations (i.e., prediction error).

Asymmetric processing of prediction errors

Behaviorally, participants updated their self-beliefs more from positive than negative PEs. To test whether this behavioral asymmetry is reflected in vmPFC (and other regions), we compared activation for positive PEs to negative PEs (i.e., positive PEs > negative PEs). Results revealed significant clusters of activation in vmPFC, bilateral superior temporal sulcus, precuneus, posterior cingulate, bilateral orbitofrontal cortex, and dorsal medial prefrontal cortex (Fig. 6, top; Table 5). The vmPFC activation is consistent with our hypothesis that the vmPFC may facilitate positively biased self-concept updating and is broadly consistent with other research on self and social feedback processing (Sharot et al., 2007; Somerville et al., 2010; Korn et al., 2012; Hughes and Beer, 2013; Hughes and Zaki, 2015). In particular, one candidate mechanism for involvement of vmPFC in processing positive self-relevant feedback may be because of its role in reward processing more generally (Rangel and Hare, 2010; Levy and Glimcher, 2012; Roy et al., 2012; Bartra et al., 2013; Tamir and Hughes, 2018). The involvement of temporoparietal and dorsal medial frontal regions in processing positive over negative PEs is consistent with regions found in mentalizing (Mitchell, 2009; Koster-Hale and Saxe, 2013; Kliemann and Adolphs, 2018) and in processing inconsistent information and updating impressions (Ma et al., 2012; Mende-Siedlecki and Todorov, 2016; Hughes et al., 2017; Charpentier and O'Doherty, 2018; Park et al., 2020a).

We also tested for regions that activated more for less-favorable PEs (i.e., negative PEs > positive PEs). We observed clusters in primary and secondary somatosensory cortex, postcentral gyrus, opercular cortex, and bilateral insular cortex (Fig. 6, bottom; Table 5). The regions that track less-favorable PEs are consistent with regions identified in social rejection and social pain-related response (Kross et al., 2011; Eisenberger, 2012; Woo et al., 2014b). In tandem, the asymmetries in brain processing of PEs observed here may support asymmetrical learning and an overall positive self-concept.

vmPFC response to feedback depends on outdegree centrality

Both our previous work (Elder et al., 2022) and behavior in the current task demonstrate that people tend to update higher

Table 4. Clusters associated with overall prediction errors, their components, and the conjunction between prediction error components and prediction error

Cluster no.	Positive association Region	Peak MNI coordinates x	y	z	Size	T	p
Prediction error (PE) = (1)							
1	Lingual gyrus	−14	−78	8	33203	17.9	0.001
2	Bilateral accumbens/putamen	−14	0	−14	1157	7.32	0.006
3	Right supramarginal gyrus/intraparietal sulcus	36	−38	38	688	5.15	0.014
4	Right anterior middle temporal gyrus	64	−6	−10	346	5.7	0.031
5	Dorsal lateral prefrontal cortex	22	48	42	285	5.32	0.041
6	Ventral lateral prefrontal cortex	42	54	0	264	5.66	0.046
Prediction error components (feedback, expectation) = (1, −1)							
1	Postcentral gyrus	40	−22	56	4081	9.57	0.001
2	Temporal occipital fusiform cortex	−32	−48	−22	2904	6.81	0.001
3	left inferior frontal gyrus	−44	8	28	2620	8.08	0.001
4	Left lateral superior occipital cortex	−26	−68	30	1876	6.35	0.002
5	Ventral medial prefrontal cortex	0	40	−16	1599	6.85	0.002
6	Supplementary motor area	−4	4	56	717	6.25	0.007
7	Right inferior frontal gyrus	50	14	32	510	5.7	0.011
8	Posterior cingulate cortex	0	−50	16	435	5.49	0.015
9	Occipital fusiform gyrus	22	−72	−16	245	4.81	0.034
10	Right angular gyrus	46	−52	20	226	4.85	0.039
Conjunction, PE ∩ Feedback > RL-SE							
1	Temporal occipital fusiform cortex	−46	−54	−28	1695	0.001	
2	Superior lateral occipital cortex	−28	−76	20	1691	0.002	
3	Medial prefrontal cortex	0	40	−24	1316	0.002	
4	Middle frontal gyrus	−40	28	16	703	0.001	
5	Middle frontal gyrus	−30	−6	42	674	0.001	
6	Supramarginal gyrus	62	−16	32	531	0.014	
7	Posterior cingulate cortex	−2	−44	4	323	0.015	
8	Inferior middle temporal gyrus	−58	−48	−16	132	0.001	
9	Supplementary motor area	0	4	50	95	0.007	
10	Occipital fusiform gyrus	26	−66	−30	63	0.034	
11	Lateral prefrontal cortex	−48	46	−10	59	0.001	
12	Posterior middle temporal gyrus	−62	−20	−24	53	0.001	

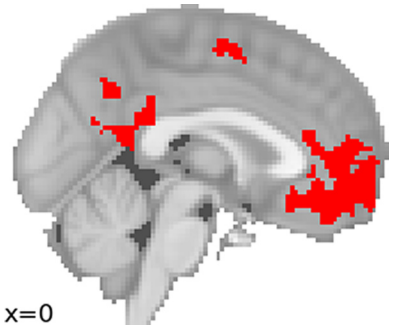

Figure 5. Statistical map of regions identified in conjunction analysis identifying intersection between prediction error contrast and feedback greater than RL-based social expectation contrast. Ventromedial prefrontal cortex and posterior cingulate cortex regions are positively associated with prediction errors, as well as the difference between feedback and expected feedback, and thus appear to be distinctly associated with prediction error rather than merely feedback.

outdegree traits less as a function of social feedback as a way of maintaining self-concept coherence. We also find that the vmPFC responds more strongly to positive than negative PEs as a way of maintaining self-concept positivity. Together, the asymmetrical vmPFC response to PEs and constrained self-updating as a function of outdegree could be reflected in a couple of ways. One possibility is that the vmPFC may respond less strongly to negative PEs and more strongly to positive PEs as outdegree increases, potentially diminishing the influence of negative PEs in the self-updating process. An alternative possibility is that before computing PE in the learning process vmPFC may de-emphasize negative feedback as outdegree increases. Doing so would reduce the impact of negative feedback with many implications by allowing fewer negative self-views across the self-concept without contradiction.

A whole-brain analysis testing whether outdegree modulates the processing of feedback and PE in the brain did not reveal any regions. Thus, we constrained our test to a vmPFC ROI found in our previous research that was associated with outdegree centrality (Elder et al., 2023). Consistent with the idea that the vmPFC may be involved in constraining updating from negative feedback for higher outdegree traits, we identified a subcluster within vmPFC (−6, 40, −14; $k = 28$, $t = 3.54$, $p = 0.041$) that showed a significant interaction between outdegree centrality and feedback. We observed that as outdegree centrality increases, vmPFC activity also increases for more positive feedback and decreases for more negative feedback (Fig. 4). We also observed a similar but marginally significant interaction of outdegree with PE in vmPFC (−6, 40, −14; $k = 18$, $t = 3.19$, $p = 0.058$), suggesting that the responsiveness to outdegree centrality during feedback processing may precede the computation of how the feedback differs from expectations (i.e., PE; Fig. 4). Together, this asymmetrical response to feedback as a function of outdegree centrality may reflect discarding negative feedback with many implications for the self-concept to minimize the negative self-views people feel committed toward. Conversely, people may be motivated to attend to positive feedback that bears many implications on other self-views.

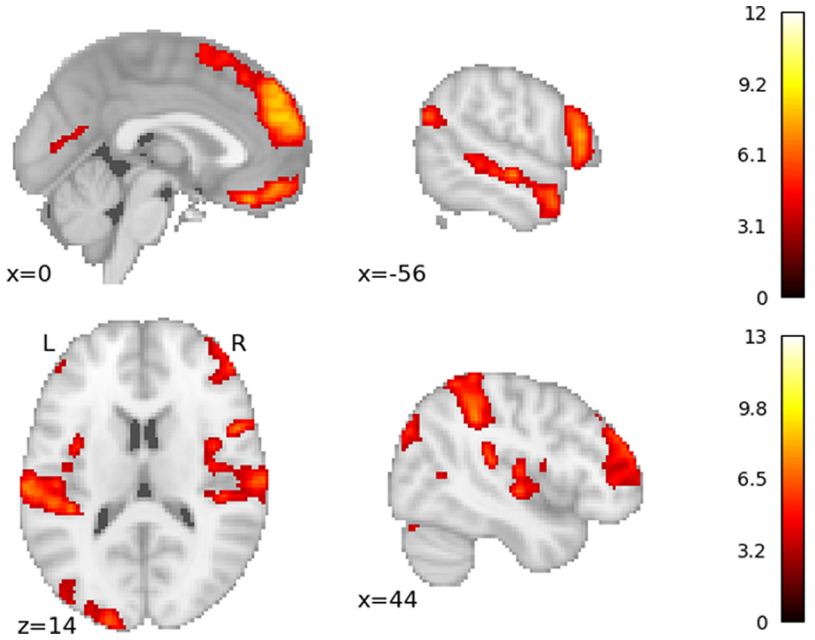

Figure 6. Statistical map of regions that are associated with positive PEs greater than negative PEs (top) and negative PEs greater than positive PEs (bottom) prediction errors. Mentalizing regions are associated with the difference between positive and negative prediction errors, whereas somatosensory regions are associated with the difference between negative and positive prediction errors. Note: Nonparametric thresholding at $\alpha = 0.001$ to correct for multiple comparisons. Right, Color bar depicts *t*-statistic magnitude.

Updating processes

Discarding negative feedback for higher outdegree traits may be one mechanism by which the brain governs self-concept updating. However, it is critical to also test whether activation at feedback predicts later changes in self-views. Does the response of the brain during feedback translate to actual changes in self-views as a function of outdegree centrality?

dmPFC predicts positive relative to negative updating

We first tested how brain activation during feedback predicted positive, negative, or no changes in self-views. We modeled participants' change from initial self-evaluations to re-evaluations as three separate regressors, positive change (change greater than zero), negative change (change less than zero), and no change (dummy indicators for trials in which self-evaluations remained unchanged across self-evaluations). We then tested whether brain activity during feedback more strongly predicts positive than negative change (Positive Change > Negative Change). We found regions in dmPFC and bilateral inferior frontal gyrus (IFG) that were more strongly associated with positive change than negative change (Table 6; Fig. 7). In contrast, we identified no regions that were more associated with negative than positive change. This may reflect the asymmetric trial-by-trial learning rates we observe, whereby updating is skewed toward generating more positive over more negative self-views. Together, findings suggest that activity in these brain regions may be involved in incorporating positive feedback that promotes positive change and discarding negative feedback that would otherwise promote negative change.

Neural mechanisms of outdegree-dependent resistance to change

Given that people resist updating self-views for higher outdegree traits, we next sought to test what neural mechanisms during feedback processing might predict this resistance to change in self-evaluations. We modeled the interaction between outdegree centrality and self-concept maintenance (i.e., trials in which people do vs do not update self-views). We previously found that vmPFC response to feedback was constrained by outdegree, so we next examined whether this constrained response to feedback would translate into resisting self-evaluation change as well. Indeed, in an ROI analysis, we identified a cluster in vmPFC (−2, 46, −16; $k = 41$, $k = 3.35$, $p = 0.028$) where activity decreases when people resist updating self-views on higher outdegree traits. This mirrors our previous finding that vmPFC activity decreases in response to higher outdegree negative feedback. It may be that vmPFC activity is attenuated for higher outdegree negative feedback that is being discarded in favor of maintaining existing self-views.

In whole-brain analyses, we found that activation in dmPFC/presupplementary motor area (pre-SMA) and left inferior frontal gyrus increases when people maintain versus change self-views on higher outdegree traits but little difference in activity for lower outdegree traits (Fig. 8; Table 7). These findings reflect that these regions activate more strongly when people maintain versus update their self-views on higher outdegree (but not lower outdegree) traits. These regions are known to be involved in cognitive control (Badre et al., 2009; Badre and Nee, 2018) and controlled semantic retrieval (Badre and Wagner, 2002; Lambon Ralph et al., 2017; Jackson, 2021), and may be gating the updating of the self-concept from feedback. Specifically, the more implications a trait has for the self-concept, the more control is exerted to resist changing self-views for this trait.

Retrieval processes

In addition to making explicit predictions for how the brain supports the processing of feedback and the updating of self-views from feedback, our model makes predictions for how the brain may engage in retrieval and decision processes during self-evaluations.

vmPFC tracks the similarity of a trait to past evaluated traits

Our model suggests that during self-evaluations people will retrieve information about past traits encountered in the task based on their relational similarity to the current trait to remain consistent in the self-evaluations over time. We define the sum of this similarity to past traits as a familiarity of a trait (Nosofsky, 1988) and use this familiarity measure to test which brain regions may be processing similarity to past traits during self-evaluations. Results revealed significant clusters in vmPFC (2, 30, −22; $k = 638$, $t = 5.88$, $p = 0.009$), posterior middle and superior temporal gyrus (−50, −30, −4; $k = 1385$, $t = 6.88$, $p = 0.003$), and right middle frontal gyrus (44, 6, 32; $k = 611$, $t = 6.17$, $p = 0.009$) that were negatively associated with familiarity at self-evaluation (Fig. 9; Table 8 for other clusters). This suggests that these brain regions are sensitive to the specific history of traits shown so far in the task, such that when a trait is less familiar, there is greater response across vmPFC, middle temporal gyrus, and middle frontal gyrus regions.

These results are consistent with past work showing that the posterior middle temporal gyrus is associated with semantic

Table 5. Clusters associated with positive prediction errors > negative prediction errors contrast

Asymmetrical prediction error contrast (positive PE, negative PE) = (1, −1)							
		Peak MNI coordinates					
Cluster no.	Positive Association Region	*x*	*y*	*z*	Size	*T*	*p*
1	Left lateral orbital frontal cortex	−48	22	−8	3902	12.1	0.001
2	Superior frontal gyrus	0	54	26	3406	8.4	0.001
3	Right lateral orbital frontal cortex	46	22	−10	1924	8.26	0.001
4	Intracalcarine cortex	−10	−82	6	830	12.3	0.005
5	Ventral medial prefrontal cortex	0	34	−22	440	7.67	0.014
6	Right posterior middle temporal gyrus	50	−22	−8	252	5.69	0.039
7	Lateral occipital cortex	−56	−66	26	214	6.1	0.049
Cluster no.	Negative Association Region	*x*	*y*	*z*	Size	*T*	*p*
1	Postcentral gyrus/right somatosensory	64	−30	44	20662	8.38	0.000
2	Right temporoocipital middle temporal gyrus	62	−58	0	397	6.86	0.019
3	Left middle frontal gyrus	−44	38	30	311	5.75	0.028
4	Occipital fusiform gyrus	18	−76	−16	286	5.6	0.032

Table 6. Clusters associated with positive change > negative change contrast

Asymmetrical change contrast (positive change, negative change) = (1, −1)							
		Peak MNI coordinates					
Cluster no.	Positive Association Region	*x*	*y*	*z*	Size	*T*	*p*
1	Middle frontal gyrus	−34	−2	66	3209	6.77	0.006
2	Presupplementary area dorsal medial prefrontal cortex	0	32	42	1560	6.35	0.018
3	Middle frontal gyrus	30	18	62	1245	5.03	0.026
4	Right lateral orbital frontal cortex	48	22	−10	706	6.23	0.048

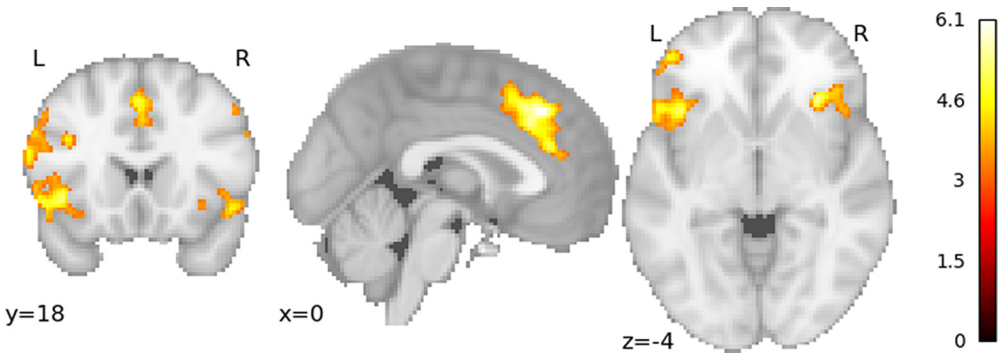

Figure 7. Statistical map of regions that are associated with positive change greater than negative change in self-evaluations, calculated as the difference between self-evaluations during the re-evaluation phase versus the learning phase. Presupplementary motor area and inferior frontal gyrus are positively associated with the difference between positive changes in self-views and negative changes in self-views. Note: Nonparametric thresholding at $\alpha = 0.001$ to correct for multiple comparisons. Right, Color bar depicts *t*-statistic magnitude.

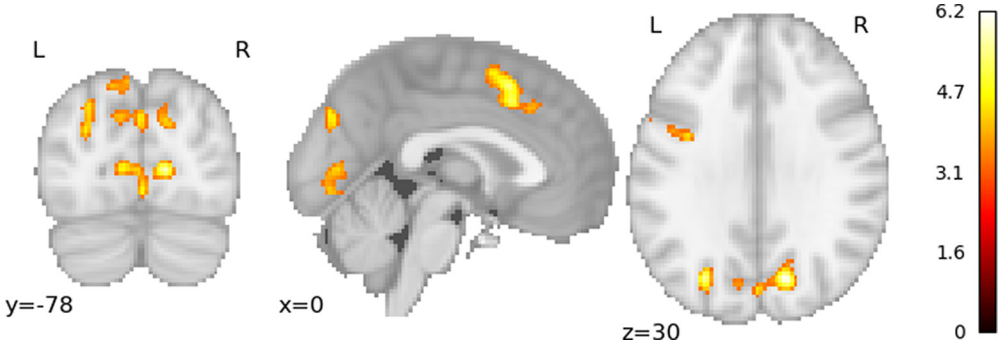

Figure 8. Statistical map of regions that exhibit greater activation when people maintain versus change their self-evaluations for higher outdegree centrality traits. Note: Nonparametric thresholding at $\alpha = 0.001$ to correct for multiple comparisons. Right, Color bar depicts *t*-statistic magnitude.

Table 7. Clusters associated with outdegree-dependent maintaining of self-evaluations

Outdegree-dependent resistance to change							
	Positive Association	Peak MNI coordinates					
Cluster no.	Region	*x*	*y*	*z*	Size	*T*	*p*
1	Intracalcarine cortex	14	−78	8	4464	6.55	0.002
2	Left middle frontal gyrus	−52	8	40	662	5.51	0.045
3	Presupplementary motor area dorsal medial prefrontal cortex	0	16	42	650	5.42	0.045

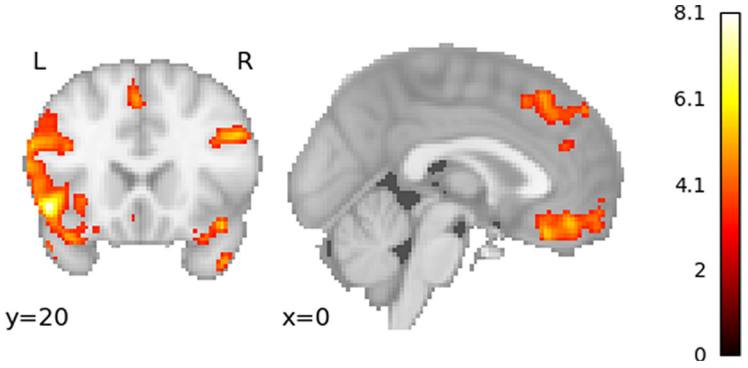

Figure 9. Statistical map of regions that exhibit less activity for familiarity (i.e., greater response to novelty). Ventromedial prefrontal cortex and middle frontal gyrus are associated with less aggregate similarity to previously self-evaluated traits. Note: Nonparametric thresholding at $\alpha = 0.001$ to correct for multiple comparisons. Right, Color bar depicts *t*-statistic magnitude.

retrieval (Davey et al., 2015, 2016), and middle frontal gyrus/lateral prefrontal cortex is associated with novelty, recollection, and familiarity-based retrieval (Friedman et al., 2001; Kishiyama et al., 2009). These results are also broadly consistent with the involvement of the vmPFC in processing novelty (Garrido et al., 2015) and detecting whether concepts are compatible and congruent in nonsocial domains (van Kesteren et al., 2012). To the extent that there are fewer or less overall similar previous observations to draw on, a current trait is less familiar. In such instances, the vmPFC, middle temporal gyrus, and related regions may provide a motivational novelty signal to help deploy processing resources to exemplars from sparse or less-clear areas of the network.

Angular gyrus responds to certainty of feedback

The familiarity analysis tests how brain activation reflects similarity to past traits, but this does not describe how the brain processes the certainty of expected feedback of a trait given the history of learned traits. For example, a trait could be very similar to several past traits, but if these traits all received different feedback, there would be more uncertainty about what the feedback would be for the current trait compared with if they all received similar feedback. To test how the brain processes this decisional uncertainty, we computed uncertainty as the likelihood of feedback given the similarity to prior traits that received feedback (Davis et al., 2012b). First, we computed probabilities of feedback categories based on how similar a current trait is to traits that received a given feedback category, such that a feedback category is more probable if prior traits that received that feedback are more similar to the current trait. Then, uncertainty is computed based on the probability of all feedback categories, and uncertainty is highest if all feedback values have equivalent probabilities because of similar traits receiving different feedback ratings. In a whole-brain analysis examining regions that correlate with uncertainty, we find that bilateral angular gyrus (right, 60, −50, 42; $k = 828$, $t = 6.38$, $p = 0.004$; right, 34, −42, 40; $k = 260$, $t = 5.35$, $p = 0.026$; left, −58, −48, 46; $k = 297$, $t = 4.94$, $p = 0.024$) and other regions (Fig. 10; Table 9) are negatively associated with uncertainty and that no regions were positively associated with uncertainty.

Our finding that angular gyrus activation is greater for traits that have more certain expected feedback is consistent with its role as a hub for integrating contextual information into specific events (Seghier, 2013) and for schematic inference (Gilboa and Marlatte, 2017). Although past studies on uncertainty using perceptual and economic decision tasks have focused on regions such as lateral PFC (Davis et al., 2017; FeldmanHall et al., 2019), studies using decision tasks requiring attention to well-learned semantic relations generally focus on the angular gyrus (Sachs et al., 2008; Seghier, 2013; Davis and Yee, 2019; Kuhnke et al., 2023). Indeed, we also previously found the angular gyrus tracks trait structure during self-evaluations (Elder et al., 2023).

Discussion

People generally strive to maintain the positivity and coherence of their interconnected self-concepts, and the interdependencies among people's self-beliefs bear important implications for how they update their self-views as a function of everyday social experiences. Here, we implement the first instance of a reinforcement learning model integrated directly into a network space to characterize the neural mechanisms by which people update interrelated self-views from social feedback and how they propagate this feedback across a system of self-beliefs. Doing so allows us to illustrate how feedback not only affects specific self-views in isolation but also propagates across trait dependencies to affect the broader system of self-views more holistically. Consistent with our hypothesis that people will process feedback differently for traits that are more central in the network and thus key for preserving coherence, we found that the vmPFC responds differently to feedback for traits with more dependencies (i.e., higher outdegree), and people tend to change their self-views less readily for these traits, suggesting that outdegree may modulate both how the brain responds to feedback and whether people decide to update their self-views from feedback. Together, our results provide insight into how the brain uses semantic relations among self-beliefs when learning from social feedback, and how such processes provide constraints that promote self-concept positivity and coherence.

Our results offer key insights into how beliefs about dependency relationships among traits shape learning about the self-concept and how this is mirrored by neural processing. By

Table 8. Clusters associated with familiarity

Familiarity contrast, familiarity = (1)							
		Peak MNI coordinates					
Cluster no.	Negative association Region	*x*	*y*	*z*	Size	*T*	*p*
1	Left lateral orbital frontal cortex	−48	20	−8	3161	7.9	0.000
2	Left posterior middle temporal gyrus	−50	−30	−4	1385	6.88	0.003
3	Right precentral gyrus	40	−22	66	1331	8.11	0.002
4	Ventral medial prefrontal cortex	2	30	−22	638	5.88	0.009
5	Right middle frontal gyrus	44	6	32	611	6.17	0.009
6	Superior frontal gyrus	−2	30	46	482	5.34	0.014
7	Lateral occipital cortex	−40	−72	30	422	5.24	0.019
8	Right lateral orbital frontal cortex/anterior temporal lobe	42	20	−18	325	5.46	0.025

developing a model of how self-beliefs relate to one another (Elder et al., 2023), we extend past work examining neural mechanisms of updating self-views from feedback, which has largely considered self-beliefs as isolated and unrelated instances (Eisenberger et al., 2011; Korn et al., 2012; Hughes and Beer, 2013; Will et al., 2017; Kawamichi et al., 2018). We show that people resist changing self-views on traits with more implications on other traits to maintain self-concept coherence (Elder et al., 2022) and identify neural computations involved in maintaining positivity and coherence. First, we replicate findings showing that vmPFC is preferentially tuned to positive over negative feedback (Somerville et al., 2010; Korn et al., 2012; Yang et al., 2016; Yoon et al., 2018) and further demonstrate that this asymmetric response to feedback is partially driven by the number of implications this feedback has for other dependent traits. Moreover, we found that this outdegree-dependent encoding of feedback constrains self-updating. In particular, vmPFC responses during feedback were attenuated when people maintained their self-evaluations on higher outdegree traits. This mirrors the finding that vmPFC exhibits less activity to negative feedback on higher outdegree traits, which are also associated with less overall self-updating. Finally, we found that dmPFC and IFG exhibited greater activity when resisting change relative to maintaining self-views for higher outdegree traits. In the current context, dmPFC and IFG may restrict the updating of self-evaluations based on the dependency relations of the traits, reflective of controlled semantic retrieval (Noonan et al., 2013; Jackson, 2021). Together, these findings highlight some of the neural computations by which people maintain a coherent and positive self-concept by selectively updating self-views from feedback as a function of their number of dependents.

To maintain self-concept coherence and avoid contradictions among self-beliefs, people must infer the dependencies among traits and incorporate that information into their self-evaluations. Indeed, we found that the vmPFC was negatively associated with our model-based familiarity measure, an aggregate measure of the similarity of a trait to previously evaluated traits in the task. The vmPFC may be involved in organizing and navigating the self-concept, just as it navigates other cognitive (Behrens et al., 2018) and spatial (Moser et al., 2008) maps that people use to explore structured environments (Schiller et al., 2015). In the current context, the vmPFC may signal the novelty of a current decision in a structured space (Hampton et al., 2006; Schuck et al., 2016; Kobayashi and Hsu, 2019; Park et al., 2020b; Knudsen and Wallis, 2022) by evaluating its similarity to past experiences based on their shared structural relationships. People may use structural relationships to infer expected feedback for a given decision, and by encoding its novelty, the vmPFC may aid in generalizing past experiences to the decision at hand.

Although our theory is influenced by past research on RL, decision-making, and categorization, it is important to consider how self-evaluations differ from standard learning and decision-making contexts. In most learning tasks, subjects are explicitly incentivized to optimize their behavior to the reward contingencies and determine which options provide more reward. In contrast, participants in our task were not instructed to align behavior with feedback, yet they nonetheless aligned their self-evaluations with past feedback to similar traits and predicted upcoming feedback about a trait when self-evaluating, taking into account the variability of past feedback they have received. This decision-making process was indexed by our model-based

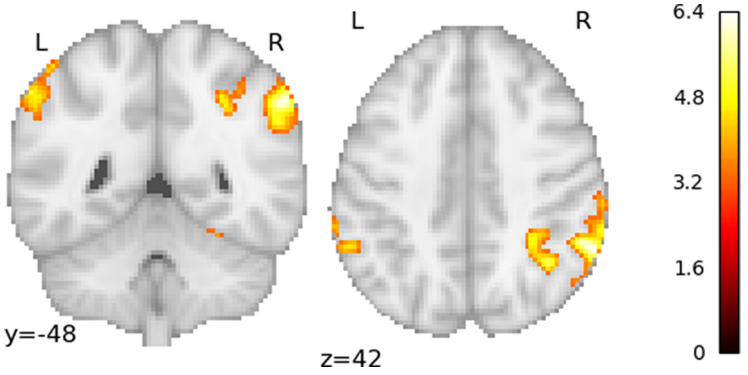

Figure 10. Nonparametric cluster-corrected regions negatively associated with uncertainty (i.e., greater response to certainty). Angular gyrus is associated with the certainty of expected feedback, given the similarity to prior traits that received feedback. Note: Nonparametric thresholding at $\alpha = 0.001$ to correct for multiple comparisons. Right, Color bar depicts *t*-statistic magnitude.

Table 9. Clusters associated with uncertainty

Entropy contrast, entropy = (1)							
		Peak MNI coordinates					
Cluster no.	Negative association Region	*x*	*y*	*z*	Size	*T*	*p*
1	Right angular gyrus	60	−50	42	828	6.38	0.004
2	Left angular gyrus	−58	−48	46	297	4.94	0.024
3	Occipital fusiform gyrus	26	−68	−16	261	5.99	0.027
4	Right superior parietal lobule	34	−42	40	260	5.35	0.026
5	Right lateral prefrontal cortex	40	44	4	173	5.79	0.048

uncertainty measure, which was negatively associated with angular gyrus, a region involved in making judgments by integrating across well-learned semantic relations and multisensory inputs (Seghier, 2013; Bonnici et al., 2018; Ramanan et al., 2018; Rugg and King, 2018; Ramanan and Bellana, 2019) and the automatic retrieval of semantic information (Davey et al., 2015). Even when people are not required to learn about contingencies between stimuli and feedback, they nonetheless evaluate the certainty of expected feedback by considering such contingencies, and aligning with others' views of them.

As one model of how people structure social knowledge, our trait network model suggests that people have directed, causal beliefs about the semantic dependencies between traits, which they use to maintain coherent (noncontradictory) beliefs when self-evaluating and incorporating social feedback. Our model is thus distinct from a recent model of learning from social knowledge structures (Frolichs et al., 2022), which is based on statistical associations between traits gathered from the Big Five model of personality (Digman, 1997). Models of statistical association would not immediately predict core findings of our belief-based model of differences in how people evaluate and update traits with higher numbers of dependencies. Likewise, because associations are not directional like our dependency network, these models could not predict differential effects for outdegree versus indegree centrality, or tendencies to update via backward-propagation versus forward-propagation. Although the Big Five model of personality used in Frolichs et al. (2022) is a model that can characterize how traits are statistically associated across individuals in the population, it is not intended to be a model of how people reason about causal commitments. To characterize how people maintain coherence among their self-views and stability in their self-concept, people must reason about causal commitments that support counterfactual reasoning (e.g., I would not be witty if I were not outgoing; Zhou et al., 2023). Future studies should aim to bridge these models and identify when directed or undirected mental models provide a better account of social learning and inference.

We made a number of pragmatic choices when designing this task that open new avenues for future research. First, we constructed two discrete networks, one containing positive traits and another containing negative traits, and focused here on the positive trait network, given our interest to examine how feedback propagates to all traits within a network. Including both complete networks would not have been possible because of time constraints and participant fatigue. Future research can test whether the current mechanisms extend to learning about negative traits. We also described feedback to participants as coming from admissions committee members. This raises interesting questions about whether learning effects vary based on status or other features of the sources of feedback. Future studies might also compare how people learn about themselves and others to examine the differences between self-relevant relative to other-relevant learning (Korn et al., 2012).

The self-concept is a dynamic mental structure, with different self-aspects activated across varying contexts (Markus and Kunda, 1986; Markus and Wurf, 1987; McConnell, 2011). Here, we provide the first evidence of the neural mechanisms supporting this dynamic process and a formal model for how this working self-concept is activated by different experiences. By developing a deeper structure of beliefs about dependency relations within the self-concept, we can understand how people dynamically update their self-beliefs. People learn from feedback, propagate that feedback across a system of beliefs, and constrain their learning based on the number of trait implications, which is retrieved for subsequent self-reflections via relational-matching processes. Importantly, asymmetries in self-learning are mirrored at the neural level by parallel effects of network structure on brain activation. Our work highlights the importance of incorporating relational structure into how we understand people's self-beliefs and changes to the working self-concept as a function of experience and social feedback.

References

Anderson N (1968) Likableness ratings of 555 personality-trait words. J Pers Soc Psychol 9:272–279.

Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12:26–41.

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412.

Badre D, Wagner AD (2002) Semantic retrieval, mnemonic control, and prefrontal cortex. Behav Cogn Neurosci Rev 1:206–218.

Badre D, Nee DE (2018) Frontal cortex and the hierarchical control of behavior. Trends Cogn Sci 22:170–188.

Badre D, Hoffman J, Cooney JW, D'Esposito M (2009) Hierarchical cognitive control deficits following damage to the human frontal lobe. Nat Neurosci 12:515–522.

Baram AB, Muller TH, Nili H, Garvert MM, Behrens TEJ (2021) Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. Neuron 109:713–723.e7.

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. J Mem Lang 68:255–278.

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412–427.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Soft 67:1–48.

Beckmann CF, Jenkinson M, Smith SM (2003) General multilevel linear modeling for group analysis in FMRI. Neuroimage 20:1052–1063.

Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z (2018) What is a cognitive map? Organizing knowledge for flexible behavior. Neuron 100:490–509.

Bonnici HM, Cheke LG, Green DAE, FitzGerald THMB, Simons JS (2018) Specifying a causal role for angular gyrus in autobiographical memory. J Neurosci 38:10438–10443.

Carter CS, Lesh TA, Barch DM (2016) Thresholds, power, and sample sizes in clinical neuroimaging. Biol Psychiatry Cogn Neurosci Neuroimaging 1:99–100.

Charpentier CJ, O'Doherty JP (2018) The application of computational models to social neuroscience: promises and pitfalls. Soc Neurosci 13:637–647.

Chen SY, Urminsky O, Bartels DM (2016) Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. Psychol Sci 27:1398–1406.

Cockburn J, Man V, Cunningham WA, O'Doherty JP (2022) Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. Neuron 110:2691–2702.e8.

Corlett PR, Mollick JA, Kober H (2022) Meta-analysis of human prediction error for incentives, perception, cognition, and action. Neuropsychopharmacology 47:1339–1349.

Cox RW, Hyde JS (1997) Software tools for analysis and visualization of fMRI data. NMR Biomed 10:171–178.

Davey J, Cornelissen PL, Thompson HE, Sonkusare S, Hallam G, Smallwood J, Jefferies E (2015) Automatic and controlled semantic retrieval: tms reveals distinct contributions of posterior middle temporal gyrus and angular gyrus. J Neurosci 35:15230–15239.

Davey J, Thompson HE, Hallam G, Karapanagiotidis T, Murphy C, De Caso I, Krieger-Redwood K, Bernhardt BC, Smallwood J, Jefferies E (2016) Exploring the role of the posterior middle temporal gyrus in semantic cognition: integration of anterior temporal lobe with executive processes. Neuroimage 137:165–177.

Davis CP, Yee E (2019) Features, labels, space, and time: factors supporting taxonomic relationships in the anterior temporal lobe and thematic relationships in the angular gyrus. Lang Cogn Neurosci 34:1347–1357.
Davis T, Love BC, Preston AR (2012a) Learning the exception to the rule: model-based fMRI reveals specialized representations for surprising category members. Cereb Cortex 22:260–273.
Davis T, Love BC, Preston AR (2012b) Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI. J Exp Psychol Learn Mem Cogn 38:821–839.
Davis T, Xue G, Love BC, Preston AR, Poldrack RA (2014) Global neural pattern similarity as a common basis for categorization and recognition memory. J Neurosci 34:7472–7484.
Davis T, Goldwater M, Giron J (2017) From concrete examples to abstract relations: the rostrolateral prefrontal cortex integrates novel examples into relational categories. Cereb Cortex 27:2652–2670.
Daw ND (2011) Trial-by-trial data analysis using computational models. Decision making, affect, and learning: Attention and performance XXIII, 23:3–38.
Digman JM (1997) Higher-order factors of the Big Five. J Pers Soc Psychol 73:1246–1256.
Edwards LJ, Muller KE, Wolfinger RD, Qaqish BF, Schabenberger O (2008) An R2 statistic for fixed effects in the linear mixed model. Stat Med 27:6137–6157.
Eisenberger NI (2012) The neural bases of social pain: evidence for shared representations with physical pain. Psychosom Med 74:126–135.
Eisenberger NI, Inagaki TK, Muscatell KA, Byrne Haltom KE, Leary MR (2011) The neural sociometer: brain mechanisms underlying state self-esteem. J Cogn Neurosci 23:3448–3455.
Elder J, Davis T, Hughes BL (2022) Learning about the self: motives for coherence and positivity constrain learning from self-relevant social feedback. Psychol Sci 33:629–647.
Elder J, Cheung B, Davis T, Hughes B (2023) Mapping the self: a network approach for understanding psychological and neural representations of self-concept structure. J Pers Soc Psychol 124:237–263.
Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ (2019) fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods 16:111–116.
Evans CEL, Christian MS, Cleghorn CL, Greenwood DC, Cade JE (2012) Systematic review and meta-analysis of school-based interventions to improve daily fruit and vegetable intake in children aged 5 to 12 y. Am J Clin Nutr 96:889–901.
FeldmanHall O, Shenhav A (2019) Resolving uncertainty in a social world. Nat Hum Behav 3:426–435.
FeldmanHall O, Glimcher P, Baker AL, Phelps EA (2019) The functional roles of the amygdala and prefrontal cortex in processing uncertainty. J Cogn Neurosci 31:1742–1754.
Friedman D, Cycowicz YM, Gaeta H (2001) The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci Biobehav Rev 25:355–373.
Frolichs KMM, Rosenblau G, Korn CW (2022) Incorporating social knowledge structures into computational models. Nat Commun 13:6205.
Garrido MI, Barnes GR, Kumaran D, Maguire EA, Dolan RJ (2015) Ventromedial prefrontal cortex drives hippocampal theta oscillations induced by mismatch computations. Neuroimage 120:362–370.
Gershman SJ (2019) How to never be wrong. Psychon Bull Rev 26:13–28.
Gershman SJ, Niv Y (2015) Novelty and inductive generalization in human reinforcement learning. Top Cogn Sci 7:391–415.
Gilboa A, Marlatte H (2017) Neurobiology of schemas and schema-mediated memory. Trends Cogn Sci 21:618–631.
Gillund G, Shiffrin RM (1984) A retrieval model for both recognition and recall. Psychol Rev 91:1–67.
Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS (2011) Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform 5:13.
Green P, MacLeod CJ (2016) SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol Evol 7:493–498.
Greve DN, Fischl B (2009) Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63–72.
Hampson SE, Goldberg LR, John OP (1987) Category-breadth and social-desirability values for 573 personality terms. Eur J Pers 1:241–258.
Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26:8360–8367.
Hirsh JB, Mar RA, Peterson JB (2012) Psychological entropy: a framework for understanding uncertainty-related anxiety. Psychol Rev 119:304–320.
Holmes A, Friston KJ (1998) Generalisability, random effects and population inference. Neuroimage 7:S754.
Hughes BL, Beer JS (2013) Protecting the self: the effect of social-evaluative threat on neural representations of self. J Cogn Neurosci 25:613–622.
Hughes BL, Zaki J (2015) The neuroscience of motivated cognition. Trends Cogn Sci 19:62–64.
Hughes BL, Ambady N, Zaki J (2017) Trusting outgroup, but not ingroup members, requires control: neural and behavioral evidence. Soc Cogn Affect Neurosci 12:372–381.
Jackson RL (2021) The neural correlates of semantic control revisited. Neuroimage 224:117444.
Jaeger BC, Edwards LJ, Das K, Sen PK (2017) An R2 statistic for fixed effects in the generalized linear mixed model. J Appl Stat 44:1086–1105.
Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156.
Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.
Jocham G, Brodersen KH, Constantinescu AO, Kahn MC, Ianni AM, Walton ME, Rushworth MFS, Behrens TEJ (2016) Reward-guided learning with and without causal attribution. Neuron 90:177–190.
Kawamichi H, Sugawara SK, Hamano YH, Kitada R, Nakagawa E, Kochiyama T, Sadato N (2018) Neural correlates underlying change in state self-esteem. Sci Rep 8:1798.
Kirby DM, Gardner RC (1972) Ethnic stereotypes: norms on 208 words typically used in their assessment. Can J Psychol 26:140–154.
Kishiyama MM, Yonelinas AP, Knight RT (2009) Novelty enhancements in memory are dependent on lateral prefrontal cortex. J Neurosci 29:8114–8118.
Kleiner M, Brainard DH, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in Psychtoolbox-3? Perception 36:1–16.
Kliemann D, Adolphs R (2018) The social neuroscience of mentalizing: challenges and recommendations. Curr Opin Psychol 24:1–6.
Knudsen EB, Wallis JD (2022) Taking stock of value in the orbitofrontal cortex. Nat Rev Neurosci 23:428–438.
Kobayashi K, Hsu M (2019) Common neural code for reward and information value. Proc Natl Acad Sci U S A 116:13061–13066.
Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR (2012) Positively biased processing of self-relevant social feedback. J Neurosci 32:16832–16844.
Koster-Hale J, Saxe R (2013) Theory of mind: a neural prediction problem. Neuron 79:836–848.
Kross E, Berman MG, Mischel W, Smith EE, Wager TD (2011) Social rejection shares somatosensory representations with physical pain. Proc Natl Acad Sci U S A 108:6270–6275.
Kuhnke P, Chapman CA, Cheung VKM, Turker S, Graessner A, Martin S, Williams KA, Hartwigsen G (2023) The role of the angular gyrus in semantic cognition: a synthesis of five functional neuroimaging studies. Brain Struct Funct 228:273–291.
Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: tests in linear mixed effects models. J Stat Soft 82:1–26.
Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT (2017) The neural and computational bases of semantic cognition. Nat Rev Neurosci 18:42–55.
Lanczos C (1964) A precision approximation of the gamma function. J SIAM Numer Anal Ser B 1:86–96.
Levy DJ, Glimcher PW (2012) The root of all value: a neural common currency for choice. Curr Opin Neurobiol 22:1027–1038.
Lieberman MD, Straccia MA, Meyer ML, Du M, Tan KM (2019) Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): causal, multivariate, and reverse inference evidence. Neurosci Biobehav Rev 99:311–328.
Lockwood PL, Klein-Flügge MC (2020) Computational modelling of social cognition and behaviour—a reinforcement learning primer. Soc Cogn Affect Neurosci 16:761–771.

Lüdecke D (2018) Ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. Journal of Open Source Software 3:772.
Ma N, Vandekerckhove M, Baetens K, Van Overwalle F, Seurinck R, Fias W (2012) Inconsistencies in spontaneous and intentional trait inferences. Soc Cogn Affect Neurosci 7:937–950.
Markus H, Kunda Z (1986) Stability and malleability of the self-concept. J Pers Soc Psychol 51:858–866.
Markus H, Wurf E (1987) The dynamic self-concept: a social psychological perspective. Annu Rev Psychol 38:299–337.
McConnell AR (2011) The multiple self-aspects framework: self-concept representation and its implications. Pers Soc Psychol Rev 15:3–27.
Mende-Siedlecki P, Todorov A (2016) Neural dissociations between meaningful and mere inconsistency in impression updating. Soc Cogn Affect Neurosci 11:1489–1500.
Mitchell JP (2009) Inferences about mental states. Philos Trans R Soc Lond B Biol Sci 364:1309–1316.
Moser EI, Kropff E, Moser M-B (2008) Place cells, grid cells, and the brain's spatial representation system. Annu Rev Neurosci 31:69–89.
Nash JC, Varadhan R (2011) Unifying optimization algorithms to aid software system users: optimx for R. J Stat Soft 43:1–14.
Nichols T, Brett M, Andersson J, Wager T, Poline J-B (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660.
Noonan KA, Jefferies E, Visser M, Lambon Ralph MA (2013) Going beyond inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. J Cogn Neurosci 25:1824–1850.
Nosofsky RM (1988) Similarity, frequency, and category representations. J Exp Psychol Learn Mem Cogn 14:54–65.
Palminteri S, Lefebvre G, Kilford EJ, Blakemore S-J (2017) Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing. PLOS Comput Biol 13:e1005684.
Park B, Fareri D, Delgado M, Young L (2020a) The role of right temporoparietal junction in processing social prediction error across relationship contexts. Soc Cogn Affect Neurosci 16:772–781.
Park SA, Miller DS, Nili H, Ranganath C, Boorman ED ED (2020b) Map making: constructing, combining, and inferring on abstract cognitive maps. Neuron 107:1226–1238.e8.
Pons P, Latapy M (2005) Computing communities in large networks using random walks. arXiv. Advance online publication. Retrieved April 10, 2023. .
Ramanan S, Bellana B (2019) A domain-general role for the angular gyrus in retrieving internal representations of the external world. J Neurosci 39:2978–2980.
Ramanan S, Piguet O, Irish M (2018) Rethinking the role of the angular gyrus in remembering the past and imagining the future: the contextual integration model. Neuroscientist 24:342–352.
Rangel A, Hare T (2010) Neural computations associated with goal-directed choice. Curr Opin Neurobiol 20:262–270.
Read SJ, Marcus-Newhall A (1993) Explanatory coherence in social explanations: a parallel distributed processing account. J Pers Soc Psychol 65:429–447.
Rescorla RA, Wagner AR (1972) A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II: current research and theory (Black AH, Prokasy WF, eds), pp 64–99. East Norwalk, CT: Appleton-Century-Crofts.
Roy M, Shohamy D, Wager TD (2012) Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn Sci 16:147–156.
Rudebeck PH, Saunders RC, Lundgren DA, Murray EA (2017) Specialized representations of value in the orbital and ventrolateral prefrontal cortex: desirability versus availability of outcomes. Neuron 95:1208–1220.e5.
Rugg MD, King DR (2018) Ventral lateral parietal cortex and episodic memory retrieval. Cortex 107:238–250.
Sachs O, Weis S, Krings T, Huber W, Kircher T (2008) Categorical and thematic knowledge representation in the brain: neural correlates of taxonomic and thematic conceptual relations. Neuropsychologia 46:409–418.
Schiller D, Eichenbaum H, Buffalo EA, Davachi L, Foster DJ, Leutgeb S, Ranganath C (2015) Memory and space: towards an understanding of the cognitive map. J Neurosci 35:13904–13911.
Schuck NW, Cai MB, Wilson RC, Niv Y (2016) Human orbitofrontal cortex represents a cognitive map of state space. Neuron 91:1402–1412.
Seghier ML (2013) The angular gyrus: multiple functions and multiple subdivisions. Neuroscientist 19:43–61.
Shannon CE (1948) A mathematical theory of communication. Bell System Technical J 27:379–423.
Sharot T, Riccardi AM, Raio CM, Phelps EA (2007) Neural mechanisms mediating optimism bias. Nature 450:102–105.
Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, Petersen SE (2014) Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Hum Brain Mapp 35:1981–1996.
Somerville LH, Kelley WM, Heatherton TF (2010) Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. Cereb Cortex 20:3005–3013.
Tamir DI, Hughes BL (2018) Social rewards: from basic social building blocks to complex social behavior. Perspect Psychol Sci 13:700–717.
Tan KM, Daitch AL, Pinheiro-Chagas P, Fox KCR, Parvizi J, Lieberman MD (2022) Electrocorticographic evidence of a common neurocognitive sequence for mentalizing about the self and others. Nat Commun 13:1919.
Thagard P (1989) Explanatory coherence. Behav Brain Sci 12:435–467.
Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29:1310–1320.
Vaidya AR, Badre D (2020) Neural systems for memory-based value judgment and decision-making. J Cogn Neurosci 32:1896–1923.
van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN (2012) How schema and novelty augment memory formation. Trends Neurosci 35:211–219.
Wagner DD, Haxby JV, Heatherton TF (2012) The representation of self and person knowledge in the medial prefrontal cortex. Wiley Interdiscip Rev Cogn Sci 3:451–470.
Will GJ, Rutledge RB, Moutoussis M, Dolan RJ (2017) Neural and computational processes underlying dynamic changes in self-esteem. Elife 6: e28098.
Wilson RC, Collins AG (2019) Ten simple rules for the computational modeling of behavioral data. Elife 8:e49547.
Woo C-W, Krishnan A, Wager TD (2014a) Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage 91:412–419.
Woo C-W, Koban L, Kross E, Lindquist MA, Banich MT, Ruzic L, Andrews-Hanna JR, Wager TD (2014b) Separate neural representations for physical pain and social rejection. Nat Commun 5:5380.
Woolrich M (2008) Robust group analysis using outlier inference. Neuroimage 41:286–301.
Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370–1386.
Woolrich M, Behrens TEJ, Beckmann CF, Jenkinson M, Smith SM (2004) Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage 21:1732–1747.
Yang J, Xu X, Chen Y, Shi Z, Han S (2016) Trait self-esteem and neural activities related to self-evaluation and social feedback. Sci Rep 6:20274.
Yoon L, Somerville LH, Kim H (2018) Development of MPFC function mediates shifts in self-protective behavior provoked by social feedback. Nat Commun 9:3086.
Zeithamova D, Mack ML, Braunlich K, Davis T, Seger CA, van Kesteren MTR, Wutz A (2019) Brain mechanisms of concept learning. J Neurosci 39:8259–8266.
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. Soc Cogn Affect Neurosci 15:695–707.
Zhang Y, Brady M, Smith S (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20:45–57.
Zhou L, Smith K, Tenenbaum J, Gerstenberg T (2023) Mental Jenga: a counterfactual simulation model of physical support. PsyArXiv. Advance online publication. Retrieved April 10, 2023. .