Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rapid offline integration of learned experiences is shaped by valence and biases future decisions

Benjamin Becker

ben_becker@gmx.de

The University of Hong Kong https://orcid.org/0000-0002-9014-9671

Ting Xu

University of Electronic Science and Technology of China, Chengdu, China https://orcid.org/0000-0003-1590-9025

Zhiyi Chen

Third Military Medical University https://orcid.org/0000-0003-1744-4647

Lei Zhang

University of Birmingham https://orcid.org/0000-0002-9586-595X

Xianyang Gan

https://orcid.org/0000-0001-9960-8523

Feng Zhou

Ran Zhang

https://orcid.org/0009-0006-0099-4094

Weihua Zhao

University of Electronic Science and Technology of China https://orcid.org/0000-0003-1780-4606

Kun Fu

https://orcid.org/0009-0001-4475-3851

Lan Wang

https://orcid.org/0009-0001-1180-3939

Dezhong Yao

University of Electronic Science and Technology of China https://orcid.org/0000-0002-8042-879X

Tingyong Feng

Southwest University https://orcid.org/0000-0001-9278-6474

Article

Keywords: Reward, Loss, Offline integration, Decision-making, Functional magnetic resonance imaging

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10900455/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. L.Z. is an Editorial Board Member for Communications Biology, but was not involved in the editorial review of, nor the decision to publish this article. The other authors declare no competing interests.

Abstract

Rapid offline reactivation of recently learned experiences support adaptive behaviors in changing environments, but whether positive and negative experiences are differentially reactivated immediately after learning and how this shapes decisions remains unclear. Combining fMRI, multivariate predictive modeling and a probabilistic reward-learning task with learning and transfer stages, we examined how gain- and loss-related experiences are encoded during learning, reactivated during post-learning rest, and influence subsequent choices. We used computational modeling to dissociate gain- and loss-related learned experiences, spanning observable stimulus and outcome information as well as latent expected values and prediction errors. Gain- and loss-related experiences were encoded in dissociable neural expressions and were differentially reactivated during post-learning rest. While gain-related experiences exhibited stronger overall reactivation, loss-related reactivation specifically predicted accurate loss avoidance in the subsequent transfer. Moreover, coordinated engagement of a distributed memory network encompassing the posterior cingulate cortex, middle temporal gyrus and hippocampus increased across post-learning rest periods, consistent with systems-level consolidation supporting adaptive loss avoidance. Together, these findings provide evidence that rapid offline reactivation is valence-sensitive and functionally asymmetric, with negative experiences exerting a disproportionate influence on subsequent avoidance despite weaker reinstatement, a pattern that may have broader relevance for understanding persistent negative thoughts in affective disorders.

INTRODUCTION

Humans possess a remarkable ability to continue processing recent experiences even after they are no longer directly perceptible. After giving a presentation, for example, moments such as nailing a tough question, an awkward pause or an unexpected reaction from the audience may shape how we prepare for the next one, even without deliberate reflection. Such continued processing thus allow newly acquired information to be further integrated and guide subsequent adaptive behaviors. One candidate process supporting this rapid integration is the offline reactivation after learning. Once encoded into memory, recently acquired experiences can be spontaneously reinstated during periods of awake rest or sleep (1, 2). Such post-encoding reactivation has been well documented for spatial experiences in rodents, with recently traversed paths spontaneously reinstated during awake rest and sleep periods (3, 4), as well as for non-spatial experiences in humans, including recently encountered objects and episodic events (5, 6). However, it remains unclear how rapid offline reactivation differentially reinstates distinct components of newly learned experiences, whether this process is shaped by experience valence and whether such selective weighting influences future decisions.

Emerging evidence suggests that offline reactivation is not a passive replay of past events, but a rapid and selective process through which newly acquired information is differentially integrated after learning (1, 2). Within reinforcement learning (RL) contexts, animal studies utilizing spatial navigation tasks have revealed that representations of rewarded spatial locations are preferentially reactivated compared with unrewarded experiences (7–9). Initial observations in humans further substantiate this discovery by

demonstrating preferential reactivation of neural representations for highly rewarded objects (10), even though the objects are incidentally associated with rewarding feedbacks (11). However, reward is not the only form of motivationally salient information that may shape offline processing. Aversive experiences, such as encounters with danger, can also be preferentially replayed and support subsequent avoidance behavior in animals (12). These findings suggest that both rewarding and aversive experiences may modulate offline reactivation according to their relevance for adaptive approach and avoidance behaviors (13). However, most previous studies have examined those reward or punishment-related experiences in isolation, leaving unclear how experiences with opposing valence are differentially integrated when acquired within the same learning environment. Understanding how rewarding and aversive experiences are selectively reactivated immediately after learning may provide insight into why certain experiences exert disproportionate influences on future cognition and behaviors. For example, excessive weighting of negative experiences may contribute to maladaptive avoidance and persistent threat-related expectations in anxiety disorders (14), whereas excessive weighting of drug-associated experiences may enhance the motivational salience of drug cues and promote maladaptive approach behaviors in addiction (15, 16).

Offline reactivation was initially observed in studies showing that recently experienced spatial trajectories can be reinstated in the mammalian brain during subsequent periods of rest or sleep (4, 17–19). Early investigations selectively characterized this phenomenon as the reinstatement of hippocampal neural activity (20, 21), with hippocampal pyramidal neurons, known as place cells, exhibiting sequential activity patterns that resemble those observed during active exploration (22, 23). However hippocampal activity patterns during offline period can also incorporate non-spatial information, including sensory features such as sounds (24), as well as internal states including reward expectation (25–27) and prediction errors (PEs) (28, 29). Recent theoretical models further suggest that hippocampal reactivation can initiate cascading activation across interconnected brain systems, particularly the default mode network (30), facilitating the transformation and integration of newly acquired experiences. Consistent with this view, many nodes of DMN exhibit coordinated activation with brain-wide reactivation events, including the medial prefrontal cortex (mPFC), which can recapitulate recent behavior-related activity sequences through receiving hippocampal signals (31, 32), and parietal regions involved in the reactivation of modular motion memory representations (33). Moreover, reactivation of aversive outcome representations has been associated with theta activity in the medial temporal lobe (MTL), amygdala and mPFC, supporting aversive value updating (34), whereas increased MTL and anterior cingulate cortex (ACC) activity during offline re-evaluation of learned experiences played a key role in optimizing subsequent replanning behaviors to obtain rewards (29). However, despite evidence that hippocampal-cortical networks support the offline integration of learned experiences, it remains unclear whether experiences with different valence are differentially integrated across these networks and how such differences influence subsequent behaviors.

To address these questions, we combined multivariate predictive decoding of BOLD fMRI signals with computational modeling approaches in a probabilistic reward learning (PRL) paradigm incorporating brief post-learning rest periods. The PRL paradigm consisted of an initial learning phase, during which

participants acquired probabilistic stimulus–outcome associations involving monetary gain and loss feedbacks, followed by a transfer phase in which they generalized learned values to novel stimulus combinations to optimize reward-seeking and loss-avoidance choices (35, 36). This sequential design enabled us to examine how gain- and loss-related experiences acquired within the same learning environment are differentially integrated during offline periods and subsequently influence adaptive decision-making. Using computational and predictive modeling approaches, we quantified latent learning variables, including expected value and PEs, and derived multivariate neural representations of both observable (stimulus and outcome) and latent (expected value and PEs) components of learned experiences across gain and loss contexts. Multivariate neural classifiers trained separately on gain- and loss-related experiences were then applied to post-learning rest periods to characterize which components of these valenced experiences were spontaneously reactivated and whether such reactivation predicted subsequent transfer performance. This framework allowed us to investigate how the brain selectively integrates rewarding and aversive experiences immediately after learning and how their rapid offline reactivation contributes to future reward-seeking and loss-avoidance decisions. We hypothesized that: (1) distinct components of gain- and loss-related experiences would exhibit dissociable neural representations, enabling their differential offline reactivation; (2) offline reactivation would be modulated by valence of learned experience, reflecting the selective weighting of rewarding and aversive information; and (3) neural representations within regions implicated in offline memory integration and reward value evaluation (e.g., hippocampus, MTL, mPFC and ventral striatum) (37–39) will be observed during rest periods, and guide reward-seeking or loss-avoidance choices during transfer.

Results

Successful acquisition and transfer of probabilistic reward-loss contingencies

We employed a classical PRL task consisting of sequential learning and transfer stages, enabling participants to acquire probabilistic stimulus-outcome associations and apply those learned information to make optimal choices when the stimuli were recombined into novel pairs (Fig. 1a) (35, 36). During the learning stage, participants were repeatedly presented with three stimulus pairs (AB, CD and EF), in which each stimulus was associated with a distinct probability of gain or loss outcomes. To characterize trial-by-trial learning behavior, choices were coded as optimal (1) when participants selected the stimulus with a higher probability of gain outcomes within each pair (e.g., choosing A over B in the AB pair) and suboptimal (0) otherwise. A Bayesian linear multilevel (BLM) model was then applied to examine the effects of learning periods (early vs. late learning), stimulus pairs (AB, CD and EF), and gender on trial-by-trial choices. The results revealed significant effects of stimulus pairs ($\beta = 0.07$, 95% CI [0.04, 0.10]) and learning periods ($\beta = 0.07$, 95% CI [0.05, 0.09], Fig. 1b), whereas no significant effect of gender was observed ($\beta = -0.01$, 95% CI [− 0.06, 0.04], Fig. 1c). Specifically, participants increased the choice of stimuli with higher probability of gain outcomes across learning periods, indicating successful acquisition of probabilistic reward-loss contingencies.

We next examined whether participants could use previously learned reward–loss contingencies to guide choices in the transfer stage, where learned stimuli were recombined into novel pairs without feedback. This design allowed us to test whether participants could generalize prior learning by preferentially selecting the stimulus associated with the highest probability of gain (e.g., A) and avoiding the stimulus associated with the highest probability of loss (e.g., B). The choice accuracy for selecting A and avoiding B was comparable ($\beta = -0.03$, 95% CI [– 0.10, 0.04], Fig. 1d), but both were significantly above chance level (all > 50%, $p < 0.001$) across transfer trials. These findings indicate that participants successfully generalized learned reward-loss contingencies to novel choices, enabling adaptive approach toward rewards and avoidance of possible loss.

Computational modeling identifies latent variables underlying early probabilistic learning

While participants progressively improved their optimal choices from early to late learning period, behavioral accuracy alone does not reveal how values were updated across trials or how latent learning signals shaped subsequent decisions. To characterize the computational processes underlying choice behavior, we fitted RL models that quantified trial-wise latent variables, including expected values and PEs. Because all participants exhibited high choice accuracy during late learning (> 90%), which could reduce sensitivity for estimating individual learning parameters, subsequent computational and neural analyses focused on the early learning period (Fig. 1c). We thus compared four computational models (M1: Recorla-Wagner (RW) model with single learning rate; M2: RW model with dual learning rate; M3: Reinforcement Learning Drift Diffusion model (RL-DDM) with single learning rate; M4: RL-DDM with dual learning rate) to characterize how participants updated value representations during early learning period. The M2 was finally selected because of good model fitting performance in comparison with other three models (please see computational modeling details in **Supplemental Material**). Moreover the M2 also captures better representation of overall learning trends (the trial-wise observed data significantly correlates with simulated data, with high correlation coefficients that ranging from 0.94 to 0.99, all ps < 0.001, Fig. 2a). In consistent with research including dual learning rate (40–42), we observed an optimism bias in participants' choices such that α_{Gain} was higher than α_{Loss} on both group (Fig. 2b) and individual level ($p < 0.001$, Fig. 2b). The explore-exploit tendency (β) also showed a significantly positive correlation with α_{Gain} ($r = 0.63$, $p < 0.001$).

To assess the reliability of parameter estimation in M2, we conducted a parameter recovery analysis. For each participant, the parameters estimated from the original data were used as generating parameters to simulate 50 new datasets, and M2 was then refitted to each simulated dataset to obtain recovered parameter estimates. The recovered estimates were averaged across simulations and compared with the corresponding generating parameters. Recovered parameters showed strong correlations with the original estimates (α_{Gain}, $r = 0.87$, $p < 0.001$; β, $r = 0.83$, $p < 0.001$; correlation between α_{Gain} and β, $r = 0.54$, $p < 0.001$, Fig. 2c). Moreover, individual level α_{Gain} was higher than that of α_{Loss} (individual level, $p < 0.001$, Fig. 2c), indicating that model parameters of M2 were recoverable as desired. Critically, the trial-wise

expected values (Q value) and PEs estimated by M2 systematically tracked changes in reward contingencies (all ps < 0.001, Fig. 2d), supporting their validity as latent signals of ongoing value updating during probabilistic learning.

Distinct neural signatures of observable and latent learned experiences

Having identified expected value and PEs as latent variables that dynamically tracked probabilistic learning, we next examined whether different components of the learned experiences were represented by dissociable neural signatures during early learning. Recent efforts in applying machine-learning based multivariate predictive modeling approach to fMRI data (43) have enabled to characterize the accurate neural representations of internal mental processes such as subjective emotional experience (44, 45) and motivation (46, 47). To this end, we developed whole brain fMRI predictive patterns associated with both observable experience components (stimulus and outcome) and computational modeling-derived latent components (expected value and PEs). These learned experience-related neural signatures subsequently served as templates for testing their spontaneous reactivation during post-learning rest.

Specifically, we first identified multivariate neural patterns that discriminated gain- from loss-related stimuli and outcomes during early learning. Whole-brain linear support vector machine (SVM) classifiers reliably distinguished gain- versus loss-related stimuli (accuracy = 0.84, sensitivity = 0.83, specificity = 0.85, p < 0.001, Fig. 3a, **Fig. S1**) or outcomes (accuracy = 0.90, sensitivity = 0.91, specificity = 0.89, p < 0.001, Fig. 3b, **Fig. S1**). After bootstrap thresholding (5,000 samples) and false discovery rate (FDR) correction (p < 0.05), the stimulus-discriminative pattern encompassed distributed regions including the posterior cingulate cortex (PCC), dmPFC, ACC, insula and vmPFC (Fig. 3a), whereas the outcome-discriminative pattern was dispersed within regions encompassing the vmPFC, dmPFC, ACC, insula and putamen (Fig. 3b). These regions have been implicated in multiple components of reinforcement learning, including reward value anticipation (47–49), outcome evaluation (50, 51) and monitoring the violations of predicted reward outcomes (52, 53), as well as the aversive feelings towards unexpected loss (54, 55).

Next, we sorted out the continuous Q value and PEs with descending sequence and divided them into 5 levels for each participant to develop predictive neural patterns sensitive to track changes of Q value and PEs. Specifically, we fitted liner support vector regression (SVR) models to identify the whole brain fMRI patterns that predicted the levels of Q value and PEs with 10 x 10-fold cross validation. Applying thresholding (bootstrapped 5,000 samples) and multiple comparisons correction (FDR corrected, p < 0.05), the Q value predictive regions included right SFG and anterior insula (Fig. 3c), while the right VS and bilateral anterior insula (aINS) were PEs-predictive regions (Fig. 3d). Those neural patterns of Q value and PEs could accurately predict the true levels of Q value and PEs such that: (1) the average within-subject correlation between predicted and actual Q value and PEs (5 pairs of scalar values per subject) was r = 0.67 ± 0.02 (standard error (SE)) and r = 0.63 ± 0.01, separately; (2) the average root

mean squared error (RMSE) was 1.50 ± 0.03 and 1.47 ± 0.02, respectively; (3) and the overall (between and within-subjects) prediction-outcome (i.e., 46x5 = 230 pairs) correlation coefficient was 0.45 (averaged across ten repetitions; bootstrapped 95% CI, [0.41, 0.50]) and 0.43 (averaged across ten repetitions; bootstrapped 95% CI, [0.41, 0.46]). Two-alternative forced-choice tests further showed that these neural patterns reliably discriminated high from low and moderate from low levels of expected value and PE, with accuracies ranging from 0.85 to 0.91 and from 0.91 to 0.98, respectively.

Finally, we examined whether the neural signatures corresponding to different components of learned experience were spatially distinct. The whole-brain multivoxel pattern similarity analyses were employed to determine the spatial similarity between the SVM or SVR weight maps for stimulus, outcome, expected value and PE. Pairwise spatial pattern correlations were not significant, with their bootstrapped 95% CIs including zero, indicating largely dissociable neural representations across learned experience components (Fig. 3e). In addition, we explored spatial similarity between neural predictive maps of learned experiences and regions of interests documented in those maps. The dmPFC, PCC, vmPFC and bilateral aINS incorporated stable predictive voxels processing the experience about stimuli and outcome, while the VS was largely specific to outcome (Fig. 3f). Among multiple regions related to Q value and PEs, the SFG and right aINS were preferentially important for encoding Q value, whereas the VS and bilateral aINS encoded PEs more (Fig. 3f). Together, these findings demonstrate that different components of learned experience were represented by partially dissociable multivariate neural signatures, providing distinct neural templates for examining their subsequent offline reactivation.

Offline reactivation of learned experiences and its relation to subsequent transfer choices

To further characterize the offline reactivation of learned experiences, we applied classifiers trained on gain- and loss-related stimulus, outcome, PEs and expected value representations to each TR acquired during the rest periods via the use of L2-regularized logistic regression (penalty = 1) (29, 56). This analysis only focused on the last 10 TRs of eat rest period in order to reduce the carry-over effects from residual neural activity elicited in each transfer phase. For each TR, the classifier generated a probability estimate ranging from 0 to 1, reflecting the extent to which the activity pattern of rest period resembled the corresponding representation of learned experiences. Applying these classifiers to successive fMRI volumes during the rest periods therefore provided a dynamic estimate of the reactivation of gain- and loss-related stimulus and outcome information, as well as latent PE and expected value representations. We first examined whether reactivation changed across rest periods. Overall reactivation probabilities did not differ significantly across rest periods (all ps > 0.05, Fig. 4a). However, loss-related experiences exhibited more coherent reactivation profiles across rest periods, with correlations ranging from r = 0.74 to 0.95 (all ps < 0.001), whereas gain-related experiences did not show cross-period correspondence (r_{min} = 0.25, p = 0.18).

Considering that there were inconsistent findings regarding the dissociation of reactivation between positive and negative events (10, 11, 13), we compared the reactivation probabilities of gain- and loss-related experiences across rest periods. Gain-related experiences showed significantly greater overall reactivation than loss-related experiences ($t = 5.53$, $p < 0.001$, Fig. 4b). We further examined whether latent information, such as internally estimated expected values and PEs, was differentially reactivated relative to directly observable experience components (stimulus and outcome). Reactivation probabilities were therefore compared between observable components and latent components, revealing significantly greater reactivation of observable than latent experiences ($t = 3.38$, $p < 0.001$, Fig. 4c).

Finally, we asked whether offline reactivation during rest periods guided optimal choices in the corresponding round of transfer stage. If the reactivation contributes to carrying recently learned experiences forward into later optimal decisions, its strength should covary with participants' ability to approach rewarding options (selecting A)or avoid loss-related options (avoiding B). Across participants, greater reactivation of loss-related experiences during Rest 1 was associated with more accurate avoidance of stimulus B in the subsequent transfer stage ($r = 0.31$, two-tailed permutation $p < 0.05$, 5,000 permutations, Fig. 4d), whereas no significant association was observed between gain-related reactivation and subsequent selection of stimulus A ($r = -0.16$, $p = 0.28$, Fig. 4d). Overall, these results reveal an asymmetry between offline reactivation strength and its relationship with subsequent behavior: gain-related experiences showed stronger overall reactivation, whereas reactivation of loss-related experiences was more consistent across rest periods and selectively associated with more accurate subsequent loss-avoidance choices.

Brain systems engaging across post-learning rest

Given evidence that offline reactivation involves coordinated activity across hippocampal–cortical networks (30, 57), we next asked which brain systems showed progressively increasing engagement across post-learning rest. A whole-brain multivariate SVR analysis was used to identify a multivoxel pattern that reliably tracked progression across successive rest periods. The multivariate neural map revealed increased activation across a distributed network encompassing the PCC (15 voxels), fusiform (54 voxels) and MTG (131 voxels, $p < 0.05$, FDR corrected, Fig. 5a). A small hippocampal cluster was also detected at the same statistical threshold (x/y/z = 34/-14/-12, 4 voxels, $p < 0.05$, FDR corrected). In contrast, decreased activation was observed in the VS (11 voxels, $p < 0.05$, FDR corrected, Fig. 5a) and VMPFC ($p < 0.001$, uncorrected for exploratory purpose, Fig. 5a), consistent with the absence of reward evaluation processes. This multivariate neural map reliably tracked progression across post-learning rest, with a high average within-subject correlation ($r = 0.67 \pm 0.02$), high overall prediction-outcome correlation ($r = 0.53$, averaged across ten repetitions; bootstrapped 95% CI, [0.47, 0.58]) but low average RMSE (0.65 ± 0.02). A two-alternative forced-choice test was further applied and indicated that the neural map could distinct different rest periods with high accuracies (0.76–0.89, $p < 0.001$, Fig. 5a). Together, these findings identify a distributed hippocampal-cortical network whose activity progressively

increased across post-learning rest, consistent with its involvement in the reward re-evaluation (58) and memory integration (59).

We finally examined whether the activation patterns of brain regions consistently engaging in post-learning rest periods was associated with subsequent reward-seeking and loss-avoidance behaviors. Applying the multivoxel pattern map identified from post-learning rest periods to neural activity during transfer revealed greater pattern expression for loss-avoidance trials (avoiding B) than for reward-seeking trials (selecting A; $t = 2.30$, $p = 0.03$, **Figure .5a**). We further tested whether regional pattern expression within this network was related to individual differences in subsequent loss avoidance. Increased pattern expression in the MTG and PCC were positively associated with more accurate avoidance of stimulus B during the first transfer stage (MTG, $r = 0.32$, two tailed permutation test $p < 0.03$, PCC, $r = 0.31$, two tailed permutation test $p < 0.03$, 5000 permutated samples, Fig. 5b), whereas no significant associations were observed for the other regions (all permutation ps > 0.05). Therefore, while distributed brain systems showed systematic changes across post-learning rest, increasing engagement of the MTG and PCC was selectively associated with more accurate subsequent loss-avoidance behaviors.

Discussion

Human decision-making in uncertain environments relies not only on learning action-outcome contingencies but also on integrating past experiences to guide future behavior (60–62). Our findings reveal that rapid offline integration of learned experiences is both valence-sensitive and functionally asymmetric, with the strength, stability and behavioral relevance of reactivation varying across gain- and loss-related experiences. Behaviorally, participants successfully applied previously learned contingencies to novel stimulus combinations, supporting both reward-seeking and loss-avoidance choices, as evidenced by above-chance choice accuracy across transfer phases. To characterize the computational processes underlying this learning, we used RL models to dissociate directly observable components of experience (stimuli and outcomes) from latent variables (expected values and PEs) under gain and loss conditions. Neuroimaging analyses further showed that these different experience components were represented by partially dissociable multivariate neural signatures and exhibited distinct reactivation profiles during subsequent rest. Gain-related experiences showed stronger overall reactivation, whereas reactivation of loss-related experiences was more consistent across rest periods and selectively associated with subsequent loss-avoidance performance. Furthermore, a distributed network involving the PCC, MTG and a small hippocampal cluster increased engagements across post-learning rest, and pattern expression within the PCC and MTG was associated with more accurate subsequent loss avoidance.

While offline reactivation of past experiences is thought to support subsequent adaptive decision-making—facilitating both reward maximization and threat avoidance—the specific content prioritized during offline periods may critically influence future behavioral optimization (63). Converging evidence indicates that offline reactivation is selective: experiences associated with salient stimuli, such as

reward (64), threat or fear (65), as well as novel experiences are preferentially reinstated (66, 67). In line with these research, we observed that monetary gain-related experiences were reactivated more strongly than loss-associated experiences, whereas loss-related experiences were replayed in a more temporally stable and associated manner. This asymmetric pattern suggests that reactivation does not merely reflect passive reinstatement of prior events, but instead selectively processes experiences according to their motivational and behavioral relevance. Such prioritization aligns with the framework of "memory triage," which proposes that salient experiences are replayed earlier and more frequently, increasing their likelihood of consolidation, whereas experiences deemed less relevant receive fewer reactivations and gradually decay (2). However, salience may also partly overlap with computational constructs central to RL. Experiences that elicit large PEs are inherently unexpected and therefore highly informative for value updating, making them plausible candidates for prioritized replay. Indeed, a recent animal study has shown that Q-learning models in which replay is prioritized by RPE provide a strong account of observed learning dynamics (28). To further examine these possibilities, we applied a Q-learning framework to dissociate observable and latent computational variables, including expected value and PE under gain and loss contexts. These components were represented in partially dissociable cortical, limbic and striatal regions, and exhibited distinguishable reactivation patterns, providing a potential neural basis for the selective weighting of recent experiences during offline processing.

In addition, the reactivation of previously learned loss-related experiences significantly predicted subsequent avoidance of disadvantageous options during the transfer phase. This observation is consistent with a large body of works showing that episodic retrieval of prior threat-associated experiences can guide future decision-making, biasing animals or humans to avoid previously harmful contexts (12, 34). However, unlike those studies that examined reactivation only within a single aversive learning context, our study employed a task incorporating both reward and loss feedback. We observed reactivation of previously learned gain- and loss-related experiences simultaneously, but critically, only the offline integration of loss-related events contributed to subsequent online decisions to avoid potential loss. While replay content can be enriched for representations of both previously rewarded or threating trajectories to guide subsequent behaviors (29, 68), the relative contribution of reward- versus loss-related replay to behavioral policy updating has not been systematically dissociated. Although a recent study implemented a unified approach–avoidance conflict and found that participants were more likely to avoid when replay was relatively stronger for paths leading to reward, and more likely to approach when replay was stronger for paths leading to loss (69). It primarily focused on prospective replay—the forward simulation of possible future courses of action and their anticipated outcomes (i.e., prospective planning or "rollout" of future states) that can bias ongoing decisions (70). In contrast, what the present study investigates is most about the reflection of past learned experiences but not a prospective plan for subsequent choice. During the transfer phase, participants were presented with novel recombined stimulus pairs without any feedback, so they could not rely on forward simulation of future outcomes. Instead, their choices—to avoid less rewarding option (i.e., B) were guided by the offline reactivation of past learned loss-related experiences, as well as the potential reorganization of previously learned loss contingencies (e.g., recognizing that B was the worst option across all stimuli). Such offline

reorganizations based on the understanding of task structures may further enables successful generalization to a relatively novel environments (71).

During post-learning rest periods, activation in the PCC, fusiform and MTG—along with a very small hippocampal cluster—showed coordinated engagement, as elevated activity was consistently observed across rest periods. In contrast, the VMPFC and VS exhibited progressively decreased activation across successive rest periods. When initializing replay, the increase in PCC–MTG–hippocampal engagement may reflect a query from the hippocampus to a putative cognitive map, possibly serving to align experiences into an ordered structure. These findings also coupled with recent theoretical models that the replay of hippocampal index codes ignites brain-wide, attribute-rich representations of experience indirectly by cascading through a series of brain networks (30). However, spontaneous replay of previously acquired information could, in principle, compete with ongoing cognitive processes. This possibility may be particularly relevant during the early transfer blocks, when participants encounter novel stimulus combinations and may continue to form new memory representations despite the absence of feedback. A potential mechanism for resolving this concern may lies in the involvement of the DMN, in which the PCC, hippocampus and MTG constitute key nodes. The DMN has been proposed to encode an internal model of the world (72, 73) and may support the coordination of replay processes while minimizing interference with externally oriented cognition (74). Such function of DMN may help support further relatively stable loss-avoidance behavior observed from the first transfer block onward. Moreover it is intriguing to observe that value-evaluation regions including the VMPFC and VS showed relatively stronger activation during the first rest period, followed by a gradual decrease across subsequent rest periods. One possible explanation is that, prior to full exposure to the structure of the transfer phase, these motivational circuits remained engaged in representing reward expectations or preparatory internal states (75). As participants became familiar with the transfer structure and task demands stabilized, reliance on value-evaluation processes may have diminished, leading to reduced activation in these regions.

Here, we provide evidence that offline reactivation is both valence-sensitive and functionally asymmetric: gain- and loss-related experiences were represented by partially dissociable neural signatures and exhibited distinct reactivation dynamics across post-learning rest periods. Although gain-related experiences showed stronger and more pronounced reactivation, the reactivation of loss-related experiences remained temporally stable and robustly predicted avoidance behavior in relatively novel contexts. However, several limitations should be acknowledged. First, although our multivariate neural decoding approach identified brain regions that showed coordinated engagement during off-task rest, the temporal resolution of fMRI limits our ability to capture replay events at finer timescales. Second, while we demonstrate that offline reactivation of previously learned experiences supports subsequent decisions—partly consistent with animal models showing preferential reactivation of rewarded locations via place-cell firing sequences (57, 76)—the definition of replay in humans predominantly reflects a representational framework (e.g., reinstatement of item-specific neural patterns) rather than a neuronal-level mechanism as characterized in rodent electrophysiology. Thus, representational-level findings in humans may not directly map onto single-neuron dynamics, which highlight the need for intracranial

recordings to resolve this translational gap. Third, we examined offline reactivation primarily within motivational contexts, where reactivation appeared biased by reward and loss. However, prior studies have also linked replay to broader memory-related functions, such as creative problem-solving and the abstraction of episodic experiences (77, 78), suggesting that replay may be selectively influenced by cognitive demands in addition to motivational value. Finally, although prior work suggests that replay is dependent on specific experience, not passage of time (79, 80), we cannot rule out the possibility that the relatively brief duration of learning in the present study constrained the robustness or stability of reactivation. Future studies employing extended training paradigms would be valuable to determine whether longer engagement alters the magnitude, content or functional role of replay.

In summary, we revealed that reward-related learned experiences were more robustly reinstated, whereas reactivation of loss-related experiences was more closely associated with subsequent avoidance behavior, pointing to a functional asymmetry in how past experiences are rapidly prioritized offline. These findings extend existing RL models, which have traditionally emphasized online updating (81), by integrating offline processes into the architecture of adaptive choice. More broadly, the valence-dependent and functionally asymmetric characteristics of offline reactivation may help explain how past experiences differentially shape future behavior, and may offer insights into the pervasive negativity bias of everyday cognition and affective disorders.

Materials and Methods

Participants

Fifty healthy, right-handed human participants were recruited for this fMRI experiment from the University of Electronic Science and Technology of China (UESTC). All participants were free of any current or a history of psychiatric, neurological, or other medical disorders and with no current or regular use of psychotropic substances including nicotine, and no MRI contraindications. Four participants were excluded from the analysis due to the failure on the choice accuracy criteria (i.e., choosing 65% A in AB, 55% C in CD, 50% E in EF, in line with Frank et al., 2007) in the learning phase. Therefore remaining data of forty-six participants (23 participants per gender; female, age (mean ± SD) = 21.13 ± 2.30 years, male, age = 20.78 ± 1.98 years) were included in the behavioral and fMRI analysis. Participants all provided written informed consent and were compensated at a base rate of 60 RMB for the behavioral task during the fMRI scanning and 60 RMB per hour for fMRI sessions. This study was approved by the Research Ethics Committee of School of Life Science and Technology in the UESTC (code: 1061422101024711) and adhered to the latest revision of the Declaration of Helsinki.

Experimental paradigm

We utilized a validated and widely-used PRL paradigm (35, 36, 40) with sequential learning and transfer stages to examine whether participants could acquire probabilistic stimulus–outcome associations and apply this learned information to optimize choices when previously learned stimuli were recombined into novel pairs. A series of visual real-world objects from different categories (82) were selected to avoid

preferential processing of stimulus with similar type and were presented when participants underwent the PRL task. During the initial learning stage, participants received with one of three different pairs of six objects (butterfly-hat, carpet-camera, guitar-sunflower were randomly denoted as AB, CD and EF, respectively) on each trial in a randomized order and were instructed to learn to choose the more rewarding object of each stimulus pair based on feedback (Fig. 1a). The learning difficulty varied for the stimulus pairs in terms of reward contingency (80%:20%, 70%:30% or 60%:40%, for AB, CD and EF, respectively). A total of 120 task trials were presented during the learning stage, distributed across two fMRI runs with 60 trials each (30 trials per stimulus pair). The first and second runs corresponded to the early and late learning periods, respectively. Each task trial began with a fixation cross presented for a jittered interval ranging from 0.3s to 0.5s followed by the presentation of two objects displayed to the left and right side of the fixation cross (side was counterbalanced). Stimuli were presented until participants made a response or after 3.5s whereas the reacted choice was visually confirmed by highlighting the chosen object with a blue frame for 0.3s, followed by 0.15s feedback presentation ('gold coin' or 'grey coin' indicated the monetary gain and loss, separately) and a final fixation cross lasted from 2.7s to 3.3s randomly. Following the learning stage, participants had been informed to test acquired experience in three rounds of transfer stage, with each ahead of a 30s rest period to ensure the potential offline reactivation of learned experiences (Fig. 1a). Six objects from learning stage were reorganized to constitute fifteen stimulus pairs and each pair was displayed 2 times in each round of transfer stage (side was counterbalanced). Each transfer round started with a fixation cross lasting same duration as the time set in the learning stage and then each stimulus pair was randomly displayed until participants made a response or after 3.5s whereas a blue frame was surrounded the chosen object for 0.3s and no feedback was offered. Participants in this stage were told to choose the more rewarding object in each stimulus pair solely rely on what they had learned in the learning stage.

Quantification and Statistical analyses

The Statistical tests were performed to analyze trial-by-trial choice behavior via the use of BLM model from the R brms package (version 2.16.1). In the learning stage, we determined the effects of stimulus pair (AB pair = 1, CD pair = 0, EF pair = -1), learning periods (1st fMRI run/early learning = 1, 2nd fMRI run/late learning = 2) and gender (male = 1, female = 0) on accuracy in choosing the more rewarding object of a stimulus pair (correct = 1, incorrect = 0). Significant effects were identified when more than 95% of the posterior distribution was above or below zero. Given the object A and B were associated with the highest probabilities of monetary gain and loss during learning, respectively, selecting A over the other objects (C–F) and avoiding B during transfer reflected successful acquisition and subsequent generalization of the corresponding reward and loss contingencies (83, 84). Therefore in each round of transfer stage, we included trials containing either stimulus A or B, excluding trials in which A and B were paired together (i.e., AB or BA). We then fitted an analogous Bayesian linear multilevel (BLM) model with condition (Approach A = 1, Avoid B = − 1) as the independent variable and trial-wise choice accuracy as the dependent variable (correct = 1, incorrect = 0) to test whether participants successfully applied the reward and loss contingencies acquired during learning to support optimal choices in transfer stage. A correct response was defined as selecting A in Approach A trials or not choosing B in Avoid B trials.

Computational modeling analyses

Direct comparisons on observable variables like individual choice accuracy can enable to examine whether the hypothesized claim is significant, yet it can make researcher stumble across one potential pitfall when they try out several statistical analyses, such as p-hacking (85). In contrast, the computational model that can interpret broad qualitative features of the behavioral data with complex algorithms holds paramount promises for advancing scientific inference and has been advocated to use in the realm of psychological science (86, 87). Consistent with this recommendation, we developed four primary computational models (M1: RW model with single learning rate; M2: RW model with learning rate for positive and negative outcomes, respectively; M3: RL-DDM with single learning rate; M4: RL-DDM with learning rate for positive and negative outcomes, respectively, **Table. S1**) with hierarchical Bayesian estimation to assess which model can describe how participants encode and predict value information to facilitate rewarding choices in the learning stage accurately.

The RW model elucidates error-driven learning process in which the choices are updated by evaluating the discrepancy between actual outcome and expected outcome value (88), whereas the RLDDM quantifies the learning of subjective values towards options over time and how this skews specific decisions via inspecting both choices and response time (89). To abstain from repeated redundant procedures of model evaluation, simulation and parameter recovery and select one model reaching goodness-of-fit criterion directly, we firstly compared the two RW models according to Leave-One-Out Information Criterion (90) using the loo package in R and evaluated the two RLDDMs performance based on the Deviance Information Criterion using the hddm package in python. The model comparison results revealed that M2 and M4 with dual learning rates were the best-fitting models (**Tables. S2-S3, Fig. S2-S7**). However, M4 was excluded due to poor simulation performance in posterior predictive checks of choice accuracy (**Fig. S4**). Therefore, we retained M2 with dual learning rates and further evaluated its parameter estimation validity via the use of model evaluation, simulation and parameter recovery analyses; see **Supplementary Methods** for details.

fMRI data acquisition and pre-processing

Whole-brain fMRI data were acquired on a 3T GE Discovery MR system (General Electric Medical System, Milwaukee, WI, USA). The anatomical MRI image were acquired using a high-resolution T1-weighted sequence (Time repetition (TR) = 6ms; echo time = 2ms; flip angle = 9°; field of view = 256 × 256 mm; matrix size = 256 × 256; voxel size = 1 × 1 × 1 mm; number of slices, 156; slice thickness, 1 mm) to improve spatial normalization of the functional data and exclude subjects with apparent brain pathologies, while the functional images using blood oxygenation level-dependent contrast were obtained via a T2*-weighted echo planar imaging sequence (TR= 2000ms; echo time= 30ms; flip angle = 90°; field of view = 240 × 240 mm; voxel size = 3.75 × 3.75 × 3 mm; number of slices, 39; slice thickness, 3 mm; resolution = 64 × 64).

All fMRI data were preprocessed via the use of a standard workflow in SPM 12 (Statistical Parametric Mapping; http://www.fil.ion.ucl.ac.uk/spm/; Welcome Trust Centre for Neuroimaging). In short, the initial

5 volumes of each functional time series were disregarded to facilitate T1 equilibration, whereas the remaining images were underwent a series of preprocessing steps including: correction for acquisition time delay, realignment to rectify head motion, unwarping for magnetic field inhomogeneities correction, and alignment with the T1-weighted structural image. Subsequently, the images were normalized to the Montreal Neurological Institute standard space (interpolated to a voxel size of 2 × 2 × 2 mm) utilizing the segmentation parameters from the anatomical images and were then spatially smoothed using an isotropic Gaussian kernel with a full-width at half-maximum of 8 mm. The spatial smoothing enhances the alignment of interindividual functional neuroimaging data while preserving the sensitivity inherent in multivariate pattern analyses (91, 92).

First-level fMRI analyses

We modeled the pre-processed BOLD time series data using three general linear models (GLMs), each incorporating different regressors tailored for specific analytical objectives. The first GLM focused on the early learning stage and included regressors for stimulus-pair presentation and outcome presentation, separately modeled for gain- and loss-related events. Additional nuisance regressors modeled the highlighting period following each choice and head motion (three rotation parameters and three translation parameters). Contrast images derived from this GLM were used to develop whole-brain multivariate neural patterns distinguishing gain- from loss-related stimulus and outcome experiences. The second GLM focused on the post-learning rest periods. Separate regressors modeled the final 10 TRs of each rest period, while the initial 5 TRs were excluded to minimize potential spillover from task-evoked neural activity immediately preceding rest. Head-motion parameters were included as nuisance regressors. Contrast images derived from this GLM were used to characterize regional activity changes across rest periods and to identify a whole-brain multivariate pattern that systematically varied across post-learning rest. The third GLM focused on the transfer stage and included separate regressors for Approach A and Avoid B trials within each transfer round, together with head-motion parameters as nuisance regressors. The resulting contrast images were used to quantify the expression of the multivoxel pattern identified during post-learning rest and to compare its expression between reward-seeking (Approach A) and loss-avoidance (Avoid B) trials. All regressors were convolved with the canonical hemodynamic response function, and a high-pass filter with a cutoff period of 128s was applied to eliminate low-frequency drifts.

Multi-voxel pattern analyses

We used linear SVM algorithms implemented in the Spider toolbox (http://people.kyb.tuebingen.mpg.de/spider) to identify whole-brain multivariate patterns discriminating gain- and loss-related stimuli and outcomes. The first-level contrast images representing two conditions (gain stimuli, loss stimuli) served as essential features for distinguishing between gain and loss stimuli, whereas the contrast images denoting the gain and loss outcomes were incorporated into the pattern classifier for gain and loss outcomes. To establish the statistical significance of the voxel weights and facilitate the interpretation of the classifier outputs, bootstrap tests comprising 10,000 bootstrap samples with replacement were undertaken. These tests yielded p-values that were utilized to establish

thresholds for the classifier weights, with two-tailed, uncorrected p-values determined for each voxel based on the proportion of weights above and below zero. To evaluate the performance of SVM classifier, we employed a leave-one-subject-out cross-validation procedure with a within-participant forced-choice test, in which pattern expression (the dot product of the activation image and classifier weights) was compared between gain and loss conditions, with higher expression indicating the gain condition (stimuli or outcome).

Given that leveraging a computational model enables to get trial-wise Q value and PEs, we methodically arranged the continuous Q value and PEs in a descending sequence and segmented them into 5 levels per participant. The SVR algorithms using a linear kernel (C = 1) implemented in the Spider toolbox with individual beta maps (one per level of Q value or PE for each subject) as features were employed to develop whole brain predictive neural patterns sensitive to capture fluctuations of Q value and PEs, respectively. To rigorously evaluate the efficacy of the SVR algorithm, we implemented a 10 × 10-fold cross-validation procedure. Participants were randomly partitioned into 10 folds, with nine folds used for training and the remaining fold used for testing. This procedure was repeated until each fold had served once as the test set. The entire 10-fold cross-validation procedure was repeated 10 times using different random partitions, and prediction performance was averaged across repetitions to reduce dependence on a particular training-test split (44, 45).

With the aim of providing a comprehensive evaluation of the predictive accuracy inherent in the neural signatures of Q value and PEs, we determined overall (between- and within-subjects; 230 pairs in total) and within-subject (5 pairs per subject) Pearson correlations (r) between the cross-validated predictions and the actual Q value or PE levels to indicate the RMSE to illustrate overall prediction error (for similar model evaluation metrics, see citations (44, 45)). Furthermore, we computed the classification accuracies of the neural signatures of Q value and PEs across three delineated levels - high, moderate, and low (i.e., high versus low and moderate versus low). This analysis utilized receiver operating characteristic curves within a forced-choice classification framework, wherein neural signature responses were contrasted between two conditions assessed within the same participant (the higher was chosen as with high expected value and positive PE). Two-sided binomial tests were used to test whether the classification accuracies were higher than chance-level (50%).

To further investigate whether the diverse learning experiences display extensive neural dissociation, we conducted multi-voxel pattern similarity analysis via calculating the pattern correlations for each participant between the SVM or SVR weights for stimuli, outcome, Q value and PEs, separately, across whole brain voxels and reported 95% CIs based on bootstrap tests with 500 iterations. Finally, we explored spatial similarity between neural predictive maps of learned experiences and regions of interests documented in those maps to investigate which certain brain regions make stable contributions to the prediction of one specific learned experience across participants.

Offline reactivation probability estimation and its association with transfer performance

A plethora of works had employed multi-voxel pattern recognition methodology to test whether the neural activity of prior learned experience could be reemerged in the sleep or short awake periods, during which applying task-specific classifier to the data of rest periods can entail to assess the reoccurrence likelihood for each task event relevant brain state (62, 93). In coincide with this, a L2-regularized logistic regression (penalty = 1) was employed to train a classifier on TRs when participants process gain- and loss-related stimuli, outcome, expected value and PE during early learning (29, 56). The trained classifiers were then applied to each TR acquired during the post-learning rest periods. Analyses were restricted to the final 10 TRs of each rest period to minimize potential contamination from residual task-evoked activity. For each TR, the classifier generated a probability estimate ranging from 0 to 1, reflecting the extent to which the activity pattern of rest period resembled the corresponding representation of learned experiences. By applying the classifier to successive brain scans acquired during the rest periods, we were able to derive a dynamic estimate of the extent to which subjects replay gain and loss stimuli/outcome, as well as PE/expected value events. Considering the offline integration of past learned experience can help support optimal choices in a relatively uncertain contexts (29, 93), we further conducted permutation-based (5000 samples) correlation analyses to examine whether the reactivation probabilities of gain- and loss-related learned experiences can predict the choice accuracy for reward-seeking (Approach A) and loss-avoidance (Avoid B) conditions across participants. Resulting correlation was considered significant at $p < 0.05$.

Identification of stable neural patterns across post-learning rest periods

Whole-brain multivariate pattern analysis was further used to identify neural patterns that changed systematically across post-learning rest periods. Specifically, we used the SVR algorithms with a linear kernel (C = 1) implemented in the Spider toolbox and encompassed individual beta maps (one per level of rest period) as features to make prediction. The prediction performance was evaluated using repeated 10-fold cross-validation, following the same procedure described above. The entire cross-validation procedure was repeated 10 times with different data partitions, and prediction performance was averaged across repetitions to reduce potential bias associated with individual training–test splits. Model performance was quantified using overall and within-participant correlations between predicted and actual rest-period indices, together with RMSE. Forced-choice tests were additionally used to assess whether the resulting multivariate pattern reliably discriminated between different rest periods.

Furthermore, to determine brain regions that made stable contributions across post-learning rest periods, we performed bootstrap analyses using 5,000 samples drawn with replacement. The SVR analysis was repeated for each bootstrap sample, and voxel-wise z scores and two-tailed P values were derived from the resulting bootstrap distributions of predictive weights. Stable predictive voxels were identified using FDR correction at $P < 0.05$, with an initial whole-brain voxel-wise threshold of $P < 0.001$, uncorrected. Finally, to examine whether expression of the neural pattern identified during post-learning rest was related to subsequent transfer behavior, we applied the resulting multivoxel pattern to contrast images from the Approach A and Avoid B conditions. Pattern expression was compared between

Approach A and Avoid B conditions. We further examined whether regional pattern expression within the identified regions was associated with the choice accuracy for reward-seeking (Approach A) and loss-avoidance (Avoid B) conditions. The statistical significance of these correlations was assessed using two-tailed permutation tests with 5,000 permutations, with $P < 0.05$ considered significant.

Declarations

Competing Interest Statement

LZ. is an Editorial Board Member for Communications Biology, but was not involved in the editorial review of, nor the decision to publish this article. The other authors declare no competing interests.

Author Contributions

TX and BB originally conceptualize the study. TX, RZ, KF, LW and YG implemented the experiments and acquired data. TX analyzed the data with the important advice provided by LZ, FZ and YC. TX and BB wrote the original manuscript. LZ, HZ, YF, ZY and BB denoted to review and edit for the original manuscript. BB supervised whole study and provided important support.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 32250610208, 32271123, 32571253, 32500964), General Project of Humanities and Social Sciences Research of the Ministry of Education (25XJA190002), Natural Science Foundation of Chongqing (CSTB2025NSCQ-GPX0487), the Fundamental Research Funds for the Central Universities (SWU2609105, SWU2509744) and a start-up grant from The University of Hong Kong. Disclaimer: Any opinions, findings, conclusions or recommendations expressed in this publication do not reflect the views of the Government of the Hong Kong Special Administrative Region or the Innovation and Technology Commission. We truly thank all colleagues for the unconditional assistance during data acquisition and the invaluable discussions that help to improve the manuscript, and all volunteers who participated in this study.

Data and code availability

The raw individual-level MRI data will be available from the corresponding author upon reasonable request after publication, subject to participant consent. Data supporting the findings, including figure source data, and custom analysis scripts will be accessible via the Open Science Framework repository (https://osf.io/amvsf/) upon publication. Code for computational modeling is available at: https://github.com/mccoyb4/Parkinson_RL and https://github.com/gbiele/RLDDM, while the MATLAB code or toolboxes used in the analyses are available at https://github.com/canlab.

References

1. Kurth-Nelson, Z., et al.: Replay and compositional computation. Neuron. **111**, 454–469 (2023)
2. van Der Meer, M.A., Bendor, D.: Awake replay: off the clock but on the job. Trends Neurosci. **48**, 257–267 (2025)
3. Ormond, J., Serka, S.A., Johansen, J.P.: Enhanced reactivation of remapping place cells during aversive learning. J. Neurosci. **43**, 2153–2167 (2023)
4. Ólafsdóttir, H.F., Bush, D., Barry, C.: The role of hippocampal replay in memory and planning. Curr. Biol. **28**, R37–R50 (2018)
5. Wimmer, G.E., Liu, Y., McNamee, D.C., Dolan, R.J.: Distinct replay signatures for prospective decision-making and memory preservation. *Proceedings of the National Academy of Sciences* 120, e2205211120 (2023)
6. Staresina, B.P., Alink, A., Kriegeskorte, N., Henson, R.N.: Awake reactivation predicts memory in humans. *Proceedings of the National Academy of Sciences* 110, 21159–21164 (2013)
7. Dupret, D., O'neill, J., Pleydell-Bouverie, B., Csicsvari, J.: The reorganization and reactivation of hippocampal maps predict spatial memory performance. Nat. Neurosci. **13**, 995–1002 (2010)
8. Ambrose, R.E., Pfeiffer, B.E., Foster, D.J.: Reverse replay of hippocampal place cells is uniquely modulated by changing reward. Neuron. **91**, 1124–1136 (2016)
9. Michon, F., Sun, J.-J., Kim, C.Y., Ciliberti, D., Kloosterman, F.: Post-learning hippocampal replay selectively reinforces spatial memory for highly rewarded locations. Curr. Biol. **29**, 1436–1444 (2019)
10. Gruber, M.J., Ritchey, M., Wang, S.-F., Doss, M.K., Ranganath, C.: Post-learning hippocampal dynamics promote preferential retention of rewarding events. Neuron. **89**, 1110–1120 (2016)
11. Wimmer, G.E., Büchel, C.: Reactivation of reward-related patterns from single past episodes supports memory-based decision making. J. Neurosci. **36**, 2868–2880 (2016)
12. Wu, C.-T., Haggerty, D., Kemere, C., Ji, D.: Hippocampal awake replay in fear memory retrieval. Nat. Neurosci. **20**, 571–580 (2017)
13. Bryzgalov, D.: Hippocampal reactivations after aversive or rewarding experience: classical and deep learning approaches, Université Paris sciences et lettres. (2021)
14. Klemanski, D.H., Curtiss, J., McLaughlin, K.A., Nolen-Hoeksema, S.: Emotion regulation and the transdiagnostic role of repetitive negative thinking in adolescents with social anxiety and depression. Cogn. therapy Res. **41**, 206–219 (2017)
15. Bornstein, A.M., Pickard, H.: Chasing the first high: memory sampling in drug choice. Neuropsychopharmacology. **45**, 907–915 (2020)
16. Treanor, M., Brown, L.A., Rissman, J., Craske, M.G.: Can memories of traumatic experiences or addiction be erased or modified? A critical review of research on the disruption of memory reconsolidation and its applications. Perspect. Psychol. Sci. **12**, 290–305 (2017)

17. Joo, H.R., Frank, L.M.: The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. Nat. Rev. Neurosci. **19**, 744–757 (2018)
18. Foster, D.J.: Replay comes of age. Annu. Rev. Neurosci. **40**, 581–602 (2017)
19. Rothschild, G.: The transformation of multi-sensory experiences into memories during sleep. Neurobiol. Learn. Mem. **160**, 58–66 (2019)
20. Pfeiffer, B.E.: The content of hippocampal replay. Hippocampus. **30**, 6–18 (2020)
21. Wilson, M.A., McNaughton, B.L.: Reactivation of hippocampal ensemble memories during sleep. Science. **265**, 676–679 (1994)
22. Skaggs, W.E., McNaughton, B.L.: Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. Science. **271**, 1870–1873 (1996)
23. Kudrimoti, H.S., Barnes, C.A., McNaughton, B.L.: Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics. J. Neurosci. **19**, 4090–4101 (1999)
24. Aronov, D., Nevers, R., Tank, D.W.: Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. Nature. **543**, 719–722 (2017)
25. Ólafsdóttir, H.F., Barry, C., Saleem, A.B., Hassabis, D., Spiers, H.J.: Hippocampal place cells construct reward related sequences through unexplored space. Elife. **4**, e06063 (2015)
26. Wikenheiser, A.M., Redish, A.D.: Changes in reward contingency modulate the trial-to-trial variability of hippocampal place cells. J. Neurophysiol. **106**, 589–598 (2011)
27. Gauthier, J.L., Tank, D.W.: A dedicated population for reward coding in the hippocampus. Neuron. **99**, 179–193 (2018)
28. Roscow, E.L., Howe, T., Lepora, N.F., Jones, M.W.: Post-learning replay of hippocampal-striatal activity is biased by reward-prediction signals. Nat. Commun. **16**, 10394 (2025)
29. Momennejad, I., Otto, A.R., Daw, N.D., Norman, K.A.: Offline replay supports Plann. Hum. reinforcement Learn. elife. **7**, e32548 (2018)
30. Kaefer, K., Stella, F., McNaughton, B.L., Battaglia, F.P.: Replay, the default mode network and the cascaded memory systems model. Nat. Rev. Neurosci. **23**, 628–640 (2022)
31. Euston, D.R., Tatsuno, M., McNaughton, B.L.: Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *science* 318, 1147–1150 (2007)
32. Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., Battaglia, F.P.: Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat. Neurosci. **12**, 919–926 (2009)
33. Wilber, A.A., Skelin, I., Wu, W., McNaughton, B.L.: Laminar organization of encoding and memory reactivation in the parietal cortex. *Neuron* 95, 1406–1419. e5 (2017)
34. Wise, T., Liu, Y., Chowdhury, F., Dolan, R.J.: Model-based aversive learning in humans is supported by preferential task state reactivation. Sci. Adv. **7**, eabf9616 (2021)
35. Frank, M.J., Claus, E.D.: Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol. Rev. **113**, 300 (2006)

36. Frank, M.J., Seeberger, L.C.: O'reilly, By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. **306**, 1940–1943 (2004)
37. DuBrow, S., Sherman, B.E., Meager, M.R., Davachi, L.: Medial temporal lobe damage impairs temporal integration in episodic memory. J. Cogn. Neurosci. **36**, 2302–2316 (2024)
38. Grabenhorst, F., Rolls, E.T.: Value, pleasure and choice in the ventral prefrontal cortex. Trends Cogn. Sci. **15**, 56–67 (2011)
39. Daniel, R., Pollmann, S.: A universal role of the ventral striatum in reward-based learning: evidence from human studies. Neurobiol. Learn. Mem. **114**, 90–100 (2014)
40. Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T., Hutchison, K.E.: Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. U.S.A. **104**, 16311–16316 (2007)
41. Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., Palminteri, S.: Behavioural and neural characterization of optimistic reinforcement learning. Nat. Hum. Behav. **1**, 0067 (2017)
42. Van Slooten, J.C., Jahfari, S., Knapen, T., Theeuwes, J.: How pupil responses track value-based decision-making during and after reinforcement learning. PLoS Comput. Biol. **14**, e1006632 (2018)
43. Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. **10**, 424–430 (2006)
44. Gan, X., et al.: A neurofunctional signature of subjective disgust generalizes to oral distaste and socio-moral contexts. Nat. Hum. Behav. 1–20 (2024)
45. Zhou, F., et al.: A distributed fMRI-based signature for the subjective experience of fear. Nat. Commun. **12**, 6643 (2021)
46. Xu, T., et al.: Angiotensin blockade enhances motivational reward learning via enhancing striatal prediction error signaling and frontostriatal communication. Mol. Psychiatry. **28**, 1692–1702 (2023)
47. Speer, S.P., et al.: A multivariate brain signature for reward. NeuroImage. **271**, 119990 (2023)
48. Iigaya, K., et al.: The value of what's to come: Neural mechanisms coupling prediction error and the utility of anticipation. Sci. Adv. **6**, eaba3828 (2020)
49. Murray, E.A., Rudebeck, P.H.: Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. Nat. Rev. Neurosci. **19**, 404–417 (2018)
50. van den Bosch, R., et al.: Striatal dopamine dissociates methylphenidate effects on value-based versus surprise-based reversal learning. Nat. Commun. **13**, 4962 (2022)
51. Filimon, F., Nelson, J.D., Sejnowski, T.J., Sereno, M.I., Cottrell, G.W.: The ventral striatum dissociates information expectation, reward anticipation, and reward receipt. *Proceedings of the National Academy of Sciences* 117, 15200–15208 (2020)
52. Hoy, C.W., et al.: Asymmetric coding of reward prediction errors in human insula and dorsomedial prefrontal cortex. Nat. Commun. **14**, 8520 (2023)
53. Domenech, P., Rheims, S., Koechlin, E.: Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. Science. **369**, eabb0184 (2020)

54. Horing, B., Büchel, C.: The human insula processes both modality-independent and pain-selective learning signals. PLoS Biol. **20**, e3001540 (2022)
55. Palminteri, S., et al.: Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. Neuron. **76**, 998–1009 (2012)
56. Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A.: Category-specific cortical activity precedes retrieval during memory search. Science. **310**, 1963–1966 (2005)
57. Harvey, R.E., Robinson, H.L., Liu, C., Oliva, A., Fernandez-Ruiz, A.: Hippocampo-cortical circuits for selective memory encoding, routing, and replay. Neuron. **111**, 2076–2090 (2023). e9
58. McGuire, J.T., Kable, J.W.: Medial prefrontal cortical activity reflects dynamic re-evaluation during voluntary persistence. Nat. Neurosci. **18**, 760–766 (2015)
59. Vaz, A.P., Wittig, J.H. Jr., Inati, S.K., Zaghloul, K.A.: Backbone spiking sequence as a basis for preplay, replay, and default states in human cortex. Nat. Commun. **14**, 4723 (2023)
60. Jacobacci, F., et al.: Rapid hippocampal plasticity supports motor sequence learning. *Proceedings of the National Academy of Sciences* 117, 23898–23903 (2020)
61. Guidotti, R., Del Gratta, C., Baldassarre, A., Romani, G.L., Corbetta, M.: Visual learning induces changes in resting-state fMRI multivariate pattern of information. J. Neurosci. **35**, 9786–9798 (2015)
62. Schuck, N.W., Niv, Y.: Sequential replay of nonspatial task states in the human hippocampus. Science. **364**, eaaw5181 (2019)
63. Mattar, M.G., Daw, N.D.: Prioritized memory access explains planning and hippocampal replay. Nat. Neurosci. **21**, 1609–1617 (2018)
64. Igata, H., Ikegaya, Y., Sasaki, T.: Prioritized experience replays on a hippocampal predictive map for learning. *Proceedings of the National Academy of Sciences* 118, e2011266118 (2021)
65. Girardeau, G., Inema, I., Buzsáki, G.: Reactivations of emotional memory in the hippocampus–amygdala system during sleep. Nat. Neurosci. **20**, 1634–1642 (2017)
66. Huelin Gorriz, M., Takigawa, M., Bendor, D.: The role of experience in prioritizing hippocampal replay. Nat. Commun. **14**, 8157 (2023)
67. Eldar, E., Lièvre, G., Dayan, P., Dolan, R.J.: The roles of online and offline replay in planning. *elife* 9, e56911 (2020)
68. Shin, J.D., Tang, W., Jadhav, S.P.: Dynamics of awake hippocampal-prefrontal replay for spatial learning and memory-guided decision making. Neuron. **104**, 1110–1125 (2019). e7
69. McFadyen, J., Liu, Y., Dolan, R.J.: Differential replay of reward and punishment paths predicts approach and avoidance. Nat. Neurosci. **26**, 627–637 (2023)
70. Zhang, J., Ou, J., Liu, Y.: Replay and ripples in humans. Annu. Rev. Neurosci. **48**, 65–84 (2025)
71. Liu, Y., Dolan, R.J., Kurth-Nelson, Z., Behrens, T.E.: Human replay spontaneously reorganizes experience. Cell. **178**, 640–652 (2019)

72. Liu, Y., Nour, M.M., Schuck, N.W., Behrens, T.E., Dolan, R.J.: Decoding cognition from spontaneous neural activity. Nat. Rev. Neurosci. **23**, 204–214 (2022)
73. Behrens, T.E., et al.: What is a cognitive map? Organizing knowledge for flexible behavior. Neuron. **100**, 490–509 (2018)
74. Higgins, C., et al.: Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. Neuron. **109**, 882–893 (2021)
75. Lansink, C.S., et al.: Preferential reactivation of motivationally relevant information in the ventral striatum. J. Neurosci. **28**, 6372–6382 (2008)
76. Lee, A.K., Wilson, M.A.: Memory of sequential experience in the hippocampus during slow wave sleep. Neuron. **36**, 1183–1194 (2002)
77. Lewis, P.A., Knoblich, G., Poe, G.: How memory replay in sleep boosts creative problem-solving. Trends Cogn. Sci. **22**, 491–503 (2018)
78. McDevitt, E.A., Zhang, J., MacKenzie, K.J., Fiser, J., Mednick, S.C.: The effect of interference, offline sleep, and wake on spatial statistical learning. Neurobiol. Learn. Mem. **193**, 107650 (2022)
79. Berners-Lee, A., et al.: Hippocampal replays appear after a single experience and incorporate greater detail with more experience. Neuron. **110**, 1829–1842 (2022)
80. McKenzie, S.: Once is enough for hippocampal replay. Neuron. **110**, 1749–1751 (2022)
81. Roscow, E.L., Chua, R., Costa, R.P., Jones, M.W., Lepora, N.: Learning offline: memory replay in biological and artificial reinforcement learning. Trends Neurosci. **44**, 808–821 (2021)
82. Konkle, T., Brady, T.F., Alvarez, G.A., Oliva, A.: Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. J. Exp. Psychol. Gen. **139**, 558 (2010)
83. Chase, H., et al.: Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. Psychol. Med. **40**, 433–440 (2010)
84. Waltz, J.A., Frank, M.J., Robinson, B.M., Gold, J.M.: Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. Biol. Psychiatry. **62**, 756–764 (2007)
85. Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D.: The extent and consequences of p-hacking in science. PLoS Biol. **13**, e1002106 (2015)
86. Guest, O., Martin, A.E.: How computational modeling can force theory building in psychological science. Perspect. Psychol. Sci. **16**, 789–802 (2021)
87. Collins, A.G.E., Cockburn, J.: Beyond dichotomies in reinforcement learning. Nat. Rev. Neurosci. **21**, 576–586 (2020)
88. Yau, J.O.-Y., McNally, G.P.: The Rescorla-Wagner model, prediction error, and fear learning. Neurobiol. Learn. Mem. **203**, 107799 (2023)
89. Fontanesi, L., Gluth, S., Spektor, M.S., Rieskamp, J.: A reinforcement learning diffusion decision model for value-based decisions. Psychon. Bull. Rev. **26**, 1099–1121 (2019)

90. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. **27**, 1413–1432 (2017)
91. de Beeck, H.P.O.: Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage* 49, 1943–1948 (2010)
92. Shmuel, A., Chaimow, D., Raddatz, G., Ugurbil, K., Yacoub, E.: Mechanisms underlying decoding at 7 T: ocular dominance columns, broad structures, and macroscopic blood vessels in V1 convey information on the stimulated eye. Neuroimage. **49**, 1957–1964 (2010)
93. Sterpenich, V., et al.: Reward biases spontaneous neural reactivation during sleep. Nat. Commun. **12**, 4162 (2021)

Figures

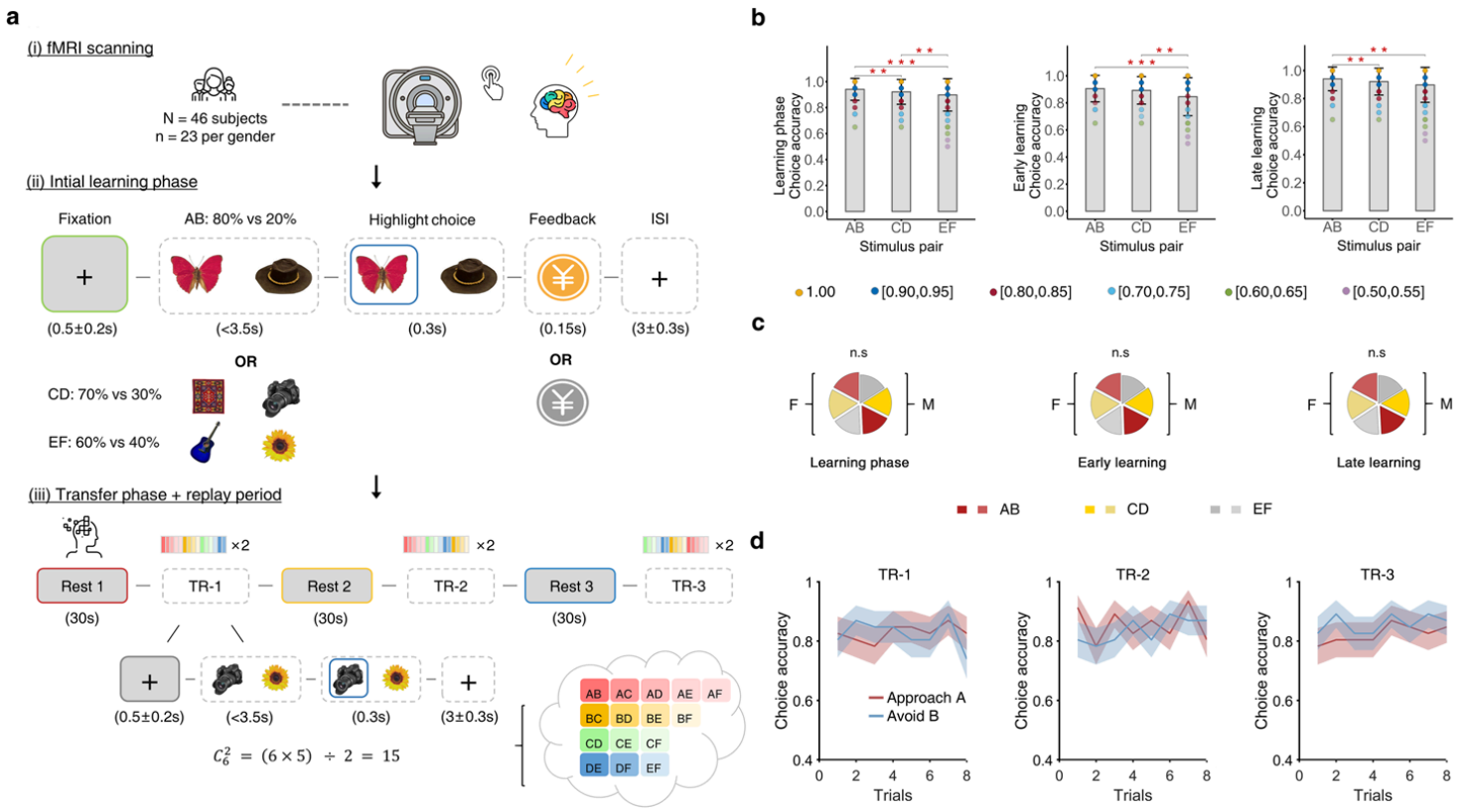

Figure 1

Experimental timeline, task design and initial behavioral results (a) During fMRI acquisition, participants completed a probabilistic reinforcement learning task comprising sequential learning and transfer stages. During the learning stage, one of three stimulus pairs (AB, CD, or EF) was presented on each trial in randomized order. Participants learned to select the better option within each pair based on probabilistic feedback (i.e., “Gold coin” or “Gray coin” images indicating monetary gain or loss, respectively). Stimulus identities and pair assignments were counterbalanced across participants to minimize stimulus-specific choice preferences. Each transfer stage was preceded by a 30-s rest period during which no explicit instructions were given. Participants were then presented twice with all possible

combinations of the previously learned stimuli in randomized order and instructed to select the better option based on their prior learned experience. (b) Choice accuracy was calculated as the proportion of trials in which participants selected the option associated with a higher probability of monetary gain within each stimulus pair. Choice accuracy differed across stimulus pairs and improved from early to late learning periods. (c) Choice accuracy across learning periods did not differ significantly between female and male participants. F, female; M, male; n.s., not significant. (d) During transfer, choice accuracy for selecting stimulus A, which was associated with the highest probability of monetary gain, and avoiding stimulus B, which was associated with the highest probability of monetary loss, was comparable across transfer phases, with both significantly above chance level (all>50%, $p<0.001$).

Abbreviations: TR, transfer.SD-standard deviation, **$p<0.01$, ***$p<0.001$, two-tailed tests.

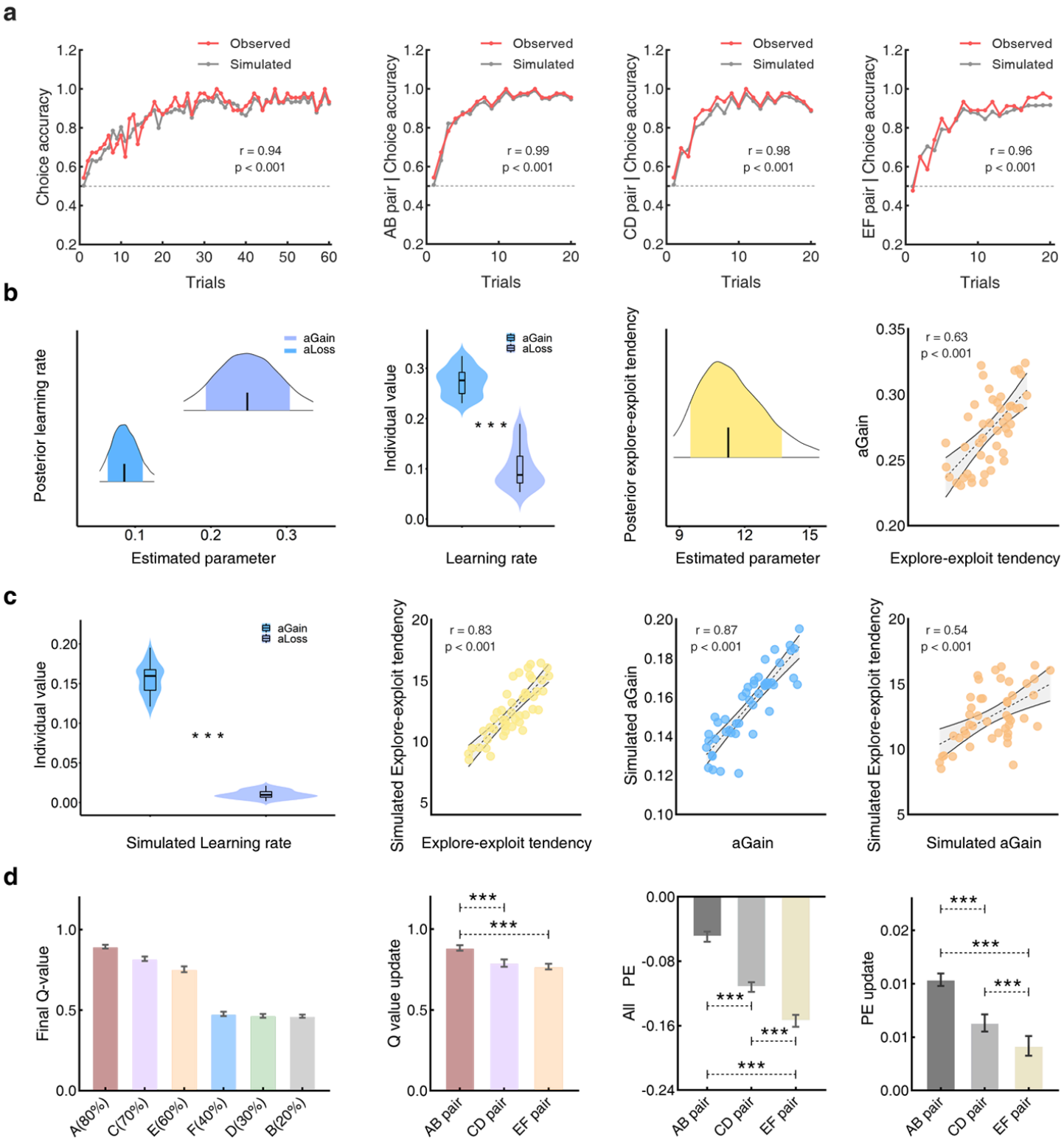

Figure 2

Computational characterization of latent value updating during early probabilistic learning (a) Comparison between simulated and observed trial-wise choices across stimulus pairs showed that the selected model closely reproduced the overall learning trajectory. Dashed lines indicate chance-level choice performance (50%). (b) In consistent with previous studies using dual learning rates, the learning rate for gain outcome (αGain) was higher than that for loss outcome (αLoss) on both group and

individual level, which indicated the “optimism bias” (40–42). Additionally the group-level estimate distribution of explore-exploit tendency parameter (β) was presented, while its individual-level estimate had positive correlation with individual-level learning rate for gain outcome. (c) Parameter recovery was assessed by refitting the selected model to 50 simulated datasets generated for each participant. The recovered parameters including learning rates (αGain, αLoss) and explore-exploit tendency (β) were checked based on procedures such that: 1) demonstrating the ‘optimism bias’ based on individual-level learning rates for gain and loss outcomes, 2) identifying correlation between the true explore-exploit tendency (β), true αGain and the corresponding recovered β, recovered αGain, separately, 3) proving positive correlation of recovered αGain and recovered β. (d) Trial-wise expected values (Q value) and prediction errors (PEs) estimated by the selected model systematically tracked changes in reward probability across stimuli and stimulus pairs. The error bars denote standard error to the mean value. ***p<0.001, two-tailed tests.

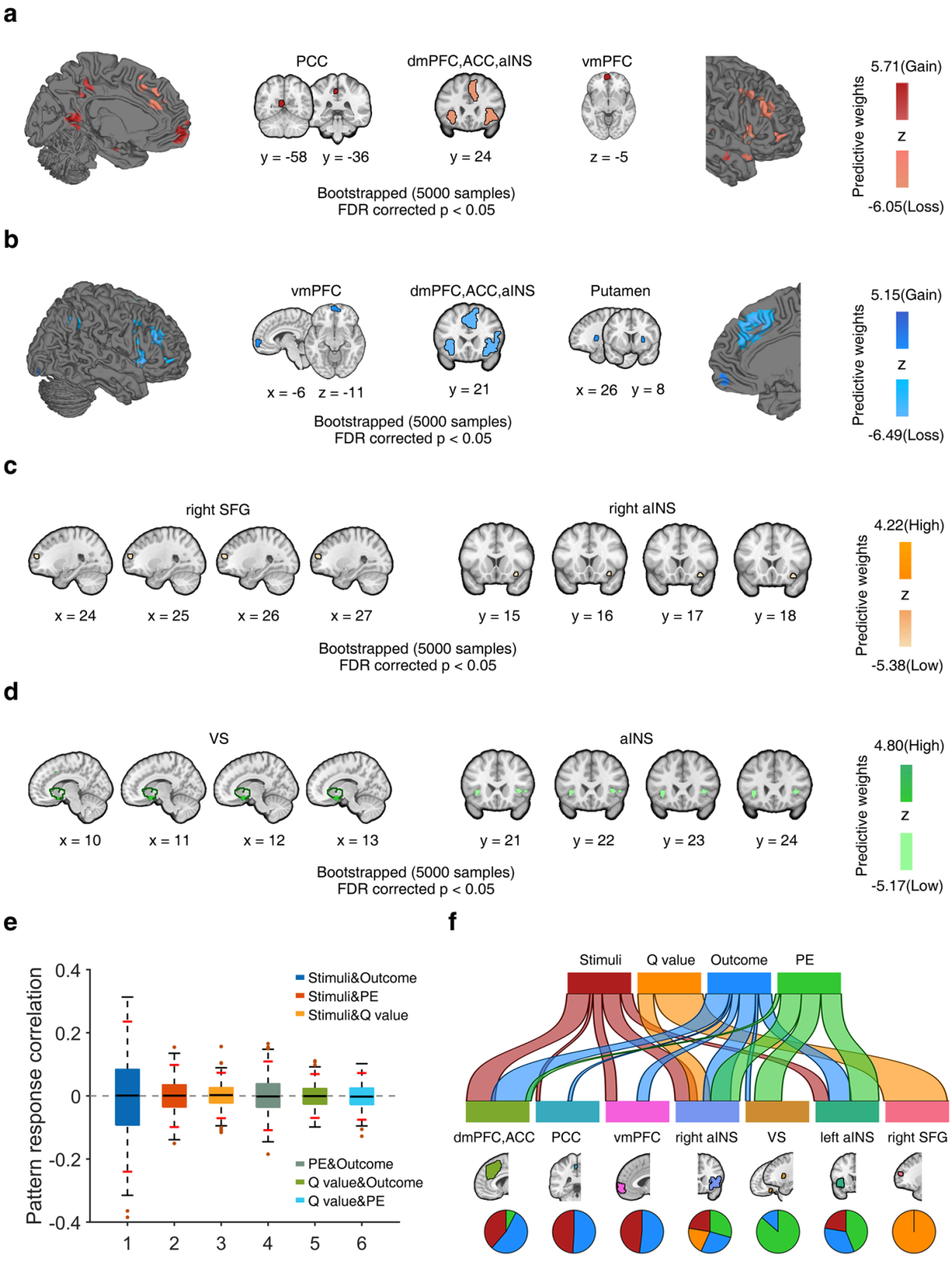

Figure 3

Multivariate neural patterns of observable and latent learned experiences (a)Whole-brain multivoxel pattern discriminating gain- and loss-related stimuli, thresholded at a false discovery rate (FDR) of $p < 0.05$ based on bootstrap distributions generated from 5,000 samples drawn with replacement. (b)Whole-brain multivoxel pattern discriminating gain- and loss-related outcomes, thresholded at FDR-corrected $p < 0.05$ based on 5,000 bootstrap samples. (c) Whole-brain multivoxel pattern predictive of expected

value (Q-value), thresholded at FDR-corrected $p < 0.05$ based on 5,000 bootstrap samples. (d) Whole-brain multivoxel pattern predictive of prediction error (PE), thresholded at FDR-corrected $p < 0.05$ based on 5,000 bootstrap samples. (e) Pairwise spatial similarities between the SVM classification weight maps and SVR regression weight maps for stimulus, outcome, Q-value, and PE representations. Short red lines indicate bootstrapped 95% confidence intervals based on 500 samples. The confidence intervals included zero, indicating little evidence for spatial similarity across these neural signatures. (f) River plots illustrate the spatial similarity, quantified using cosine similarity, between the multivoxel pattern maps and anatomical regions of interest (ROIs). Ribbon widths were normalized to the maximum cosine similarity across ROIs, with each multivoxel pattern map shown separately. Maps were thresholded at FDR-corrected $p < 0.05$, and only positive-weight voxels were retained for similarity analyses. Ribbon positions relative to the boxes are arbitrary. Pie charts show the relative contribution of each multivoxel pattern map to each ROI, defined as the percentage of voxels showing the highest cosine similarity with each pattern. Error bars indicate mean±SE where applicable.

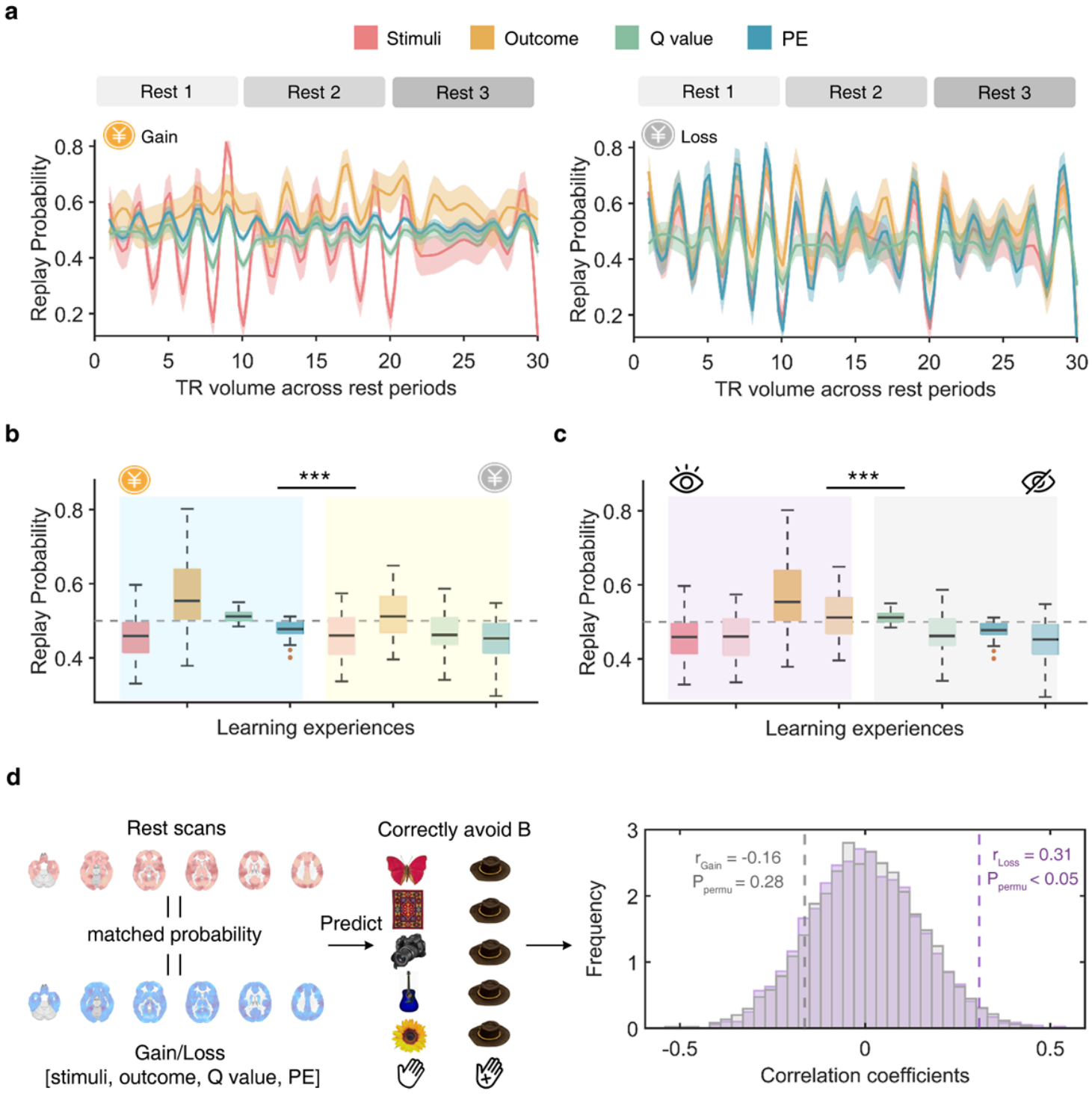

Figure 4

Differential offline reactivation of learned experiences and subsequent loss avoidance (a) Dynamic estimates of reactivation probabilities of gain- and loss-related learned experiences across rest periods. (b) Gain-related learned experiences showed significantly higher reactivation probabilities than loss-related experiences. (c) Observable experience components (stimulus and outcome) showed significantly higher reactivation probabilities than latent components (PE and Q value). (d) Greater reactivation of loss-related learned experiences was associated with higher accuracy in avoiding stimulus B during the subsequent transfer stage. *** p<0.001

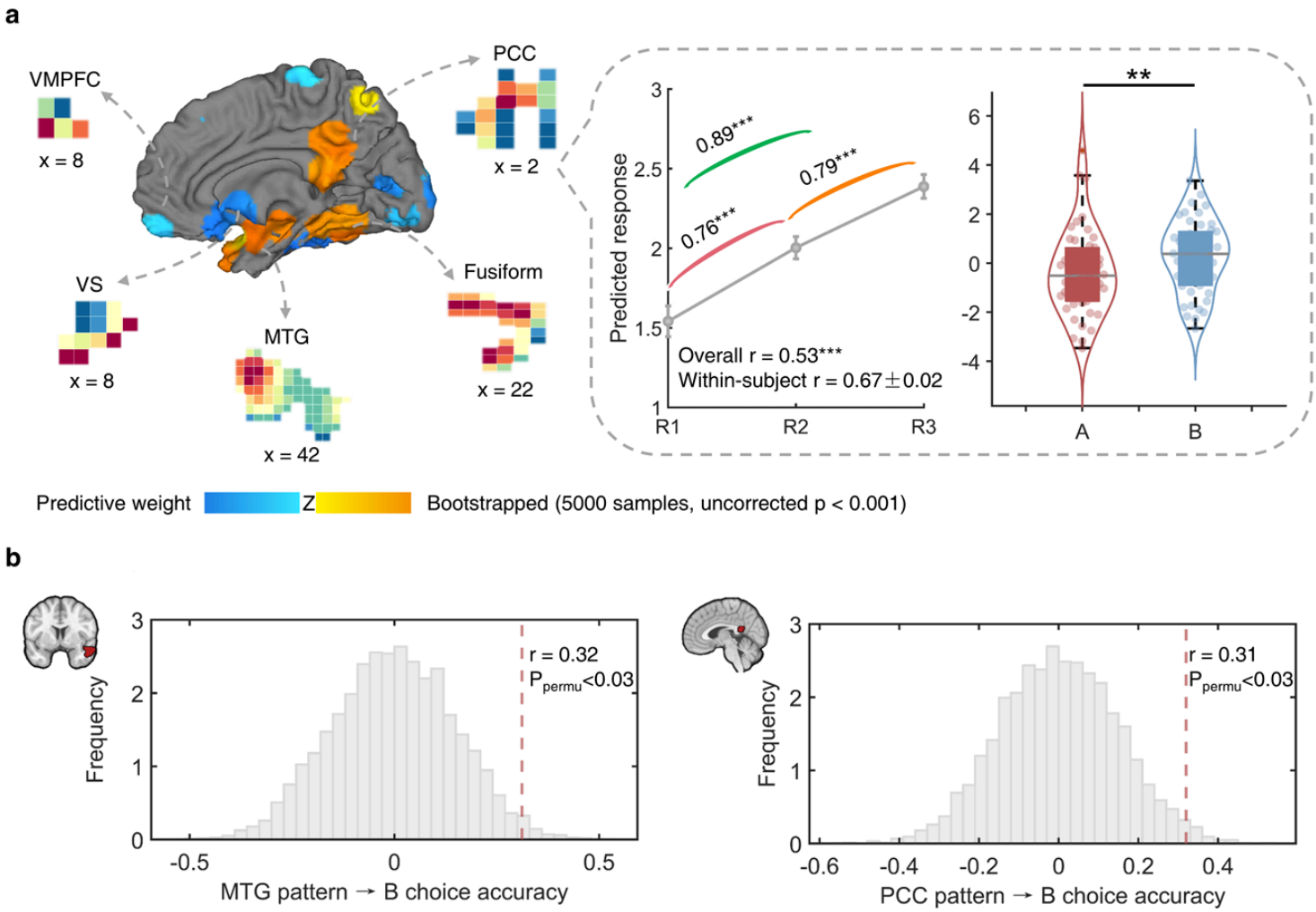

Figure 5

Brain systems engaging across post-learning rest and their association with subsequent transfer choices (a) Whole-brain multivariate pattern showing systematic changes across post-learning rest periods. For visualization, pattern weights are displayed at an uncorrected threshold of p<0.001, based on bootstrap distributions generated from 5,000 samples drawn with replacement. Consistent pattern weights were observed in the ventromedial prefrontal cortex (vmPFC), posterior cingulate cortex (PCC), ventral striatum (VS), middle temporal gyrus (MTG) and fusiform gyrus. Line plots show the predicted versus actual rest-period labels, and r denotes the Pearson correlation between predicted and observed rest-period indices. Forced-choice accuracies reflect pairwise discrimination between post-learning rest periods, with two-sided binomial tests used to assess performance above chance. Expression of this multivariate neural pattern was greater during loss-avoidance trials (avoiding B) than during reward-seeking trials. (b) Increased pattern expression in MTG and PCC was associated with higher accuracy in avoiding stimulus B during the first transfer stage. **p<0.01, ***p<0.001

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 003SI.docx
- 004nrreportingsummaryflat.pdf