Title: Prefrontal cortex state representations shape human credit assignment

Authors: Amrita Lamba[1], Matthew Nassar[2,3], Oriel FeldmanHall[1,3]

Affiliations: [1]Department of Cognitive Linguistic & Psychological Sciences, Brown University, Providence, RI, USA. [2]Department of Neuroscience, Brown University, Providence, RI, USA. [3]Carney Institute of Brain Sciences, Brown University, Providence, RI, USA.

Abstract: People learn adaptively from feedback, but the rate of such learning differs drastically across individuals and contexts. Here we examine whether this variability reflects differences in *what* is learned. Leveraging a neurocomputational approach that merges fMRI and an iterative reward learning task, we link the specificity of credit assignment—how well people are able to appropriately attribute outcomes to their causes—to the precision of neural codes in the prefrontal cortex (PFC). Participants credit task-relevant cues more precisely in social compared to nonsocial contexts, a process that is mediated by high-fidelity (i.e., distinct and consistent) neural representations in the PFC. Specifically, the medial PFC and orbitofrontal cortex work in concert to match the neural codes from feedback to those at choice, and the strength of these common neural codes determines learning success. Together this work provides a window into how neural representations drive adaptive learning.

Significance Statement: Successful learning requires selectively attributing outcomes to their cause—a process known as credit assignment. Little is known about how the brain performs this credit assignment, or how the process might differ across contexts or individuals. Functional neuroimaging analyses reveal that precise credit assignment is linked to high fidelity (i.e., distinct and consistent) neural representations of causal cues in the prefrontal cortex (PFC), which supports increased differentiation between stimuli during learning. Our results reveal *why* individuals learn differently: differences are not driven by the magnitude of learning signals (i.e., stronger prediction errors) as has been previously claimed, but by differences in the strength of neural representations to which those learning signals are attributed.

Imagine that you are applying for a job and receive varying feedback on your "pitch" from your interviewers. Which aspects of your pitch bear repeating for future interviews when similar behaviors across different settings produce a wide distribution of outcomes? This example encapsulates the inherent difficulty of accurately linking outcomes to specific actions, particularly when the causal structure of the world is unknown and decision-irrelevant outcomes also occur in close temporal proximity. In such scenarios, humans and animals are thought to group action-outcome contingencies together based on causal cues that reflect states in the environment, allowing outcomes to be selectively linked to specific states (Collins & Frank, 2013; Gershman, Norman, & Niv, 2015). However, because learners have yet to discover the underlying generative structure of outcomes, and because learning and memory systems are fallible, outcomes can be misattributed and spread to irrelevant states—a challenge known as the structural credit assignment problem (Hamid, Frank, & Moore, 2021; Sutton, 1984). In these cases, information gets smeared in memory (Vaidya & Fellows, 2016) which can result in overgeneralization. Despite this known learning challenge (Sutton, 1984), conventional reinforcement learning (RL) algorithms applied to human behavior typically assume perfect credit attribution for each outcome observed. Thus, an open question is how does the human brain successfully bind observed outcomes to the appropriate causal state to solve the credit assignment problem?

Non-human animal research hints that the prefrontal cortex (PFC) might play an integral role in credit assignment by binding state and action-value representations (Asaad, Lauro, Perge, & Eskandar, 2017), which could then be reinforced through midbrain dopaminergic signals to selectively gate reward attribution (O'Reilly & Frank, 2006). Tracking state-contingent outcome history would also be critical to properly assigning credit, which is believed to be governed by the lateral orbitofrontal cortex (lOFC; Chau et al., 2015; Jocham et al., 2016; Walton, Behrens, Buckley, Rudebeck, & Rushworth, 2010). Indeed, humans with lesions in the lOFC exhibit reduced state-contingent reward learning (Noonan, Chau, Rushworth, & Fellows, 2017) and display a greater tendency to misattribute rewards to irrelevant causal factors (Vaidya & Fellows, 2016). More recent work shows that the medial PFC and lOFC jointly track latent states (Schuck, Cai, Wilson, & Niv, 2016) by leveraging surprise signals (Nassar, McGuire, Ritz, & Kable, 2019), allowing for credit assignment to be performed for both experienced (Akaishi, Kolling, Brown, & Rushworth, 2016) and unobserved outcomes (Boorman, Witkowski, Zhang, & Park, 2021; Witkowski, Park, & Boorman, 2022).

While prior work across species suggests that the PFC is involved in tracking latent states, it is currently not known whether the configuration of neural patterns encoding causal cues reflect distinct forms of learning. For example, unlike an eligibility trace, it is possible that learners *actively* represent task-relevant states during learning, enabling selective binding of outcomes to states in memory. By allowing temporally disparate actions and outcomes to be neurally bound to the relevant state identity, these sustained neural patterns may support increased discrimination between cue-specific decision policies. If this were the case, the success of credit assignment may be contingent on the structure and fidelity—i.e., degree of distinctiveness and consistency of each representation over time—of state representations in the PFC. Thus, to effectively guide credit assignment, a distinct and persistent neural code representing the current state should emerge both during choice and when feedback is delivered. Failures to properly encode high fidelity state identities during both choice or feedback should therefore result in increased misattribution and credit spreading.

In the current study we leverage a computational neuroimaging framework by combining RL models with representational similarity analysis (RSA) to investigate whether distinct forms of credit assignment can be discerned from neural patterns in the PFC. Our modeling framework allows us to estimate the precision of credit assignment from behavior, while RSA allows us to directly measure the content and fidelity of evoked neural state representations. We then link the fidelity of these neural representations during choice and feedback to how well an individual assigns credit across different contexts. For example, during social situations humans are often able to exploit social feedback to quickly infer latent structure (Lamba, Frank, & FeldmanHall, 2020; Lau, Gershman, & Cikara, 2020; Park, Miller, & Boorman, 2021; van Baar, Nassar, Deng, & FeldmanHall, 2022), which may allow credit to be assigned more selectively to specific individuals. Alternatively, when learning in less familiar environments with an unknown causal structure (e.g., gambling) learners may assign credit less precisely. We leverage these potential differences to evaluate whether the brain adaptively adjusts the fidelity of state representations across contexts to mediate the selectivity of causal learning.

Participants played an iterative, multiplayer social learning task which requires participants to distinguish between trustworthy and untrustworthy partners when making strategic monetary decisions, as well as a matched nonsocial gambling task. We developed a RL algorithm to capture how precisely outcomes are attributed to specific states, which in our task are denoted by distinct stimulus' identities (i.e., partners and bandits)—observing that different learning profiles are due to how accurately individuals assign credit. Furthermore, some participants consistently spread credit across states, particularly in the bandit task, and when receiving negative feedback. Multivariate neural patterns in prefrontal regions, including the lOFC and mPFC, encode state representations, but do so less precisely in those who spread credit. In contrast, high-fidelity state representations were associated with greater task earnings and more precise credit assignment during choice. Neural state representations share a common geometry across choice and feedback, signifying a persistent neural code indexing state identity.

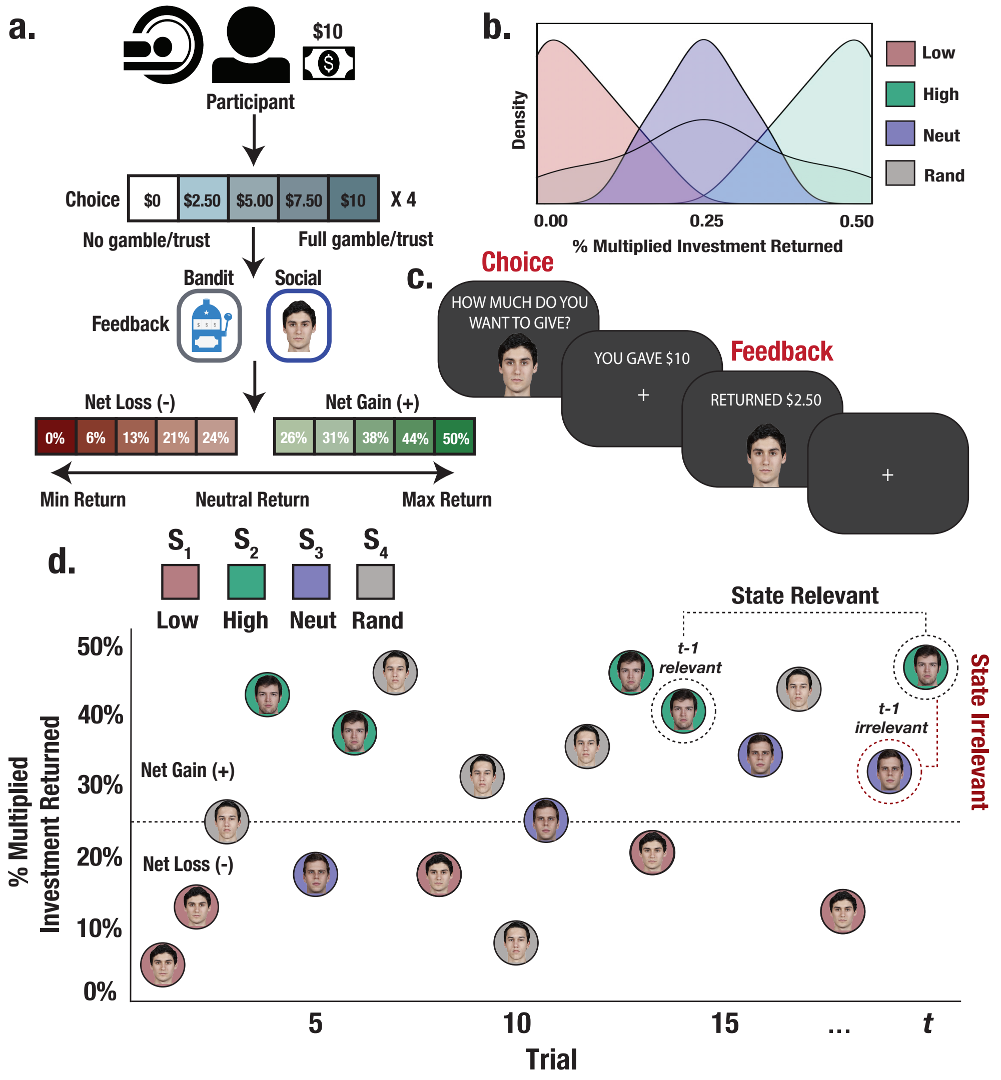


Fig. 1. ***A.*** *Social and bandit tasks. Participants played 15 trials with each partner/bandit while in the scanner. On each trial, participants were paired with one of the four stimuli (partner/bandit) and given $10 to invest using a 5-button response box to indicate their investment in $2.50 increments. The monetary investment was then quadrupled, and partners/bandits returned anywhere from 0%-50% of the money received, allowing for the possibility to double one's earnings, lose the full investment, or any outcome in between.* ***B.*** *Task reward structure. Stimuli were randomly assigned to respond with fixed reward rates generated from one of four outcome distributions. Each stimulus deterministically returned less than the participant initially invested (low), more (high), an amount close to the initial investment (neutral), or a random amount.* ***C.*** *Task event sequence. Participants were given up to 3 seconds to indicate their choice, after which they experienced a jittered inter-stimulus delay. The returned investment was then displayed on the screen for a fixed 2 second duration.* ***D.*** *Within-task stimulus presentation. Trials were randomly interleaved such that interactions with each stimulus could occur anywhere from 1 to 15 trials apart, allowing us to probe learning effects from relevant versus temporally adjacent irrelevant outcomes.*

Results

Humans are faster to learn payoff maximizing strategies in the social domain. Participants (N = 28) completed 60 trials of the Trust Game and a matched bandit task (order counterbalanced), while undergoing functional neuroimaging (fMRI). Participants made a series of monetary decisions with partners and bandits that varied in their reward rate (Fig. 1, A to D). Learning success was evaluated by maximizing monetary gains by investing the full $10 with the high return stimulus and minimizing monetary losses by investing $0 with the low return stimulus. Despite being perfectly matched across social and nonsocial conditions, participants invested more money with the high return social partner compared to the bandit (mean social investment: $7.12; mean bandit investment: $6.43; $t = -3.57$, $p < .001$; Fig. 2A) and less money with the low return partner vs. bandit (mean social investment: $1.71; mean bandit investment: $2.71; $t = 5.69$, $p < .001$), with no differences in mean investments for the neutral or random stimuli across tasks (all Ps > .1).

Investments are influenced by history of relevant and irrelevant prior outcomes. To shed light on the credit assignment problem, we used a series of time-lagged regression models to examine how participants use relevant and irrelevant outcomes to guide learning. We paired each stimulus with its previous outcome and modeled the effect of three stimulus-matched (i.e., relevant) prior interactions on current investments (Fig. 1D). We observed that the most recently experienced relevant outcome exerted the largest effect on investments (significant difference between slopes at t-1 vs. t-2: $t = -4.98$, $p < .001$; Fig. 2B), while also exerting a stronger effect for social partners, compared to the prior outcome of a bandit (a significant effect at t-1 in which slopes were larger for the social vs. bandit task: $t = 3.19$, $p = .004$; Fig. 2B). To investigate whether temporally adjacent but irrelevant outcome history biased decisions, we yoked each investment to the immediately preceding outcome, irrespective of its identity. Although these outcomes should not inform choices on the current trial given the generative task structure, we observed that recent outcomes (mean slope from t-1 through t-3) influenced decisions to gamble with bandits ($t = 1.83$, $p = 0.039$; Fig. 2B), whereas recent irrelevant outcomes were anticorrelated with choices in the social task ($t = -2.73$, $p = 0.011$), indicating that feedback from a partner in the social domain was not attributed to other partners, and these task differences in using irrelevant outcome history to guide investments were significant ($t = 3.42$, $p < .001$; Fig. 2B). We also found a valence-dependent effect of outcome history on learning. Across tasks, prior relevant outcomes that were rewarding exerted a stronger influence on investments than losses (significant difference between slopes for gains versus losses; $t = -4.27$, $p < .001$; Fig. 2C). In contrast, immediately preceding irrelevant losses disproportionately impacted decisions compared to gains ($t = 2.21$, $p = .029$; Fig. 2C), indicating that outcome misattribution is, in part, valence-dependent. Put simply, learning from relevant outcomes was asymmetrically driven by gains, whereas outcome misattribution stemmed disproportionately from losses.

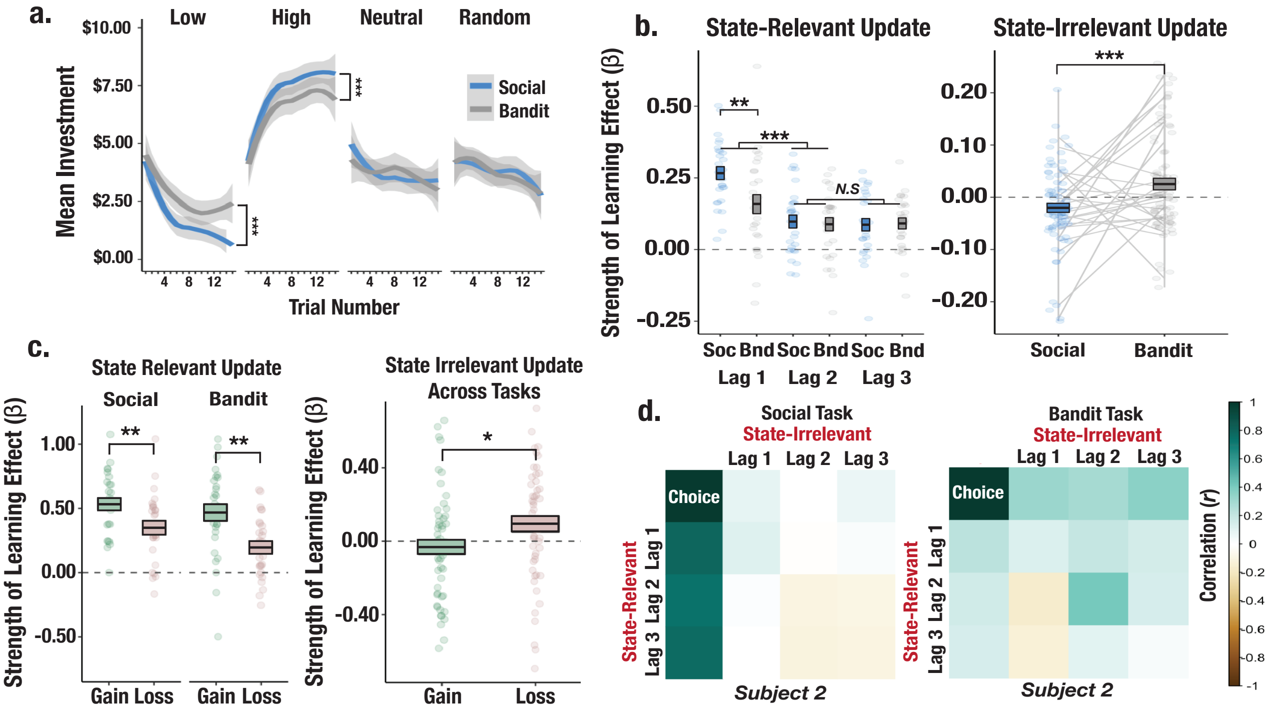


Fig. 2. A. *Learning curves from social and bandit tasks. Predicted investments over trials computed from fixed-effects regression model show faster learning in the social task. Shaded gray regions correspond to the standard error of the mean of the regression line.* ***B****. Effect of relevant and irrelevant outcome history on choice. Model terms show increased learning from the most recent relevant outcome in the social task and an increased effect of irrelevant outcomes on investments in the bandit task. The box-length denotes the standard error of the mean, and the black line corresponds to the mean beta estimate for the lag term.* ***C****. Effect of valence-dependent outcome history on choices. Relevant prior gains compared to losses exerted a greater influence on investments.* ***D****. Correlation matrix of relevant and irrelevant outcomes on investments for a prototypical participant. The participant shows a strong pattern of learning exclusively from relevant outcomes in the social task but applies irrelevant outcomes to learning in the bandit task. Asterisks (*,**,***) denote p < .05, p < .01, p < .001, respectively.*

Modeling credit assignment captures domain and valence-dependent learning asymmetries. Given that we observed asymmetrical learning profiles across domains and outcome valence, we probed whether these divergent learning profiles were due to differences in successfully assigning credit. Prior work suggests that interference effects (e.g., the influence of irrelevant outcomes), can emerge through lapses in attention and working memory that cause forgetting between task-relevant events (Collins & Frank, 2012), and through misattribution of outcomes to irrelevant causes (Vaidya & Fellows, 2016). We used model comparison to evaluate the relative contributions of these factors to task performance.

We implemented RL using a continuous choice, logistic function Q-learning algorithm with valence-dependent learning rates (see Methods). Our model set included: 1) a base model in which learning was perfectly assigned to the appropriate states, 2) a credit assignment model that includes a parameter controlling the degree to which outcomes affect the Q-value of irrelevant states, 3) a two parameter credit assignment model in which ca parameters were fit separately for positive versus negative prediction errors (PEs; v-ca model), and 4) a model in which learned values decay incrementally on each trial to capture forgetting. Behavior was best fit by the credit assignment model with valenced ca parameters (v-ca) in both the social and bandit task (see Methods and Supplement for model comparison). Consistent with the analyses above, our model reveals that people assigned credit more precisely to partners in the social task, and spread credit more diffusely across bandits in the gambling task (social task mean ca estimate = 0.74; bandit task mean ca estimate = 0.54; *F* = 10.67, *p* = .002; Fig. 3B)—an effect that was heightened for gains compared to losses (*F* = 4.72, *p* = .033; Fig. 3C). Thus, our best fitting model revealed that learning differences were not attributed to forgetting but instead were linked to credit assignment errors.

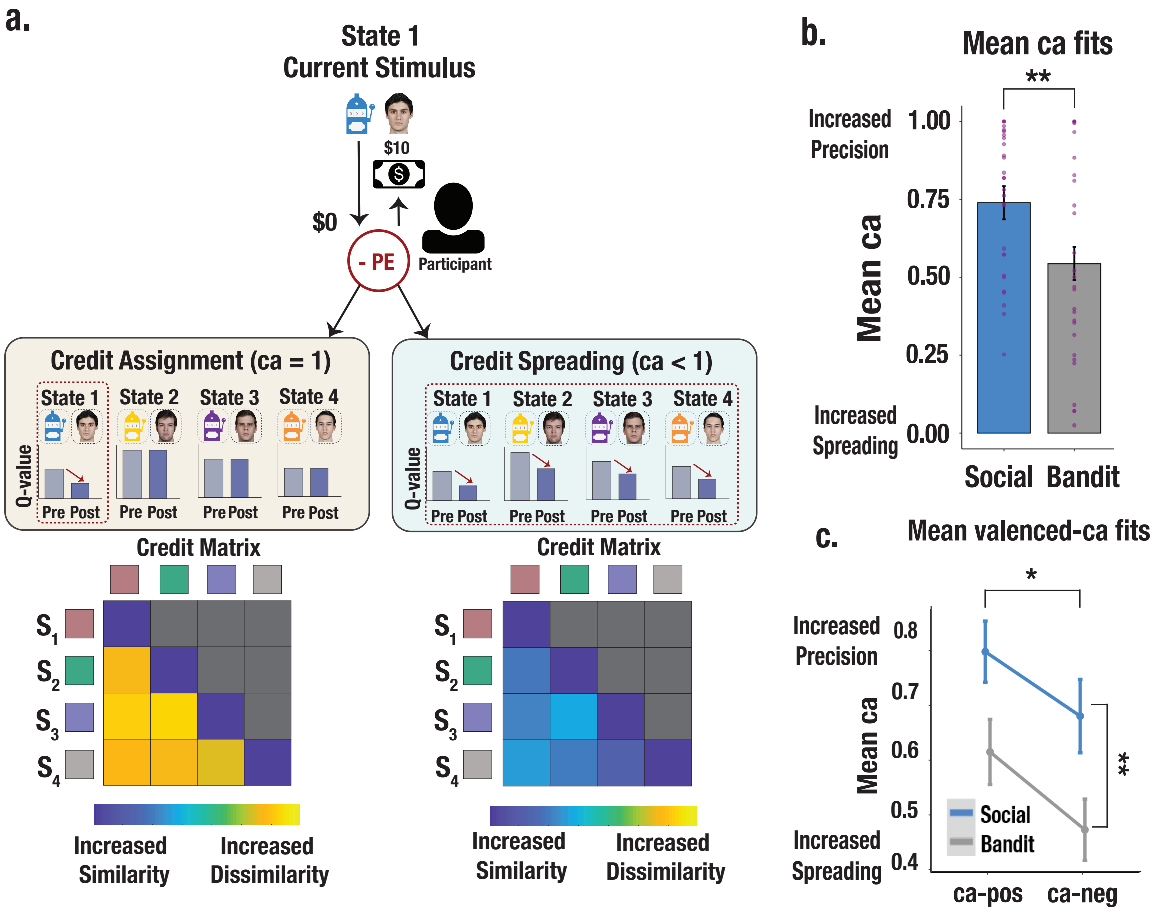


*Fig. 3. **A**. Schematic of credit assignment and spreading mechanisms. In a perfect credit assignment scenario (left side), PEs only update the expected value of the current state. A resulting credit matrix showing how credit assignment impacts the overall similarity between states shows that credit assignment values closer to 1 (perfect credit assignment) would result in no credit spreading among other states (1-ca), therefore allowing for increased differentiation in the state space. Conversely, in a credit-spreading scenario (right side), PEs are used update the expected value of the current state and all other states. A credit matrix would therefore predict higher off diagonal terms (1-ca values indicating complete spreading of outcomes across states), resulting in less differentiation between states. **B**. Credit assignment parameter estimates across social and bandit tasks. Mean credit assignment fits show more accurate credit assignment in the social task and increased spreading in the bandit task. Purple dots show individual parameter estimates and error bars denote the standard error of the mean. **C**. Valenced credit assignment parameter estimates. Parameter fits from our valenced ca model show more precise credit assignment for gains (ca-pos) and more spreading for losses (ca-neg).*

Credit assignment predicts the fidelity of state representation during choice. Identifying and characterizing the neural circuitry supporting successful credit assignment compared to credit spreading can further clarify how this mechanism is precisely deployed. We test the prediction that credit assignment requires a high fidelity (i.e., distinct, and consistent) neural representation of a stimulus' identity during choice, which should in theory support retrieval of the state-specific decision policy. Using a whole brain searchlight, we extracted single trial coefficients of the neural pattern on each trial and for each stimulus to create a neural representational dissimilarity matrix (RDM), separately for choice and feedback (Fig. 4A; see Methods). We then computed the correlation distance between each searchlight RDM and our state identity hypothesis matrix, controlling for additional regressors in regions of interest (ROIs; separate ROIs constructed from choice and feedback searchlights) that survived correction for multiple comparisons (Fig. 4B; see Methods). This enabled us to evaluate the degree to which neural patterns differed across identities, while also measuring the extent to which neural patterns were consistent across trials, thus serving as a high-fidelity "stamp" of the stimulus' identity.

Neural patterns in a constellation of brain regions including the mPFC, lOFC, and mOFC met basic criteria for providing state representations (statistically significant beta coefficients of the identity RDM; Fig. 5B: see Supplement for full list of ROIs and their coordinates) at the time of choice. Within these regions, stimulus identity was more strongly encoded in the social vs. bandit task across ROIs (social task mean $\beta = 0.019$; bandit task mean $\beta = 0.012$; $t = -3.64$, $p < .001$; Fig. 5B), consistent with the findings from our model. There was also a positive relationship between an individual's mean ca parameter estimates and the strength of the stimulus identity representation in the PFC ROIs across both tasks ($t = 3.38$, $p < .001$; Fig. 5D) which did not depend on outcome valence (ca positive: $t = 3.26$, $p = 0.0014$; ca negative: $t = 2.50$, $p = 0.014$)—revealing that individuals who assigned credit more selectively to the relevant stimulus identity also had more consistent and distinct neural representations of it. Furthermore, the more money earned in the task, the more robust these stimulus representations were in the PFC ROIs in both tasks ($t = 4.14$, $p < .001$). These results accord with the prediction that the fidelity of state representations in the PFC during choice control the precision of credit assignment by supporting increased differentiation between state-specific decision policies.

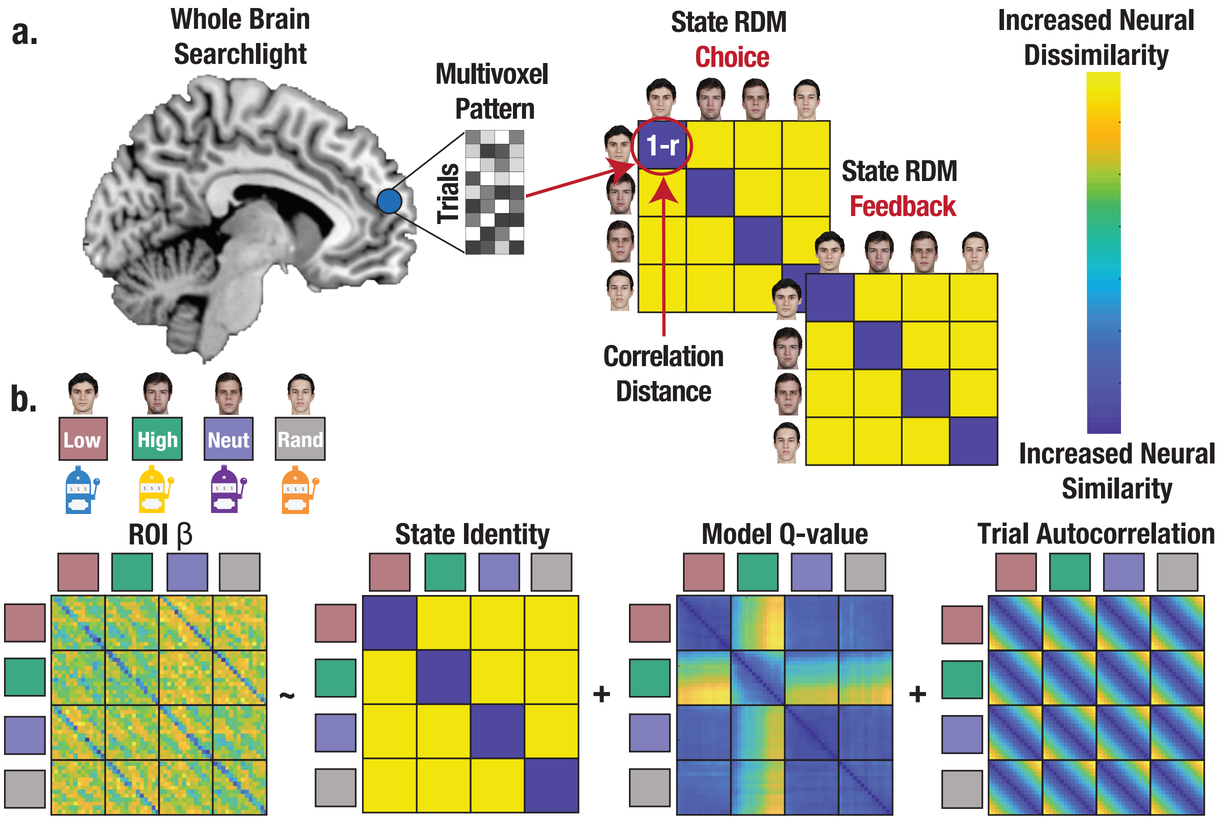
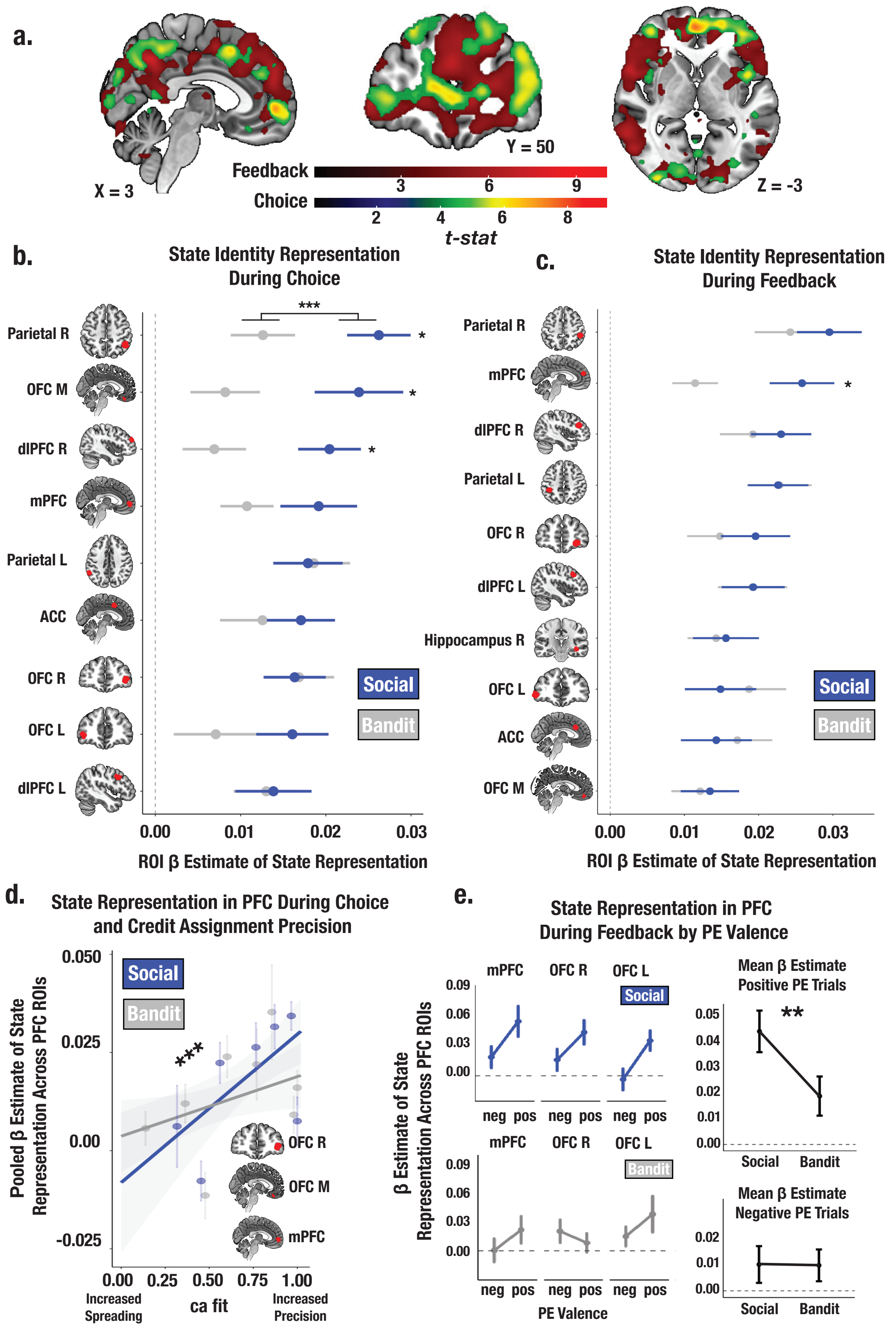


***Fig. 4. A**. Conceptual depiction of RSA methods with whole brain searchlight. Multivoxel patterns were extracted for all trials and reorganized into a correlation distance (1-r) matrix with trials nested within stimulus identities for each task. State representation was evaluated separately for choice and feedback. **B**. Regression approach estimating state representation in neural ROIs, controlling for expected value (Q) and trial autocorrelation.*

Reward enhances encoding of state representations during feedback. At the time of feedback, a stimulus' identity must be sufficiently encoded so that outcomes can be linked to the appropriate prior action. Using our searchlight approach, we identified a suite of regions providing state representations during feedback (Fig. 5C). We observed stronger identity representations in the social task uniquely in the mPFC (t = -2.41, p = 0.023; the fidelity of state representations did not differ across tasks in any additional ROIs). We then examined whether positive vs. negative PEs differentially modulate the strength of stimulus-encoding during feedback, by estimating the strength of stimulus identity encoding within our prefrontal ROIs separately for positive and negative PE trials for each participant (Methods). Across tasks we observed stronger stimulus encoding during positive vs. negative PE trials in the mPFC and lOFC (t = 3.25, p < .0013; Fig. 5E), an effect that was significantly more robust in the social compared to bandit task (t = -2.71, p = 0.008). Together, this suggests that valence asymmetries in learning emerge because reward enhances the strength of state encoding in the PFC, especially in social contexts.



a.
X = 3
Y = 50
Z = -3
Feedback
3
6
9
Choice
2
4
6
8
t-stat
b.
State Identity Representation
During Choice
***
Parietal R
*
OFC M
*
dlPFC R
*
mPFC
Parietal L
ACC
OFC R
OFC L
dlPFC L
Social
Bandit
0.00
0.01
0.02
0.03
ROI β Estimate of State Representation
c.
State Identity Representation
During Feedback
Parietal R
mPFC
*
dlPFC R
Parietal L
OFC R
dlPFC L
Hippocampus R
OFC L
ACC
OFC M
Social
Bandit
0.00
0.01
0.02
0.03
ROI β Estimate of State Representation
d.
State Representation in PFC During Choice
and Credit Assignment Precision
Pooled β Estimate of State
Representation Across PFC ROIs
0.050
0.025
0.00
-0.025
Social
Bandit
***
OFC R
OFC M
mPFC
0.00
0.25
0.50
0.75
1.00
Increased
Spreading
ca fit
Increased
Precision
e.
State Representation in PFC
During Feedback by PE Valence
β Estimate of State
Representation Across PFC ROIs
mPFC
OFC R
OFC L
Social
0.09
0.06
0.03
0.00
neg pos
neg pos
neg pos
mPFC
OFC R
OFC L
Bandit
0.09
0.06
0.03
0.00
neg pos
neg pos
neg pos
PE Valence
Mean β Estimate
Positive PE Trials
**
0.05
0.04
0.03
0.02
0.01
0.00
Social
Bandit
Mean β Estimate
Negative PE Trials
0.02
0.01
0.00
Social
Bandit

Fig. 5. ***A***. *Task-summed group-level t-map of state representation during choice and feedback; image is thresholded at the cluster-level (P_{FWE} < .05) and at peak level (p < .0001).* ***B.*** *Beta estimates of state representation across ROIs identified from choice phase searchlight, disaggregated by social and bandit tasks, indicate stronger state representation in the social task across ROIs. Individual asterisks denote a significant within-subject effect for the specified ROI.* ***C***. *State representation estimates in ROIs from feedback searchlight. Task differences only emerge in the mPFC.* ***D.*** *Predictive association between individual credit assignment parameter estimates and the fidelity of state representation in mPFC, lPFC, and mOFC during choice. Across both social and bandit tasks, credit assignment predicts the strength of state representation observed within prefrontal ROIs shown in panel B.* ***E.*** *Effect of PE valence on state encoding in the PFC during feedback. Positively valenced PEs were associated with higher-fidelity state representations across PFC ROIs, particularly in the social task.*

Successful credit assignment hinges on shared representational geometry between choice and feedback. How is information from choice and feedback integrated to support learning? We consider the possibility that credit assignment is achieved by neurally binding outcomes to states, which may occur by matching identity representations from the last relevant outcome to the next relevant choice—potentially in the mPFC and lPFC given their observed involvement. Greater alignment of neural representations across choice and feedback could preserve a common neural code of the state identity, supporting increased credit assignment precision. To examine the shared representational structure between these timepoints and the extent to which increased alignment reflects a common state identity representation, we identified conjunction ROIs from voxels that survived permutation testing in both choice and feedback searchlight analyses (Fig. 6, A to B). We then computed the degree of neural pattern similarity between choice and feedback representations in these ROIs—all localized to the PFC—and evaluated the degree to which the shared geometry preserved information about the stimulus' identity (Methods). Across both tasks, we observed a significant positive relationship between an individual's credit assignment precision (ca parameters) and the consistency of their stimulus representations across both timepoints in the mPFC and lPFC (t = 2.17, p = .033; Fig. 6C), an effect that was selectively enhanced for gains but not losses (t = 3.48, p < .001). Thus, the precision of credit assignment, particularly for rewarding outcomes, was strongly linked to increased neural binding between feedback and choice —a process that is supported by shared geometry of state representations across distinct phases of learning.

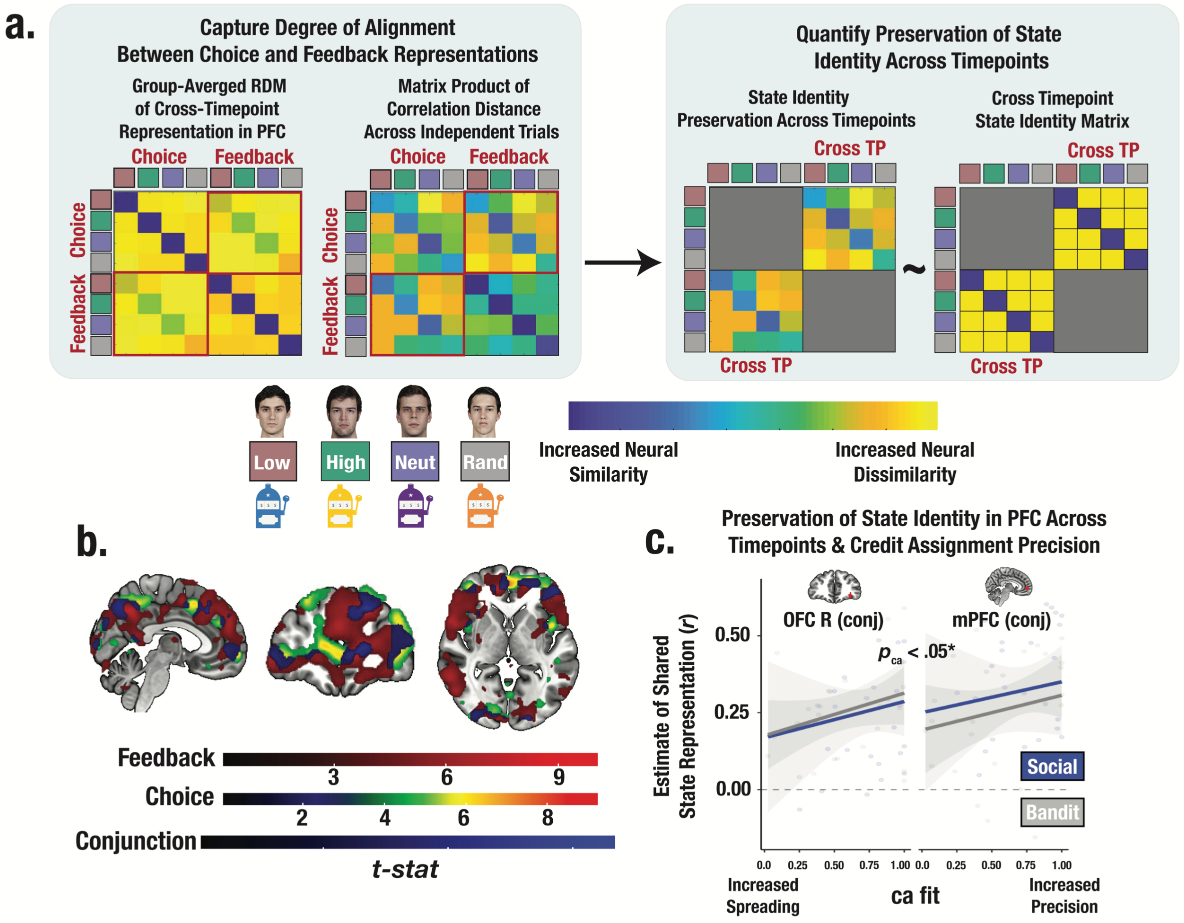


***Fig. 6. A**. Conceptual depiction of cross-timepoint RSA. For each participant, cross-timepoint matrices were constructed as the correlation distance between even and odd trial neural RDMs (Methods). Cross-timepoint cells from the matrix were selected and then correlated with a cross-timepoint identity matrix to estimate the degree to which the shared structure within the neural patterns across choice and feedback reflected a stimulus' identity. **B**. Task-summed group-level t-maps displaying results of conjunction contrasts (Methods). Group-level image is thresholded at the cluster-level ($P_{FWE} < .05$) and at peak level ($p < .0001$). **C**. Predictive association between individual credit assignment estimates and the consistency of identity representational structure in mPFC and lOFC (i.e., conjunction ROIs) across choice and feedback. * Denotes the effect of ca on the pooled estimate of shared state representation across ROIs ($p < .05$).*

Discussion

Adaptable learning systems, whether human, animal, or artificial intelligence, must be able to exploit causal structure by differentiating between an array of spatial and temporal cues, allowing learners to balance between behavioral flexibility and specificity (Asaad et al., 2017; Soto, Gershman, & Niv, 2014; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Walton et al., 2010). Structural credit assignment in particular, enables learners to integrate these contingencies into a set of learned decision policies that are flexibly recruited in response to predictive features and cues in the environment (O'Reilly & Frank, 2006; Sutton, 1984). Here we show *how* humans achieve successful structural credit assignment. First, people are more accurate when attributing outcomes to other people than they are to bandits, an effect that is enhanced for gains and attenuated for losses. In addition, our study reveals a simple neural mechanism for implementing credit assignment. State representations must be sufficiently encoded during feedback, and we find that this process is strengthened by gains. During choice, differences in learning are expressed by the fidelity of these state representations, consistent with the prediction that high fidelity state representations support increased state discrimination and improved learning specificity. This functional division of labor and coordination during distinct phases of learning aligns with our finding that the degree of shared geometry in state representations in the PFC predicts how precisely participants assign credit. Taken together, we provide novel evidence that the PFC serves as a hub for credit assignment by leveraging reward signals and organizing state representations into a shared neural code, allowing for the efficient transfer of credit from feedback to subsequent choice. Our key findings converge with prior work demonstrating that lOFC and mPFC track latent states (Schuck et al., 2016), particularly those that govern the structure of rewards so that humans can respond flexibly and adaptively to shifting environmental demands (Jocham et al., 2016; Nassar et al., 2019; Witkowski et al., 2022).

We also observed that outcome valence asymmetrically shapes the specificity of assigning credit, which results in more accurate outcome attribution for gains and increased credit spreading for losses. While past studies offer insight into credit assignment for rewards specifically (Asaad et al., 2017; Hamid et al., 2021; Jocham et al., 2016; Walton et al., 2010), no studies that we are aware of have documented asymmetrical effects of PE valence driving the precision of credit assignment. Thus, these results offer an interesting parallel to existing stimulus generalization theories. Prior work using classical conditioning paradigms, in which a neutral stimulus is paired with an aversive outcome, shows that the conditioned response is also evoked by novel stimuli (Dunsmoor & Paz, 2015; Hull, 1943; Schechtman, Laufer, & Paz, 2010). Transfer effects to a new stimulus follows a generalization gradient, in which the experienced intensity of the prior aversive outcome predicts increased threat generalization (Dunsmoor, Kroes, Braren, & Phelps, 2017; Lissek et al., 2005), and greater perceptual generalization (Davis, Walker, Miles, & Grillon, 2010; FeldmanHall et al., 2018). Although our paradigm articulates generalization through credit spreading mechanisms, from a signal detection standpoint, these findings undoubtedly dovetail with that notion that it is 'better to be safe than sorry' in the aversive domain (Dunsmoor & Paz, 2015). This may help to explain why increased credit spreading of negative PEs across irrelevant cues can become pathological and maladaptive, offering potential inroads into understanding the etiology of generalized anxiety disorders.

While this work unveils a generalizable computational and neural mechanism for structural credit assignment, there are a number of unanswered questions that future work can help address. In particular, valence-asymmetric credit assignment effects may have interesting mappings onto dopamine modulation in the striatum and amygdala. Recent work in mice finds that wave-like dopamine signals from the dorsal striatum communicate when successful actions performed during instrumental learning are necessary for performance, offering insight into the underlying neuromodulatory dynamics of credit assignment in the reward domain (Hamid et al., 2021). Conversely, prior work suggests that dopamine in the amygdala gates the selectivity of an acquired threat response, whereas inhibition of amygdala dopamine receptors is linked to threat overgeneralization (De Bundel et al., 2016). Future work should consider how midbrain dopamine modulation in the striatum and amygdala mechanistically impacts the specificity of downstream state representations. Notably, our whole brain searchlight also picked up state representations in other prominent cortical networks, such as the control network (lateral parietal, anterior cingulate, and dorsal lateral prefrontal regions), and may accord with the possibility that distinct functional networks encode abstract state representations that vary only in format to optimize for differing task demands (Vaidya & Badre, 2022). Future work could consider how prefrontal and control networks interact during successful credit assignment, particularly when abstract state representations are required to perform complex sequences of actions. To summarize, our results identify a simple and domain-general neural mechanism for credit assignment in which outcomes and states are temporally bound together in the PFC, revealing a biologically grounded model for how humans assign credit to causal cues encountered in the world. How this mechanism coordinates with other known neural systems and deviates in psychiatric disorders has yet to be uncovered.

Methods

Participants. Data was collected from 30 right-handed adults (ages 21-36; mean age = 23.5, N_{female} = 16) in the Providence, Rhode Island area. Our study protocol was approved by Brown University's Institutional Review Board (Protocol #1607001555) and all participants indicated informed consent before completing the social and the bandit tasks in the scanner. After all fMRI preprocessing steps were completed, two participants were removed from the final sample due a high degree of motion artifacts (movement > 3mm). All participants received monetary compensation ($15/hour) and additional performance-based bonus payment of up to $20.

Instructions and stimulus presentation. Prior to scanning, all participants were given instructions for the social and bandit tasks and instruction ordering was counterbalanced depending on which task participants completed first in the scanner. For the social task, participants were told they would see the faces of previous participants who had already indicated the proportion of the investment they wished to return and who had been photographed prior to leaving their session. In reality, each of the four face stimuli were drawn from the MR2 database (Strohminger et al., 2016). All face stimuli included in our task were prejudged to be equivalent on trustworthiness and attractiveness dimensions by independent raters (Strohminger et al., 2016). Slot machine stimuli varied by visually distinct colors (purple, blue, yellow, and orange).

For each task, participants were required to pass a basic comprehension check to ensure that they understood the payoff structure of each game. Stimuli were presented using Psychtoolbox in MATLAB 2017a. Each trial was designed to elapse over a 16 second duration. The trial was initiated with a choice phase with a 3-second response window in which participants indicated their investment using a 5-button response box (options: $0, $2.50, $5.00, $7.50, $10.00). After participants keyed in their response, a jittered interstimulus interval (ITI), randomly distributed between 1-5 seconds, reminded participants of their investment. The outcome was then presented for a fixed 2-second duration, following by an additional ITI, filled with however many seconds remained for the full 16 second trial duration (between 6-13 seconds). If participants failed to indicate their investment within the 3-second response window, the investment was considered $0, and a missed trial prompt appeared during the ITI. Missed trials were omitted from all behavioral and RSA analyses. Stimulus ordering was randomly interleaved, and therefore consecutive presentations of the same stimulus could occur anywhere from 1 to 15 trials apart following a right-skewed distribution, such that most consecutive stimulus interactions occurred within 1-5 trials.

Reward structure. Each stimulus in the social and bandit tasks was randomly assigned to follow one of four reward distributions, such that stimulus identities were counterbalanced across different payoff structures. The high reward stimulus always returned more than the participant initially invested and thus always resulted in a net gain, whereas the low reward stimulus always returned a lower amount resulting in a net loss. Neutral and random stimuli were designed to return a roughly equivalent amount and served as a control for outcome valence, allowing us to examine outcome attribution precision with stimuli that resulted in net gains and losses with equal frequency. Neutral and random stimuli only differed in terms of their return variance (i.e., the extremity of gains and losses). Rather than truly sampling from a payoff distribution which could have resulted in vastly different observed outcomes across participants and tasks (e.g., observing a consistent string of gains or losses on the extreme end of the distribution simply due to chance), we preselected the return rates so that the full range of the distribution was sampled from. We then applied these preselected return rates for all participants and in each task but allowed their ordering to be randomized across trials. Notably, although return rates were fixed, the payoff on each trial was still dependent on the participant's investment (see task structure in Fig. 1).

Time-lagged behavioral regression analyses. Single trial investments were modeled using a regression that included the monetary amount returned on previous trials at various time points (i.e., lags) as explanatory variables. To capture individual learning effects, we modeled investment data for each participant separately including both social and bandit tasks in the same linear regression model. Investments were modeled as a weighted sum of previous returns experienced on relevant trials (i.e., those with a matched stimulus) as well as irrelevant trials (i.e., those that immediately preceded a given investment, irrespective of stimulus identity). We fit slopes for the contribution of each lag term using previous returns from the *n*th trial back in each task (social vs. bandit), yielding 13 coefficients per participant (model equation below; *i,t* denotes each participant and trial, respectively).

$$\text{Investment}_{(i,t)} = \beta_0 + \beta_{1,2}\text{Lag}_1\text{Rel}_{(i,t)} \mid \text{task} + \beta_{3,4}\text{Lag}_2\text{Rel}_{(i,t)} \mid \text{task} + \beta_{5,6}\text{Lag}_3\text{Rel}_{(i,t)} \mid \text{task} + \\ \text{B}_{7,8}\text{Lag}_1\text{Irrel}_{(i,t)} \mid \text{task} + \beta_{9,10}\text{Lag}_2\text{Irrel}_{(i,t)} \mid \text{task} + \beta_{11,12}\text{Lag}_3\text{Irrel}_{(i,t)} \mid \text{task}$$

To model whether rewards vs. losses differentially impacted reward attribution, for each participant we separated trials into scenarios in which the previous stimulus-matched and previous irrelevant outcome resulted in a net gain (return > investment) or a net loss (return < investment). Here we considered only lag1 trials to minimize parameter tradeoffs that prevented model convergence, and furthermore ran four separate regression models quantifying the effects of lag1 returns for each combination of relevant vs. irrelevant and gain vs. loss trials.

Logistic Reinforcement Learning model. To better understand trial-to-trial changes in investing, we developed a nested set of Q-learning models to translate trial-outcomes into behavioral updates. Because choices in the task were both discrete and ordinal in their magnitude (choice options: $0, $2.50, $5.00, $7.50, $10.00) we used a logistic function to model the learned value of investing with each partner/bandit type based on trial and error.

Model investment function. In each of our models, all choices in which the participant responded were included in model fitting. We initialized Q values to a prior, estimated for each participant as a free parameter. Predicted investments for each trial were generated from a sigmoid function that included parameters to account for individual investment biases (i.e., baseline differences in investment preferences) and the slope of the relationship between Q-values and predicted investments (m; Fig. M1A), where Q*(t,j)* reflects the Q value for investing in stimulus (i.e., partner/bandit) *j* on trial *t*:

$$pred.investment_t = \frac{max\ investment}{1 + e^{-m(Q_{(t,j)} - bias)}}$$

To get the likelihood with which our model would produce all possible investments on a given trial, we assumed that the probability of a given investment would fall off according to a Gaussian probability density function (PDF) around the predicted investment (Fig. M1B):

$$p(investments) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(all\ investments - pred.investment)^2}{2\sigma^2}}$$

The width of the Gaussian distribution was fixed to a value of 1 in all models, controlling the variability in model investments. The probability of investing—*p*(investment)—generated from the Gaussian PDF was normalized on each trial such that the total probability across investments was equal to 1, and the model was fit by minimizing the negative log of the sum of *p*(investments) corresponding to the actual participant investments across trials.

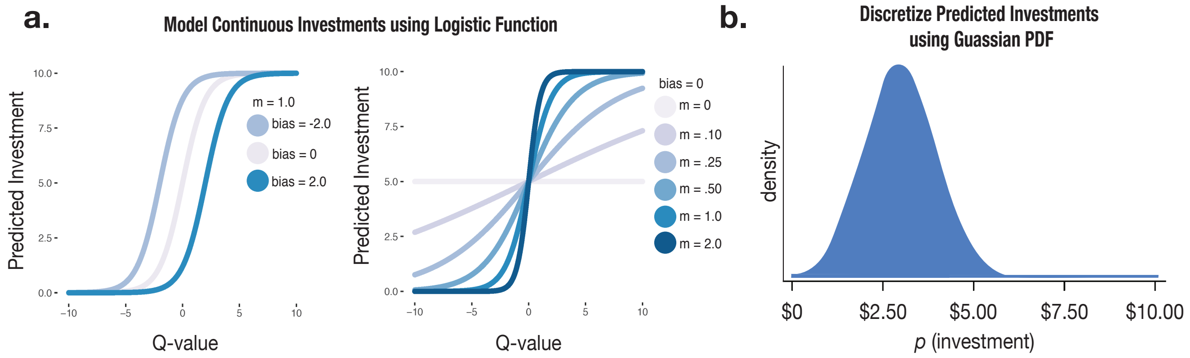


***Fig. M1**. Continuous choice, logistic Q-learning algorithm. **A**. Impact of bias and m parameter estimates on predicted investments in the Q-learning model. Higher bias parameters indicate an increased baseline tendency to invest larger monetary amounts, whereas m parameters capture the degree of evidence learners require before switching investment strategies. **B**. Schematic of Gaussian density function used to convert predicted investments from continuous to discrete ordinal scale and to compute trial log likelihood.*

Modeling learning. In our base model, we implemented RL using the delta rule to compute the reward prediction error (δ) on each trial, which updated the expected value (Q) of investing with each partner/bandit type after each outcome observation. Error-driven learning was then scaled by the learning rate (a):

$$\delta = reward_t - Q_{(t-1,j)}$$

$$Q_{(t,j)} = Q_{(t-1,j)} + a \cdot \delta$$

In the base model, outcomes were always attributed to the appropriate state, effectively performing standard model-free RL. We included additional learning rate, credit assignment, and decay parameters to the base model to construct a set of models that varied in complexity, and that provided distinct conceptual accounts of learning differences, but notably all models mapped experienced outcomes to learned values to guide future choice (see Supplement for a full list of models tested).

To capture individual differences in credit assignment vs. credit spreading we introduced a credit assignment parameter (ca), that quantified the degree to which observed outcomes on the current trial influenced the Q-value of irrelevant causes (i.e., all other partners/bandits not engaged with on the current trial), denoted with the index k.

$$Q_{(t,j)} = Q_{(t-1,j)} + \ a \cdot \delta \ \cdot ca$$

$$Q_{(t,k)} = Q_{(t-1,k)} + \frac{a \cdot \delta \ \cdot (1 - ca)}{n_k}$$

To account for valence-dependent learning effects, we fit models with valenced learning rate parameters, in which PEs greater or less than 0 were scaled by positive or negative learning rates, respectively. We also fit a model with valenced ca terms to capture valence-specific differences in reward attribution (i.e., whether credit assignment vs. spreading is dependent on observing better or worse than expected outcomes). This model was algorithmically identical to the ca model, except that two separate ca parameters were used to account for credit assignment on positive and negative PE trials, respectively.

Decay models. We modeled forgetting effects using a decay model in which forgetting was estimated as overall decay (γ) of learned Q-values to a pre-existing prior between exposures to the current stimulus, where γ and prior were fit as additional free parameters to each participant.

$$Q_{(t,j)} = Q_{(t-1,j)} + \ a \cdot \delta$$

$$Q_{(t,k)} = Q_{(t-1,k)} + \ \gamma \ \cdot (prior - Q_{(t-1,k)})$$

Model comparison. Model fits were then evaluated using the Akaike information criterion (AIC), which we computed as:

$$AIC = \ -2(BayesLL) - \ 2(n \ parameters)$$

We performed model selection by maximizing the negative AIC and minimizing Δ AIC, which was calculated as the difference between each participant's best-fitting model and every other model in the set, allowing us to evaluate model performance penalized for additional terms and the model fit advantage of each participant's best-fitting model relative to the explanatory power of each other model in the set. Thus, the best-fitting model would ideally be able to explain each participant's data approximately as well, if not better, in most instances. Model comparison was performed separately for social and bandit tasks (see Supplement for model performance and validation details).

MRI data acquisition. Data was acquired at the Brown University MRI Facility with the Siemens Prisma 3T MRI Scanner. Anatomical scans were collected using a T1-weighted sequence with $1mm^3$ isotropic voxels, 1900ms TR, flip angle = 9 degrees, 160 slices/volume, 1mm slice thickness, for a duration of 4 minutes, 1 second. Functional scans were acquired using a T2-weighted sequence with $3mm^3$ isotropic voxels, 2000ms TR, flip angle = 78 degrees, 38 slices/volume, 3mm slice thickness. Each task was divided into 2 functional runs, each consisting of 246 volumes, for a duration of 8 minutes and 12 seconds. We used a bounding box with a forward tilt along the AC-PC axis to ensure we were imaging lateral and medial OFC.

Data preprocessing. Data was preprocessed in SPM12 in the following sequence: slice-time correction, realignment, co-registration, segmentation, normalization, spatial smoothing. Images were normalized to a standard MNI template and resampled to $2mm^3$ voxels. For multivariate analyses, images were minimally smoothed using a $2mm^3$ smoothing kernel. RSA images constructed from the deconvolved time-series GLM (see below) were later smoothed using a $6mm^3$ smoothing kernel for group analyses.

Time-series GLM. For each participant we obtained time-series estimates of the BOLD signal by deconvolving the HRF using single trial regressors (Mumford & Poldrack, 2007; Ramsey et al., 2010), concatenated across task runs. We also included separate trial regressors for choice and feedback onsets within each GLM (i.e., choice and feedback onsets were modeled simultaneously) to control for potential temporal correlations in the BOLD signal resulting from consecutive task events. For choice phase regressors, we modeled voxel activations during the choice duration, which occurred within a 3-second window. Feedback phase activations were modeled during a fixed 2-second duration. We included six motions regressors derived from realignment, along roll, pitch, yaw and x,y,z dimensions to control for motion artifact. We also included additional framewise displacement (FD) regressors, using a FD threshold = 1.2 to identify noisy images. We then regressed out noise from frames above the FD threshold, including the previous images and subsequent two images, based on recommendations (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012; Power et al., 2014).

Whole brain searchlight analysis. We conducted four independent searchlight analyses across the whole brain to measure representational dissimilarity during choice and feedback phases of the task, and separately for social and bandit tasks. For each participant and each task phase (choice and feedback), we first selected the relevant trial regressors (e.g., choice regressors for all trials in which the participant responded). We then constructed a 9mm radius spherical searchlight, which we moved along x,y,z coordinates of participant-specific brain masks (binary mask of voxels with sufficient accompanying BOLD activations), with a step size of 1, such that the center of the searchlight was placed in each voxel once.

We then extracted single event (choice/feedback) beta coefficients from all voxels within the searchlight from the relevant phase regressors modeled in our deconvolved time-series GLM and noise normalized the coefficients (Walther et al., 2016). Beta coefficients from each phase regressor were then extracted and reorganized into a voxel by trial coefficient matrix (Fig. 4A). To align the searchlight neural RDM with our state identity hypothesis matrix, we reorganized the coefficient matrix by nesting trial within each stimulus type. To obtain the searchlight RDM, we then computed the correlation distance (1-*r*) between each row and column element in the coefficient matrix. Correlation distance values from the lower triangle (all elements off the identity line) of our neural RDM were then z-transformed and correlated with the lower triangle of our z-transformed predictors, using the following linear model to obtain a *t*-statistic estimating the effect of state identity for each searchlight (neural correlation distance ~ state identity + autocorrelation term). We examined the effect of the state identity and Q-value hypothesis matrices separately in our whole brain searchlight analyses to avoid any systematic correlations that could potentially emerge from predictor collinearity but allowed state identity and Q-value predictors to compete in our ROI analyses. The resulting first-level *t*-maps from the whole brain searchlights for choice and feedback and for each task were then spatially smoothed with a 6mm^3 Gaussian smoothing kernel before being submitted to second-level analyses.

To avoid constructing task-biased ROIs, we created summed *t*-maps from each participant's social and bandit searchlights using SPM's imcalc function (images created using a simple summation method; social *t*-map + bandit *t*-map). We then conducted second-level analyses on the task-combined *t*-maps for choice and feedback phases of the task. To correct for multiple comparisons, we conducted non-parametric permutation testing on the second-level analyses, using a cluster-forming threshold of $p < .0001$ and a null distribution based on 5000 permutations. Permutation testing was conducted with the SnPM package (Hayasaka & Nichols, 2003). We then created binary masks of all voxels in our second-level analyses that were significant at the cluster-level ($p_{\text{FWE}} < .05$) and at the peak level ($p < .0001$) and used the task-combined corrected *t*-maps to identify ROIs. We used a data driven approach to identify two sets of ROIs for choice and feedback by extracting coefficients from statistically significant cluster peaks in corrected *t*-maps. We limited the cluster size of our ROIs to be no larger than the size our searchlight by placing a 9mm radius sphere at the center of the local maxima and extracted all voxels from the sphere that survived permutation testing. Within each ROI from our two sets chosen from choice and feedback searchlights, coefficients from the model were then disaggregated to independently evaluate the strength of state representation in the social vs. bandit task (shown in Fig. 5, B to C), which we computed from the following model (neural correlation distance ~ state-identity RDM + Q-value RDM + autocorrelation term).

Valence-based RSA. For each participant, trial-level deltas (PEs) estimated from the valenced-ca model using the MLE optimized parameters were z-scored to control for the extremity of experienced gains and losses across participants, to ensure we had sufficient trials from each stimulus within both positive and negative valence RDMs, and to ensure a roughly equivalent amount of data in each neural RDM. The z-transformed PEs were then used to separate data into the appropriate RDM and our RSA procedure for each ROI was then separately applied for positive and negative RDMs and for choice and feedback phases.

Cross-timepoint RSA. Prior to computing the cross-timepoint correlations across task phases, we constructed a set of ROIs in candidate areas that were involved in state representation during both choice and feedback. To avoid constructing ROIs that were biased towards a particular task phase, we integrated task-combined *t*-maps (choice *t*-map + feedback *t*-map). We then masked our task and phase-combined image using a binary conjunction *t*-map of voxels that survived permutation testing in *both* our choice and feedback phase analyses, using a 9mm radius sphere centered at the local peaks to isolate voxels from statistically significant clusters and focusing specifically on voxels in the PFC. We then computed the cross-timepoint correlation within each of our conjunction ROIs, separately for each task.

Our approach for computing the cross-timepoint correlation was to first create two data sets for each participant by separating the data into even and odd trials that occurred with each partner/bandit type (e.g., set 1: all even trials with low, high, neutral, random stimuli, set 2: all odd trials with low, high, neutral, random stimuli), so that representational correspondence could be measured across consecutive stimulus-matched trials that were temporally distanced in time (trial ordering was interleaved such that stimulus-matched trials were always 1-15 trials apart in the task; Fig. 1D). We then constructed a neural RDM for each of our even and odd data sets, including choice and feedback in same matrix, and then computed the correlation distance between our even and odd trial neural RDMs. The resulting matrix product of even and odd RDMs allowed us to evaluate the shared structure of representations across task phases (the lower left and upper right quadrants of the cross-timepoint RDM; Fig. 6A) and across independent trials, therefore breaking any systematic temporal or autocorrelation signals within the neural pattern. We then correlated only the cross-timepoint quadrants of the matrix product with a cross-timepoint state identity matrix to quantify the degree of representational alignment in the neural code across choice and feedback timepoints that preserved the state identity. The degree of shared information in the neural code (quantified as Pearson correlation coefficients) was subsequently submitted to further regression analyses to evaluate the association between shared representational geometry and credit assignment precision (Fig. 6C). We constructed ROIs in the mPFC and lOFC, given that clusters in these regions emerged from our conjunction *t*-map and signaled cross-timepoint state encoding.

Acknowledgments: We thank Michael J. Frank for comments and feedback and Eric Ingram for assisting with data collection. We also thank Avinash Vaidya for helpful RSA code and discussion. **Funding:** This work was supported by NARSAD grant 26210 to O.F.H. **Author contributions:** A.L., O.F.H, conceptualized the experiment, A.L., collected the data. Data Analysis: A.L., did the data analysis under the supervision of O.F.H, and M.N., A.L., O.F., M.N. wrote the manuscript. **Competing interests:** The authors declare no competing interests.

References and Notes

Akaishi, R., Kolling, N., Brown, J. W., & Rushworth, M. (2016). Neural mechanisms of credit assignment in a multicue environment. *Journal of Neuroscience, 36*(4), 1096-1112.

Asaad, W. F., Lauro, P. M., Perge, J. A., & Eskandar, E. N. (2017). Prefrontal neurons encode a solution to the credit-assignment problem. *Journal of Neuroscience, 37*(29), 6995-7007.

Boorman, E. D., Witkowski, P. P., Zhang, Y., & Park, S. A. (2021). The orbital frontal cortex, task structure, and inference. *Behavioral Neuroscience, 135*(2), 291.

Chau, B. K., Sallet, J., Papageorgiou, G. K., Noonan, M. P., Bell, A. H., Walton, M. E., & Rushworth, M. F. (2015). Contrasting roles for orbitofrontal cortex and amygdala in credit assignment and learning in macaques. *Neuron, 87*(5), 1106-1118.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience, 35*(7), 1024-1035.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol Rev, 120*(1), 190.

Davis, M., Walker, D. L., Miles, L., & Grillon, C. (2010). Phasic vs sustained fear in rats and humans: role of the extended amygdala in fear vs anxiety. *Neuropsychopharmacology, 35*(1), 105-135.

De Bundel, D., Zussy, C., Espallergues, J., Gerfen, C. R., Girault, J.-A., & Valjent, E. (2016). Dopamine D2 receptors gate generalization of conditioned threat responses through mTORC1 signaling in the extended amygdala. *Molecular psychiatry, 21*(11), 1545-1553.

Dunsmoor, J. E., Kroes, M. C., Braren, S. H., & Phelps, E. A. (2017). Threat intensity widens fear generalization gradients. *Behavioral Neuroscience, 131*(2), 168.

Dunsmoor, J. E., & Paz, R. (2015). Fear generalization and anxiety: behavioral and neural mechanisms. *Biological psychiatry, 78*(5), 336-343.

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences, 115*(7), E1690-E1697.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences, 5*, 43-50.

Hamid, A. A., Frank, M. J., & Moore, C. I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell, 184*(10), 2733-2749. e2716.

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *Neuroimage, 20*(4), 2343-2356.

Hull, C. L. (1943). Principles of behavior: An introduction to behavior theory.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., . . . Behrens, T. E. (2016). Reward-guided learning with and without causal attribution. *Neuron, 90*(1), 177-190.

Lamba, A., Frank, M. J., & FeldmanHall, O. (2020). Anxiety impedes adaptive social learning under uncertainty. *Psychological science, 31*(5), 592-603.

Lau, T., Gershman, S. J., & Cikara, M. (2020). Social structure learning in human anterior insula. *Elife, 9*, e53162.

Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: a meta-analysis. *Behaviour research and therapy, 43*(11), 1391-1424.

Mumford, J. A., & Poldrack, R. A. (2007). Modeling group fMRI data. *Soc Cogn Affect Neurosci, 2*(3), 251-257.

Nassar, M. R., McGuire, J. T., Ritz, H., & Kable, J. W. (2019). Dissociable forms of uncertainty-driven representational change across the human brain. *Journal of Neuroscience, 39*(9), 1688-1698.

Noonan, M. P., Chau, B. K., Rushworth, M. F., & Fellows, L. K. (2017). Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision-making in humans. *Journal of Neuroscience, 37*(29), 7023-7035.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation, 18*(2), 283-328.

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature neuroscience, 24*(9), 1292-1301.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage, 59*(3), 2142-2154.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage, 84*, 320-341.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *Neuroimage, 49*(2), 1545-1558.

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience, 30*(31), 10460-10464.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron, 91*(6), 1402-1412.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychol Rev, 121*(3), 526.

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behav Res Methods, 48*(3), 1197-1204.

Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*: University of Massachusetts Amherst.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285.

Vaidya, A. R., & Badre, D. (2022). Abstract task representations for inference and control. *Trends in Cognitive Sciences*.

Vaidya, A. R., & Fellows, L. K. (2016). Necessary contributions of human frontal lobe subregions to reward learning in a dynamic, multidimensional environment. *Journal of Neuroscience, 36*(38), 9843-9858.

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour, 6*(3), 404-414.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage, 137*, 188-200.

Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron, 65*(6), 927-939.

Witkowski, P. P., Park, S. A., & Boorman, E. D. (2022). Neural mechanisms of credit assignment for inferred relationships in a structured world. *Neuron*.

SUPPLEMENT

Prefrontal cortex state representations shape human credit assignment

Amrita Lamba, Matthew Nassar, Oriel FeldmanHall

Reinforcement Learning Model

Reinforcement Learning models and parameters. We built and tested a set of 6 nested RL models which provided differing cognitive accounts of learning and interference effects (see Methods). Each RL model in our final set included a subset of the parameters listed in Table S1, with their specific configurations noted in Table S2. To ensure adequate model identifiability for model comparison we verified that classification rates from the inverse confusion matrix were within a reasonable range (Fig. S1).

Model comparison. Model comparison was performed by minimizing Δ AIC across models (participant's best-fitting model AIC – comparison model AIC). Mean AIC and Δ AIC for each model are reported in Tables S3 to S4, respectively. As shown in Fig. S2 to S3, both the v-ca model (V-CA,V-LR) and the single ca-term model (V-LR,CA) better captured performance than the decay model, or at least equally as well in most instances. Further, the performance advantage of the V-CA model over the single term CA model was directly dependent on the degree of ca asymmetry between gains and losses (Fig. S3, D), allowing us to further explain valence-asymmetric credit assignment effects. As shown in the MLE predictive check in Fig. S4 and parameter recovery plots in Fig. S6, the v-ca model was able to capture task performance and parameters were reliably estimated. The distribution of parameter estimates for the v-ca model are provided in Fig. S5.

ROI coordinates

Regions of Interest. A full description of our ROI identification and construction process is detailed in the Methods section. MNI coordinates, estimated effect size, and cluster size for each ROI from the task-summed t-maps are reported for each phase in Tables S5 to S7. All peak coordinates were cross-referenced with Neurosynth and the cluster mass for each ROI was restricted to a 9mm-radius spherical searchlight size to prevent ROIs from spanning multiple anatomical boundaries. All reported effects below survived correction for multiple comparisons using our permutation testing procedure (Methods).

Model Parameter	Parameter Description	Upper-bound	Lower-bound
bias	Baseline investment tendencies.	2.0 – Baseline tendency to invest full amount.	0 – Baseline tendency to invest minimal amount.
m	Slope of logistic function associating trial Q-value and predicted investments.	2.0 – Little evidence needed to shift investment strategies.	0.2 – More evidence required before shifting strategies.
prior	Bias parameter on trial 1 before observing any outcomes.	2.0 – Invested fully on trial 1.	0 – Invested minimally on trial 1.
a	Learning rate quantifying degree of update to state Q-value from PE term.	1.0 – PEs maximally updated state Q-value.	0 – No influence of PEs on state Q-value.
a_{pos}	Learning rate for gains.	1.0 – PEs > 0 maximally updated state Q-value.	0 – PEs > 0 did not update state Q-value.
a_{neg}	Learning rate for losses.	1.0 – PEs < 0 maximally updated state Q-value.	0 – PEs < 0 did not update state Q-value.
ca	Credit assignment parameter capturing spread of credit across states.	1.0 – PEs only update Q-value of relevant state.	0 – PEs update the Q-value of all states.
ca_{pos}	Credit assignment for gains.	1.0 – PEs > 0 only update the Q-value of the relevant state.	0 – PEs > 0 update the Q-value of all states.
ca_{neg}	Credit assignment for losses.	1.0 – PEs < 0 only update the Q-value of the relevant state.	0 – PEs < 0 update the Q-value of all states.
decay	Degree of forgetting	1.0 – Complete decay of Q-values.	0 – No forgetting.

Table S1: *Logistic Q-learning algorithm parameters indicating model behavior at upper and lower bounds.*

Model No.	***Model***	***bias***	***m***	***prior***	a	a_{pos}	a_{neg}	***decay***	***ca***	ca_{pos}	ca_{neg}	***No. params***
1	Baseline	×	×		×							3
2	Decay	×	×	×	×			×				5
3	V-LR, Decay	×	×	×		×	×	×				6
4	V-LR, CA	×	×	×		×	×		×			6
5	V-CA	×	×	×	×					×	×	6
6	V-CA, V-LR	×	×	×		×	×			×	×	7

Table S2: *List of RL Models included in model comparison, including their respective free parameters indicated with the* ×*. V denotes valenced terms for either the learning rate (LR) or credit assignment (CA) parameters in the model.*

Model No	Model	Task	Mean AIC	SE
1	Baseline	Social	-324.93	30.84
		Bandit	-349.11	21.00
2	Decay	Social	-232.06	19.51
		Bandit	-285.99	17.75
3	V-LR, Decay	Social	-203.83	16.68
		Bandit	-277.41	18.55
4	V-LR, CA	Social	-202.68	17.00
		Bandit	-259.47	17.30
5	V-CA	Social	-208.54	16.75
		Bandit	-273.25	17.93
6	V-CA, V-LR	Social	-200.18	16.65
		Bandit	-258.25	16.99

Table S3: *Mean AIC and SE for each model. Mean AIC for V-CA, V-LR was the max in the set. Mean values are plotted below in Fig. S2.*

Model No	Model	Task	Δ AIC	SE
1	Baseline	Social	127.66	25.17
		Bandit	92.85	14.98
2	Decay	Social	34.79	9.35
		Bandit	29.73	5.28
3	V-LR, Decay	Social	6.56	4.77
		Bandit	21.16	5.23
4	V-LR, CA	Social	5.40	4.94
		Bandit	3.32	4.79
5	V-CA	Social	11.26	3.32
		Bandit	16.99	5.71
6	V-CA, V-LR	Social	2.91	4.88
		Bandit	2.00	4.56

***Table S4:** Model Comparison. Model Comparison was performed by minimizing Δ AIC, which was computed as the difference between each participants best-fitting model and each model in the set (see Methods). Δ AIC was the lowest for the V-CA, V-LR model, indicating that the model captured the behavioral data better than other models in the set, and in instances in which a participant's data was better fit by another model the V-CA, V-LR model could explain the data equally as well. Δ AIC with individual points is plotted below in Fig. S2.*

Region	***Peak MNI Coordinates***			Z_{peak}	***k***
	x	**y**	**z**		
Parietal R	46	-54	52	5.36	331
OFC M	-4	36	-24	4.02	81
dlPFC R	40	52	26	4.81	254
mPFC	2	54	-2	5.56	321
Parietal L	-58	-50	36	4.78	292
ACC	-2	10	42	5.14	291
OFC R	42	52	6	4.70	260
OFC L	-40	40	-4	4.71	253
dlPFC L	-42	8	40	4.36	267

Table S5. *Choice Phase ROI coordinates.*

Region	***Peak MNI Coordinates***			***Z_{peak}***	***k***
	x	**y**	**z**		
Parietal R	48	-38	54	6.45	388
mPFC	0	52	22	5.32	388
dlPFC R	42	28	28	5.50	389
Parietal L	-38	-38	54	5.16	377
OFC R	26	56	-10	4.72	238
dlPFC L	-40	20	32	5.57	327
Hippocampus R	40	-18	-10	4.89	213
OFC L	-48	46	-4	5.18	275
ACC	-2	24	32	4.75	330
OFC M	0	46	-20	4.33	142

Table S6: *Feedback Phase ROI coordinates.*

Region	***Peak MNI Coordinates***			***Z_{peak}***	***k***
	x	**y**	**z**		
OFC R	32	46	-10	5.26	106
mPFC	6	50	-2	5.94	158

Table S7. *Conjunction ROI coordinates.*

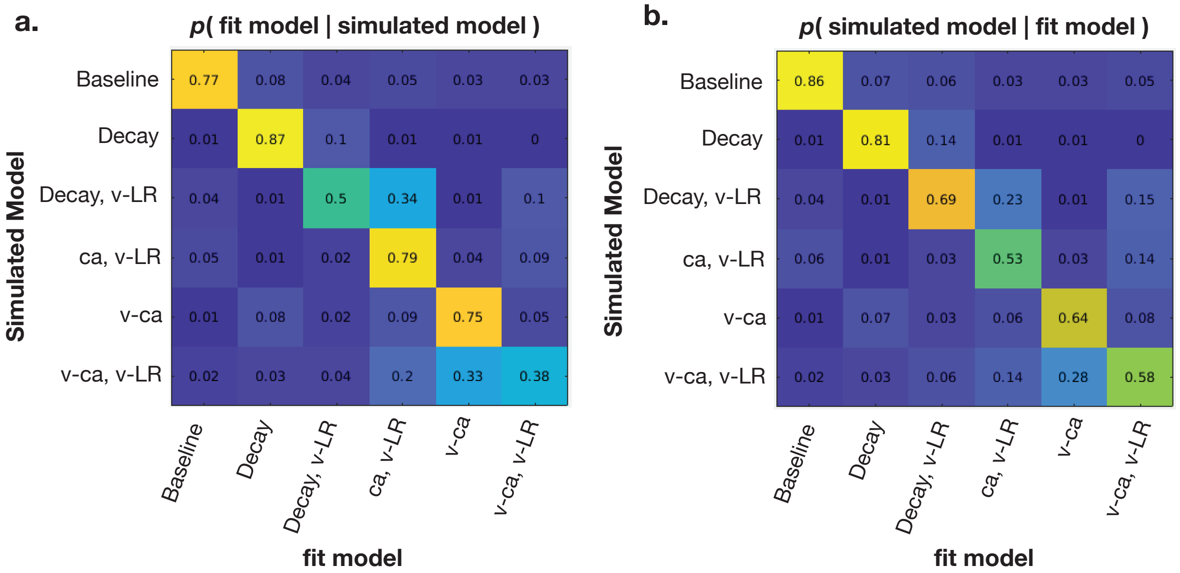


Fig. S1.** Model identifiability.* ***A: *Model confusion matrix. Model confusability was evaluated by simulating 100 participants per model. For each simulated participant, free parameters were randomly sampled from a uniform distribution (Table S2) and trial-to-trial investments were generated under the sampled parameterization. Each model was fit to each simulated participant using 20 iterations of gradient descent, and classification rates were computed as the frequency with which each participant was best fit by the correct generative model which we evaluated by maximizing the negative AIC (rate of winning model/true model).* ***B:*** *Inverse confusion matrix based on same data but showing the probability that each generative model gave rise to a given "best fitting" model.*

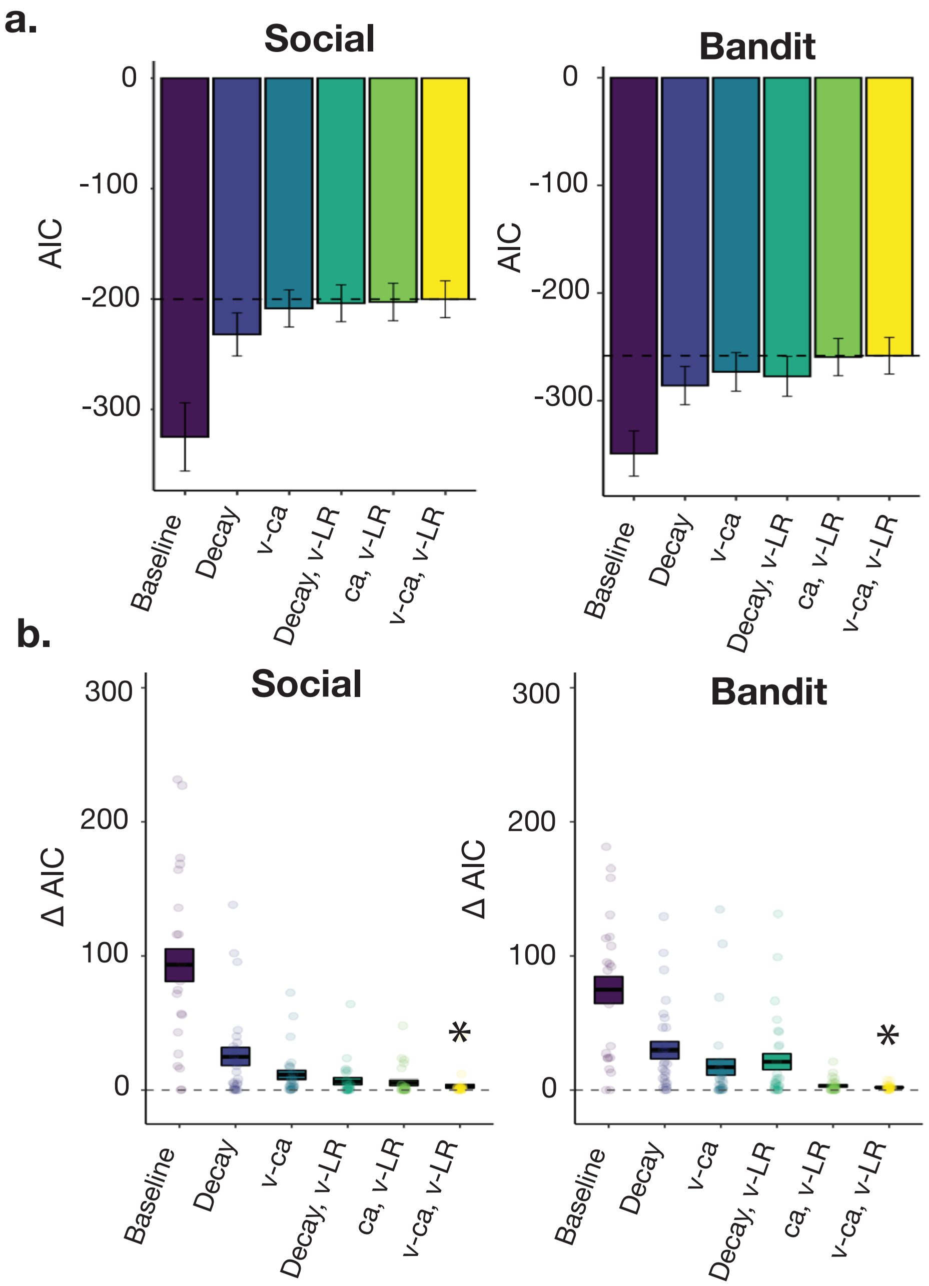


Fig. S2. *Model performance and comparison.* ***A:*** *Mean AIC for each model. Error bars correspond to the standard error of the mean. The dotted black line shows the mean AIC for the V-CA, V-LR model for comparison.* ***B:*** *Mean Δ AIC for each model. The length of each bar shows the standard error of the mean, and individual points correspond to the difference between the AIC for each participant's best-fitting model and the model denoted on the x-axis.*
** Denotes the model with the min Δ AIC in the set.*

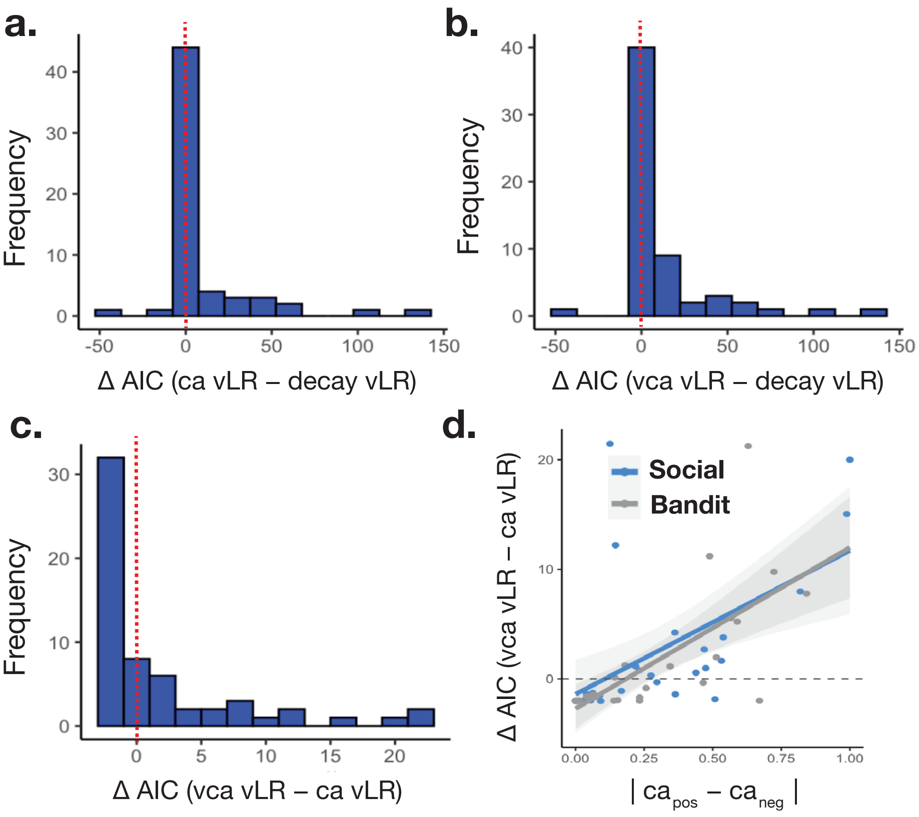


Fig. S3. *Δ AIC for models of interest.* ***A:*** *Model fit difference between the ca model and the decay model with V-LR terms. The histogram shows the model fit advantage of the ca model across participants, in which values above 0 indicate better fits to the ca model. Notably, almost all participants are better fit, or are equally fit, by the ca model.* ***B:*** *Model fit difference between the v-ca model and decay model. Again, almost all participants are better fit, or are equally fit, by the v-ca model.* ***C:*** *Model fit difference between the ca and v-ca models.* ***D:*** *The model fit advantage of the v-ca model compared to the single term ca model was associated with the degree of credit assignment asymmetry between gains and losses.*

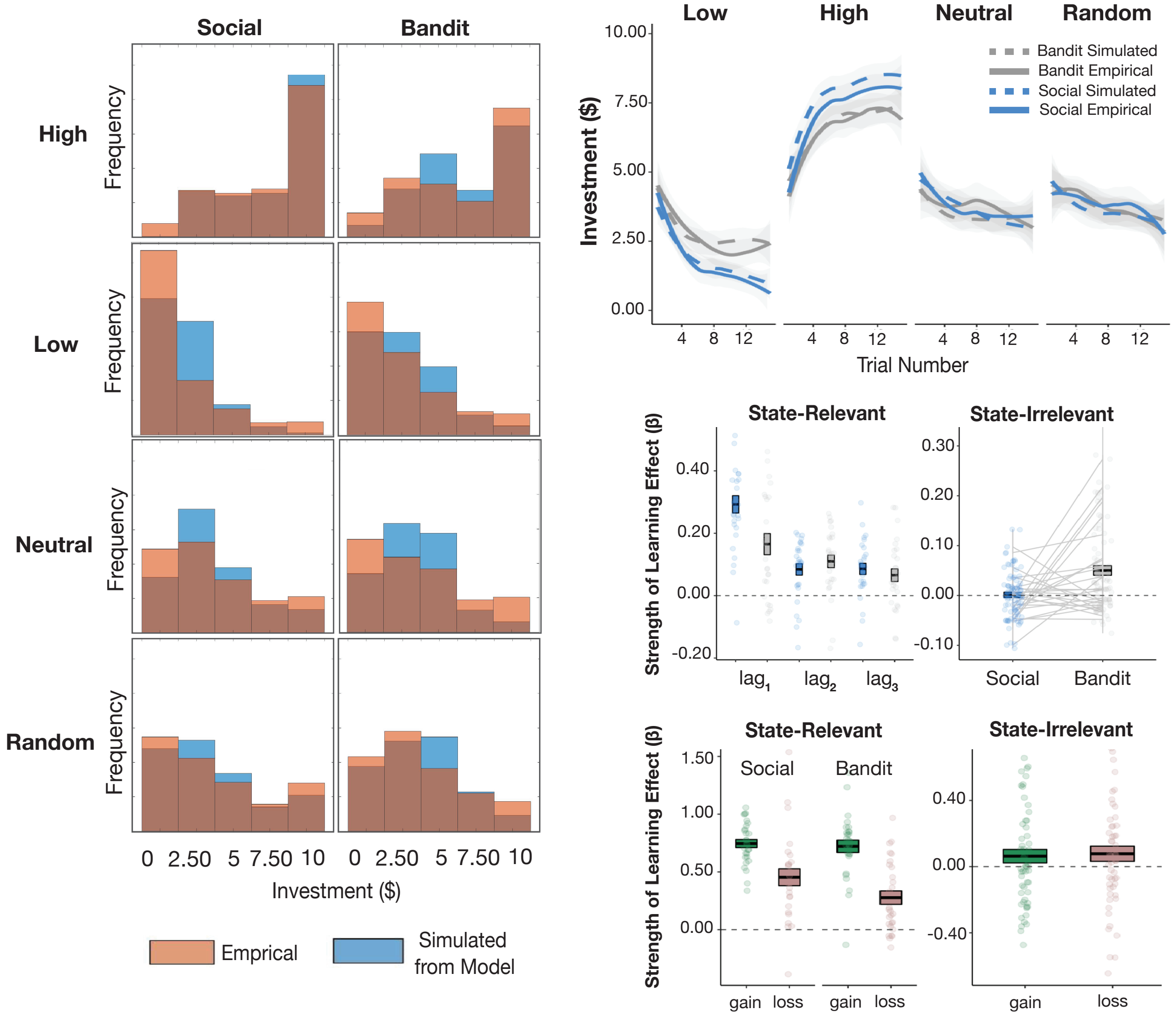


Fig. S4. *MLE Predictive Check. For each participant, the set of MLE-optimized parameters from the V-CA,V-LR model were used to simulate data from the model. Model-generated data is plotted above and reproduces behavioral effects.*

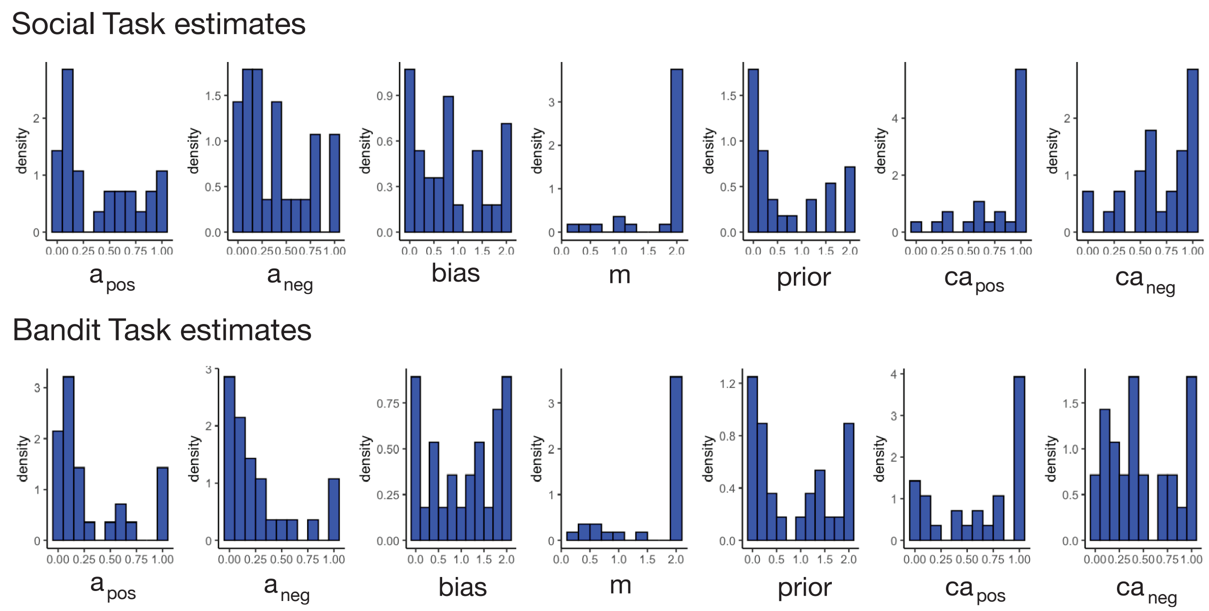


Fig. S5. *Distribution of parameter estimates from the v-ca model (V-CA, V-LR).*

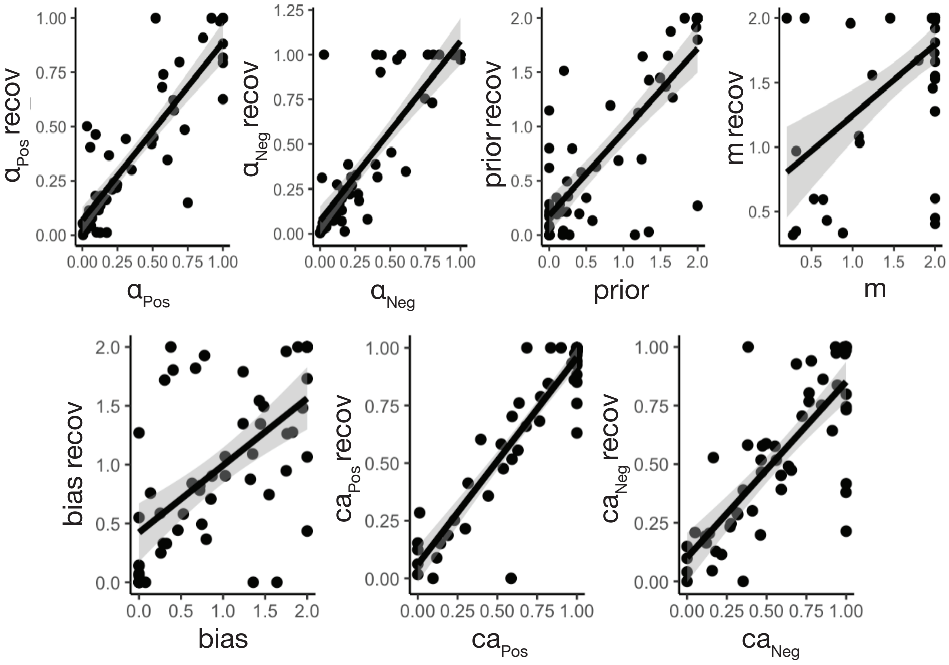


Fig. S6. *Parameter recovery. Estimated parameters from the v-ca model could be reliably recovered from model-generated data from the MLE predictive check, particularly for parameters of interest (ca and LR).*