Article

Current Biology

Semantic structures facilitate threat memory integration throughout the medial temporal lobe and medial prefrontal cortex

Highlights

- Mechanisms of indirect threat learning are revealed by MVPA of fMRI
- Emotional memory integration promotes generalization across semantic categories
- Amygdala and medial prefrontal cortex reinstate indirect threat memory
- Durable threat representation changes found in hippocampus and perirhinal cortex

Authors

Samuel E. Cooper, Augustin C. Hennings, Sophia A. Bibb, Jarrod A. Lewis-Peacock, Joseph E. Dunsmoor

Correspondence

samuel.cooper@austin.utexas.edu (S.E.C.), joseph.dunsmoor@austin.utexas.edu (J.E.D.)

In brief

How do new emotional experiences integrate with prior knowledge? Using sensory preconditioning to produce "threat learning by proxy," Cooper et al. show that threat memories integrate with and generalize across categories. The amygdala and medial prefrontal cortex show transient generalization, whereas the hippocampus and perirhinal cortex are more durable.

Cooper et al., 2024, Current Biology *34*, 3522–3536
August 5, 2024
https://doi.org/10.1016/j.cub.2024.06.071

CellPress

CellPress

Current Biology

Article

Semantic structures facilitate threat memory integration throughout the medial temporal lobe and medial prefrontal cortex

Samuel E. Cooper,[1,*] Augustin C. Hennings,[2] Sophia A. Bibb,[3] Jarrod A. Lewis-Peacock,[1,4,5,6] and Joseph E. Dunsmoor[1,5,6,7,*]

[1]Department of Psychiatry and Behavioral Sciences, University of Texas at Austin, Austin, TX, USA
[2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[3]Neuroscience Graduate Program, Ohio State University, Columbus, OH, USA
[4]Department of Psychology, University of Texas at Austin, Austin, TX, USA
[5]Center for Learning and Memory, University of Texas at Austin, Austin, TX, USA
[6]Department of Neuroscience, University of Texas at Austin, Austin, TX, USA
[7]Lead contact
*Correspondence: samuel.cooper@austin.utexas.edu (S.E.C.), joseph.dunsmoor@austin.utexas.edu (J.E.D.)
https://doi.org/10.1016/j.cub.2024.06.071

SUMMARY

Emotional experiences can profoundly impact our conceptual model of the world, modifying how we represent and remember a host of information even indirectly associated with that experienced in the past. Yet, how a new emotional experience infiltrates and spreads across pre-existing semantic knowledge structures (e.g., categories) is unknown. We used a modified aversive sensory preconditioning paradigm in fMRI (n = 35) to investigate whether threat memories integrate with a pre-established category to alter the representation of the entire category. We observed selective but transient changes in the representation of conceptually related items in the amygdala, medial prefrontal cortex, and occipitotemporal cortex following threat conditioning to a simple cue (geometric shape) pre-associated with a different, but related, set of category exemplars. These representational changes persisted beyond 24 h in the hippocampus and perirhinal cortex. Reactivation of the semantic category during threat conditioning, combined with activation of the hippocampus or medial prefrontal cortex, was predictive of subsequent amygdala reactivity toward novel category members at test. This provides evidence for online integration of emotional experiences into semantic categories, which then promotes threat generalization. Behaviorally, threat conditioning by proxy selectively and retroactively enhanced recognition memory and increased the perceived typicality of the semantic category indirectly associated with threat. These findings detail a complex route through which new emotional learning generalizes by modifying semantic structures built up over time and stored in memory as conceptual knowledge.

INTRODUCTION

Imagine you developed a fear of dogs after a terrifying encounter at a relative's house. As time goes by, you realize that not only do you avoid your relative's dog but also parks, hiking trails, and certain friends' houses—all locations that did not previously cause anxiety but where you know, through experience, that dogs might be off leash. This illustrates the complex relationships that humans draw upon to integrate emotional experiences into pre-existing knowledge structures, allowing us to draw meaningful inferences about the possibility of danger in the absence of direct knowledge. This cognitive process conforms to long-standing principles from learning theory[1] and is operationalized by paradigms wherein memories are indirectly modified through reinforcement of related stimuli, known as higher-order conditioning.[2,3]

A flexible learning and memory system that can efficiently update prior stimulus representations and semantic (knowledge) structures, or schemas,[4] with new learning is clearly adaptive[5]—we can predict the potential for harm without experiencing the negative consequences directly (e.g., if you avoid parks, you lower even the minuscule possibility of another dog bite at those locations). Conversely, an experience of threat that indiscriminately generalizes to an entire semantic structure can spread threat associations to harmless stimuli or situations only tangentially related to the negative experience (e.g., avoiding other commonly domesticated animals) (see Figure 1A for a visualized example). This maladaptive form of generalization is characteristic of many anxiety-related disorders, such as post-traumatic stress disorder (PTSD) and obsessive-compulsive disorder.[6–8] Although the ability to modify pre-existing semantic structures with new learning is a hallmark of human cognition, the neurobehavioral mechanisms by which threat learning might do this are poorly understood.

There have been at least two primary experimental approaches for studying how a new experience updates our

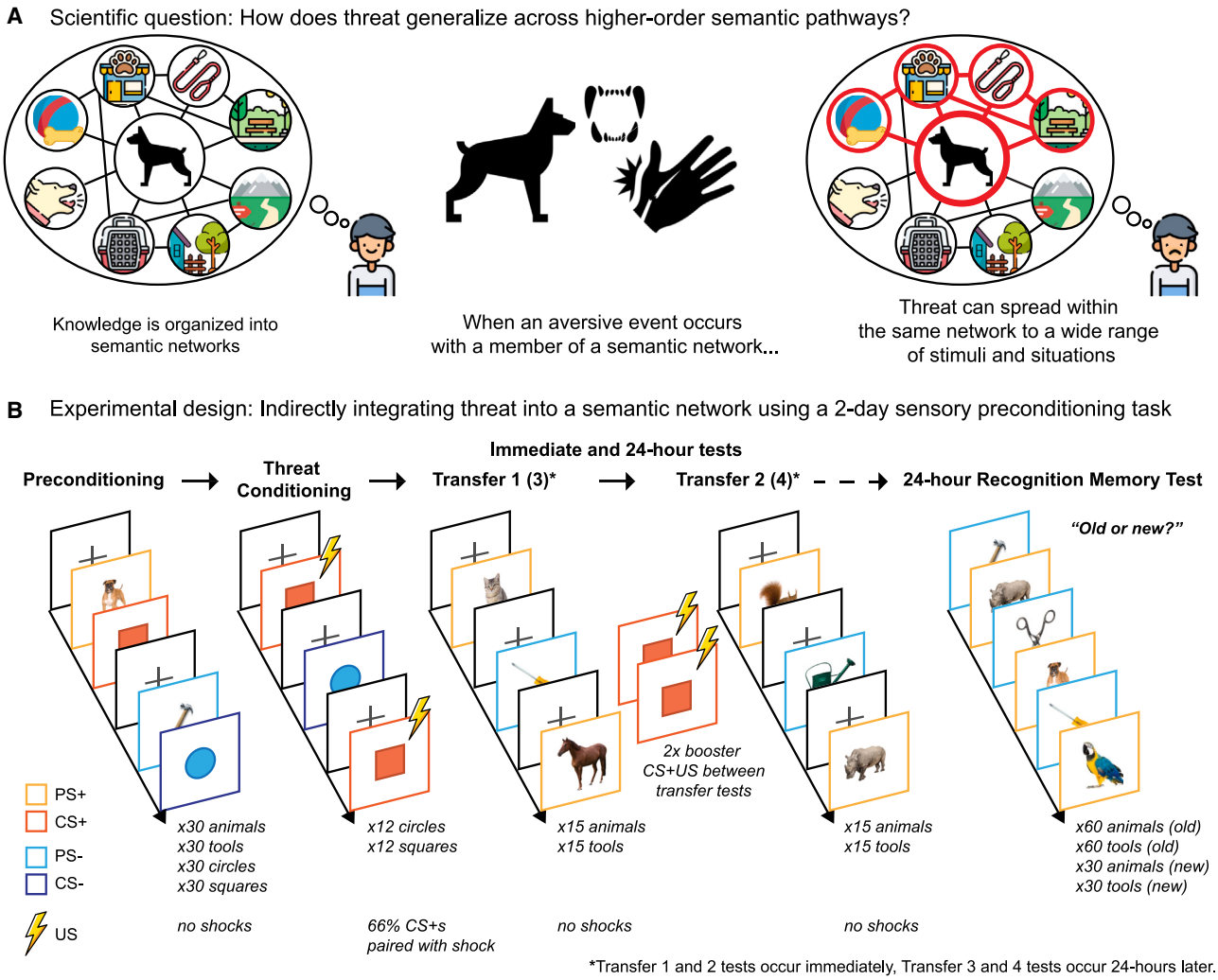

Figure 1. Conceptual and experimental overview of modified aversive sensory preconditioning procedure

(A) Conceptual overview. Threat associations can generalize from a traumatic incident (e.g., a dog bite) across similar conceptual associations, referred to as a semantic structure or network (e.g., things related to dogs, such as parks or pet stores), even when stimuli from these structures were not present nor perceptually resemble aspects of the traumatic incident.

(B) 2-day aversive sensory preconditioning task structure. Day 1: semantic categories (animal, tools) were paired with one of two neutral shapes (circle, square) during a preconditioning phase. One shape (CS+) was then paired with the US (shock) during a subsequent threat conditioning phase; the category paired with this shape is labeled the PS+, the other shape/category are CS− and PS−, respectively. Transfer phases immediately after conditioning (transfer 1 and 2) test for threat (CS+-US) associations generalizing to novel PS+ category items. Day 2: transfer tests were repeated with novel category exemplars (transfer 3 and 4) and followed by a recognition memory test for day 1 stimuli.

CS+, conditioned threat cue; CS−, conditioned safety cue; PS+, preconditioned threat cue; PS−, preconditioned safety cue; SCR, skin conductance response; US, unconditioned stimulus.

mental model of the world by modulating memories related to those experiences. One approach involves animal learning models that incorporate sensory preconditioning protocols. Sensory preconditioning is a well-established protocol in which animals first undergo a preconditioning phase where they learn an association between at least two arbitrary and affectively neutral stimuli (e.g., tone and light) in the absence of meaningful reinforcement (i.e., latent learning).[2,9–12] Then, one of the stimuli is used as a conditioned stimulus (CS, the light) in a learning (conditioning) phase where the animal learns it predicts a biologically salient unconditioned stimulus (US; e.g., a shock in threat conditioning). Finally, a transfer phase (also referred to as a retrieval phase) tests whether the preconditioned stimulus (PS, the tone) elicits a conditioned response (e.g., freezing) similar to that elicited by the CS during initial learning. Neurobiological research shows consistent involvement of the hippocampus, perirhinal cortex (PRC), and orbitofrontal cortex in learning and retrieval in sensory preconditioning tasks.[13] In *aversive* sensory preconditioning, the PRC cooperates with the basolateral amygdala (BLA) to coordinate indirect PS-US threat associations.[14,15] The other primary experimental approach involves episodic memory tasks in humans that require novel inferences about pairs or groups of previously encoded stimuli based on new learning. These tasks consistently

show engagement of the hippocampus and medial prefrontal cortex (mPFC) for integrating across overlapping stimuli or events to draw novel inferences.[5,16–19]

Integrating a new memory of threat with broad semantic structures could lead to widespread changes in how related concepts are perceived, appraised, and remembered.[20,21] In this way, the transfer of emotional value could go beyond directly experienced instances of a stimulus in the past and spread more widely through underlying associations built up over time and stored in memory as semantic structures. Using the earlier example, it is unnecessary to have prior experience with the vicious dog outside of your relative's house. Pre-existing knowledge of where dogs are likely to be encountered is sufficient to motivate avoidance of those locations after the attack. Neuroimaging research shows that threat conditioning modulates cortical representations of object concepts that are directly associated with threat.[22,23] However, whether and how neural mechanisms designed to integrate discrete memories provide a route to *indirectly* implant emotional value into existing semantic structures is unknown.

A major question concerns *when* memory integration occurs for events that overlap with previous experiences.[5,16,17,24] An online integration (i.e., mediated learning) account proposes that the mental representation of the PS is reactivated on CS trials during conditioning, thereby integrating the PS and US representations at the time of emotional learning.[15] Evidence from aversive sensory preconditioning tasks in rodents suggests that online integration requires the PRC, as temporary lesions of the PRC spare direct threat conditioning but prevent the transfer of threat learning and responding to the PS.[13] Additionally, non-aversive memory integration tasks in humans show that hippocampal activity at the time of learning predicts successful inference and preferences toward the paired preconditioned cues at test[18] (but see Wang et al.[25]).

Alternatively, the retrieval account of memory integration[26] emphasizes processes during the transfer phase. In this account, PS presentations elicit retrieval of the previously associated CS, which brings with it the representation of the US to inform behavior (i.e., chaining) without a mediated PS-US representation. Appetitive sensory preconditioning work suggests the orbitofrontal cortex is required to retrieve indirectly acquired positive-value information of the PS to predict novel outcomes during retrieval.[27]

These two accounts are not mutually exclusive. Dynamic behavioral demands might necessitate memory integration at the time of conditioning in some instances or at transfer test (retrieval) in others.[26] For example, online integration at the time of emotional learning might modify memory representations of specific instances of the PS pre-associated with the CS in preparation for reencountering that specific PS.[15,18,19] Retrieval-based integration might rely on pattern completion processes, subserved by the hippocampus,[24] or relational reasoning, subserved by the hippocampus and mPFC,[16] to draw upon a broader network of relational links when encountering novel instances of the PS at a transfer test.

Here, we investigated the neurobehavioral mechanisms by which emotional learning indirectly modulates the representation of object concepts through memory integration. We used fMRI while participants completed a novel 2-day aversive sensory preconditioning task and applied multivariate pattern analysis (MVPA) to test different accounts for how regions in the medial temporal lobe and the mPFC facilitate integration of a threat memory with a previously associated category. During preconditioning, trial-unique (non-repeating) exemplars from two semantic categories (animals or tools) served as PSs and were paired with a shape CSs (square or circle). Then, during threat conditioning, one shape (CS+) predicted an aversive electric shock, and the other was safe (CS−). Next, novel category exemplars from the PS categories (now referred to as PS+ and PS−, indirectly associated with the CS+ and CS−, respectively) were presented alone during two transfer tests, separated by a brief reminder of the CS-US association.[28] The next day, participants completed two more transfer tests in the scanner to assess whether representational changes persist over a 24-h period. The experiment concluded with a recognition memory test and subjective ratings of category typicality for the PSs encoded before and after threat conditioning on day 1.

Synthesizing research on aversive sensory preconditioning and non-affective memory integration, we predicted that aversive sensory preconditioning with trial-unique exemplars would be sufficient to generate a category-level association with the CS, which would then selectively modulate patterns of neural similarity for novel category members at test following direct CS-US learning. Specifically, we hypothesized that medial temporal lobe regions primarily implicated in aversive sensory preconditioning (PRC, hippocampus, amygdala) and cortical memory integration regions (mPFC) would demonstrate stronger pattern similarity for the PS+ category in comparison with the PS− category at test. We also hypothesized that category-selective occipitotemporal cortex regions would demonstrate the same pattern, reflecting their role in tracking semantic relationships.

To investigate the mechanisms underlying potential category-level modulation, we tested reinstatement of threat-specific CS+ neural patterns on PS+ trials at test (retrieval account). We hypothesized that this form of reinstatement would be evident in the medial temporal lobe and mPFC. We also tracked neural reactivation of the PS+ category during conditioning on CS+ trials, which would putatively support the recombination of the PS category and the CS at the time of threat conditioning (online integration). For this analysis, our hypothesis was limited to predicting that reactivation in category-selective occipitotemporal regions would relate to BLA activity to the PS+ at test, given strong rodent evidence for the role of the BLA in online integration. As an additional exploratory test, we also investigated the potentially moderating role of individual differences in neural activity in the hippocampus and mPFC, both key memory integration regions, in the relationship between increased category reactivation and threat-related BLA activity.

RESULTS

Behavioral results

Threat conditioning

Confirming successful differential threat acquisition, mean skin conductance responses (SCRs) were greater for the CS+ relative to the CS− ($\beta = 0.45$, $t_{wald}(34) = 4.97$, $p < 0.001$, 95% confidence interval [CI] [0.26, 0.63]), as were mean shock expectancy ratings

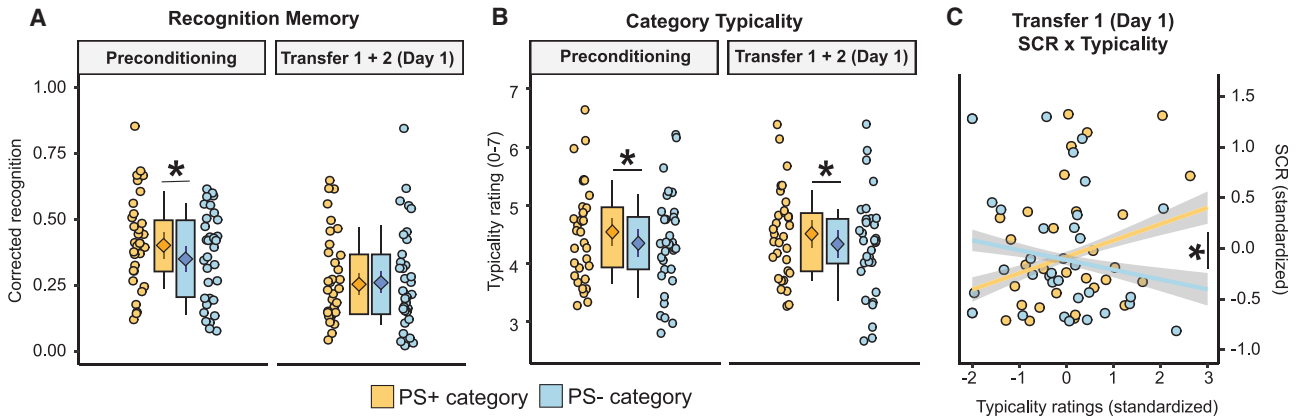

Figure 2. Behavioral and physiological evidence for successful aversive sensory preconditioning

A retroactive bias toward increased memory for PS+ (vs. PS−) items encoding during preconditioning is behavioral evidence for successful mediated learning (A). Further evidence includes increased reported mean typicality for PS+ (vs. PS−) items (B) and an immediate transfer 1 × SCR interaction (C), such that higher typicality for the PS+ category was associated with increased SCR (positive slope) to PS+s during transfer. For box-and-whisker plots, the box represents the middle 50% of the individual data points (fitted values; for recognition data, values are transformed back to the response scale). The shaded point and error bars inside the box represent the mixed-effects regression estimated marginal mean and 95% confidence intervals. Also see Figure S1 for behavioral and physiological data from the threat conditioning phase.

CS+, conditioned threat cue; CS−, conditioned safety cue; PS+, preconditioned threat cue; PS−, preconditioned safety cue; SCR, skin conductance response. *$p < 0.05$.

($\beta = 2.16$, $t_{wald}(33.5) = 41.18$, $p < 0.001$, 95% CI [2.05, 2.26]). See Figure S1 for plotted results.

Transfer tests

Mean SCRs were not significantly different between the PS+ and PS− during the immediate (day 1) or 24-h (day 2) transfer test ($ps \geq 0.085$). Given prior evidence that typicality influences category-level threat generalization,[29] we tested whether individual differences in participants' mean typicality for the PS+ category predicted arousal toward PS+ items at test. During the transfer phase (transfer 1) immediately after conditioning, category typicality significantly positively moderated the relationship between stimulus and SCRs ($\beta = 0.117$, $t_{wald}(1,490) = 1.99$, $p = 0.046$, 95% CI [0.01, 0.23]) (Figure 2C). This association between retrospective typicality ratings and within-session arousal was selective to the PS+ category. Typicality did significantly moderate in other experimental phases ($ps \geq 0.376$).

Participants did not report elevated shock expectancy ratings during the transfer tests. Mean ratings indicated an overall low likelihood of receiving a shock on either PS+ (day 1: $M = 1.175$; day 2: $M = 1.045$) or PS− (day 1: $M = 1.45$; day 2: $M = 1.14$) trials. Ratings were nominally enhanced on PS− trials during immediate transfer 1 and 24-h transfer 3 (transfer 1: $\beta = -0.227$, $t_{wald}(1,950) = -4.987$, $p < 0.001$, 95% CI [−0.31, −0.13]; transfer 3: $\beta = -0.152$, $t_{wald}(1,943) = -3.035$, $p = 0.002$, 95% CI [−0.24, −0.05]); there was no PS+ vs. PS− difference on either immediate transfer 2 or 24-h transfer 4 (transfer 2: $\beta = 0.011$, $t_{wald}(1,950) = 0.248$, $p = 0.803$, 95% CI [−0.078, 0.10]; transfer 4: $\beta < 0.001$, $t_{wald}(1,942) = -0.001$, $p = 0.999$, 95% CI [−0.09, 0.09]).

Threat conditioning retroactively enhances memory for items pre-associated with a conditioned stimulus

24-h recognition memory performance (controlling for false alarm rate) was significantly greater for PS+ vs. PS− items encoded prior to conditioning ($\beta = 0.226$, $z_{asymp.} = 2.28$, $p = 0.022$, 95% CI [0.03, 0.42]) (see Figure 2A). There was no significant difference between PS+ and PS− items encoded during the transfer tests immediately following conditioning ($\beta = -0.032$, $z_{asymp.} = -0.299$, $p = 0.765$, 95% CI [−0.24, 0.18]).

Subjective typicality ratings

Participants rated PS+ items (relative to PS− items) as being overall more typical of their semantic category. This included items encoded during preconditioning ($\beta = 0.10$, $t_{wald}(3,826) = 2.55$, $p = 0.010$, 95% CI [0.01, 0.18]) and the day 1 transfer tests (transfers 1 and 2) ($\beta = 0.09$, $t_{wald}(3,826) = 2.32$, $p = 0.020$, 95% CI [0.02, 0.18]) (see Figure 2B). This result suggests that threat-conditioning retroactively and proactively enhanced subjective stimulus typicality, in line with prior findings.[30]

Univariate analysis of aversive sensory preconditioning

Whole-brain analyses

Univariate whole-brain fMRI analysis (voxel wise $p \leq 0.001$, cluster corrected $p < 0.05$) of the CS+ > CS− contrast for threat conditioning found significant clusters consistent with prior threat conditioning meta-analyses[31] (see Table S2; Figure S3). Significant whole-brain clusters for the PS+ > PS− or PS− > PS+ univariate contrasts were not found during transfer phases or preconditioning.

Univariate ROI analyses

Activity in the BLA ($\beta = 0.31$, $t_{wald}(170) = 2.49$, $p = 0.013$, 95% CI [0.06, 0.56]) and hippocampus ($\beta = 0.32$, $t_{wald}(170) = 2.25$, $p = 0.025$, 95% CI [0.04, 0.64]) was significantly higher for the PS+ (vs. PS−) during the transfer 1 test. Across all regions of interest (ROIs), there were no other significant PS+ > PS− or PS− > PS+ activations during the transfer tests ($ps \geq 0.075$).

Pattern similarity analysis of aversive sensory preconditioning

To examine potential modulation of category-level representations resulting from indirect threat learning, multivariate patterns

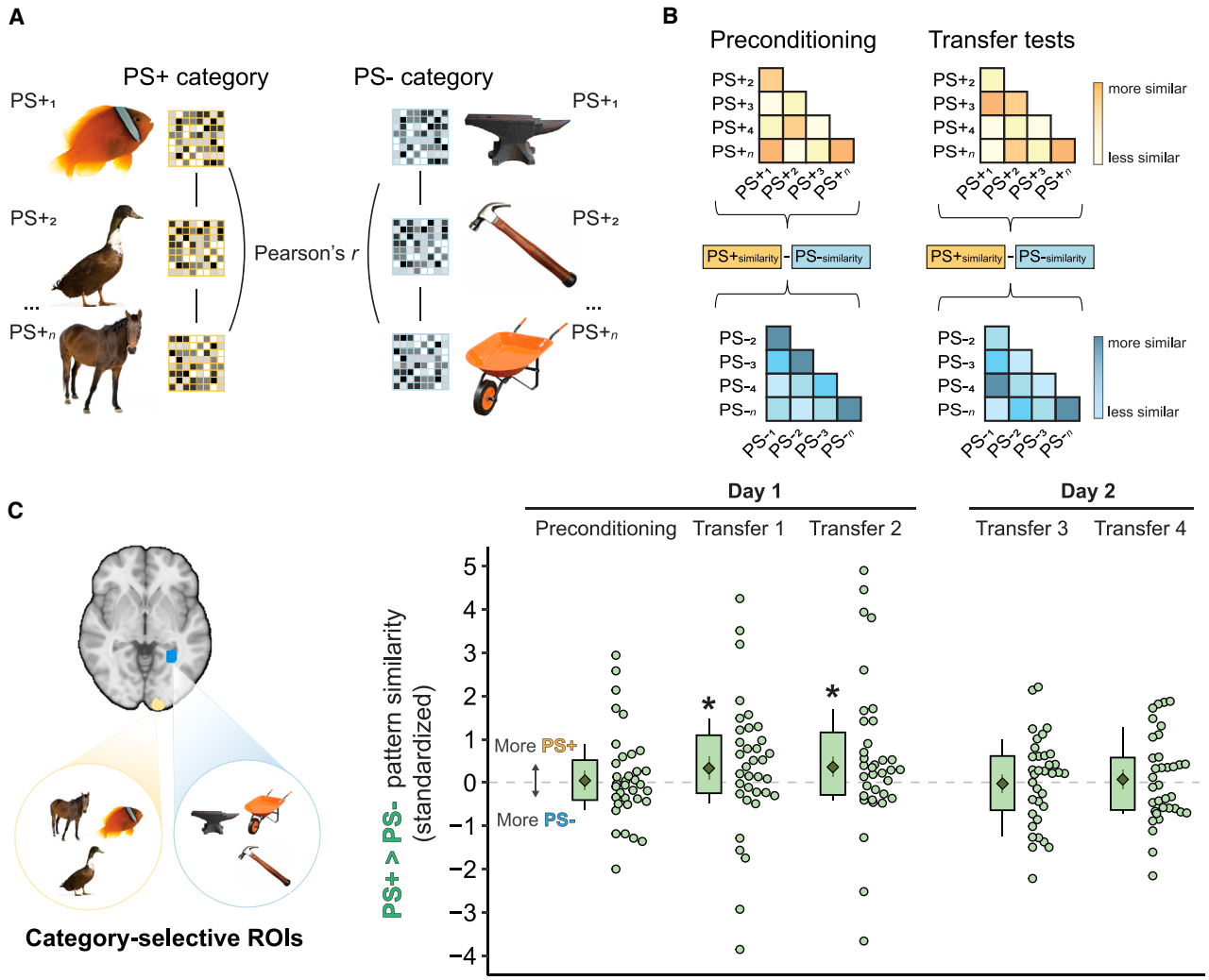

Figure 3. Schematic of within-category representational similarity analyses and category-cortex results

(A) Overview of within-category representational similarity analyses to test for category-level neural modulation after threat conditioning. For each phase, each multi-voxel pattern for each stimulus is correlated with all other stimuli from the same category for all possible pairs. Each trial is a unique category exemplar.

(B) For each phase, correlations from within each category are averaged to form an overall metric of within-category similarity. PS+ vs. PS− similarity scores from the same phase are then tested.

(C) Within-category representational similarity results for category-selective occipitotemporal regions across all task phases. Data are represented as PS+ > PS− difference scores for visualization purposes only; models/estimated marginal means incorporate separate PS+ and PS− values. Larger values indicate stronger similarity for the PS+ vs. PS−. For box-and-whisker plots, the box represents the middle 50% of the individual data points. Shaded point and error bars inside the box represent the mixed-effects regression estimated marginal mean and 95% confidence intervals. Statistical tests are conducted on the estimated marginal means visualized here, which are derived from outlier-resistant robust models that down-weight extreme individual values in tests. Also see Figure S4 for fitted values with separate PS+ and PS− values and Table S3 for full test statistics.

PS+, preconditioned threat cue; PS−, preconditioned safety cue.

*$p < 0.05$

of activation to each trial-unique PS item were correlated with the patterns from all other PS category exemplars encoded within the same experimental phase. As expected, there was no difference in pattern similarity between the PS+ and PS− categories at pre-conditioning in any *a priori* ROIs (*p*s $\geq$ 0.773; Figures 3C and 4; also see Table S3 for full test statistics and Figures S4 and S5 for plots showing PS+ and PS− values separately), providing benchmark evidence that semantic categories were not differentially represented in multi-voxel patterns of activity prior to conditioning.

Following conditioning, category-selective occipitotemporal regions (identified from the independent category localizer) exhibited enhanced pattern similarity for PS+ items vs. PS− items during day 1 transfer tests (transfer 1: $\beta = 0.33$, $t_{wald}(170) = 2.11$, $p = 0.036$, 95% CI [0.02, 0.64]; transfer 2: $\beta = 0.35$, $t_{wald}(170) = 2.28$, $p = 0.023$, 95% CI [0.04, 0.66]) (see Figure 3C). Notably, selectivity in pattern similarity between the PS+ vs. PS− categories was absent in these occipitotemporal regions after ~24 h (transfer 3: $\beta = -0.02$, $t_{wald}(170) = -0.15$, $p = 0.877$, 95% CI [−0.33, 0.28]; transfer 4: $\beta = 0.07$, $t_{wald}(170) = 0.46$, $p = 0.642$, 95% CI [−0.23, 0.38]).

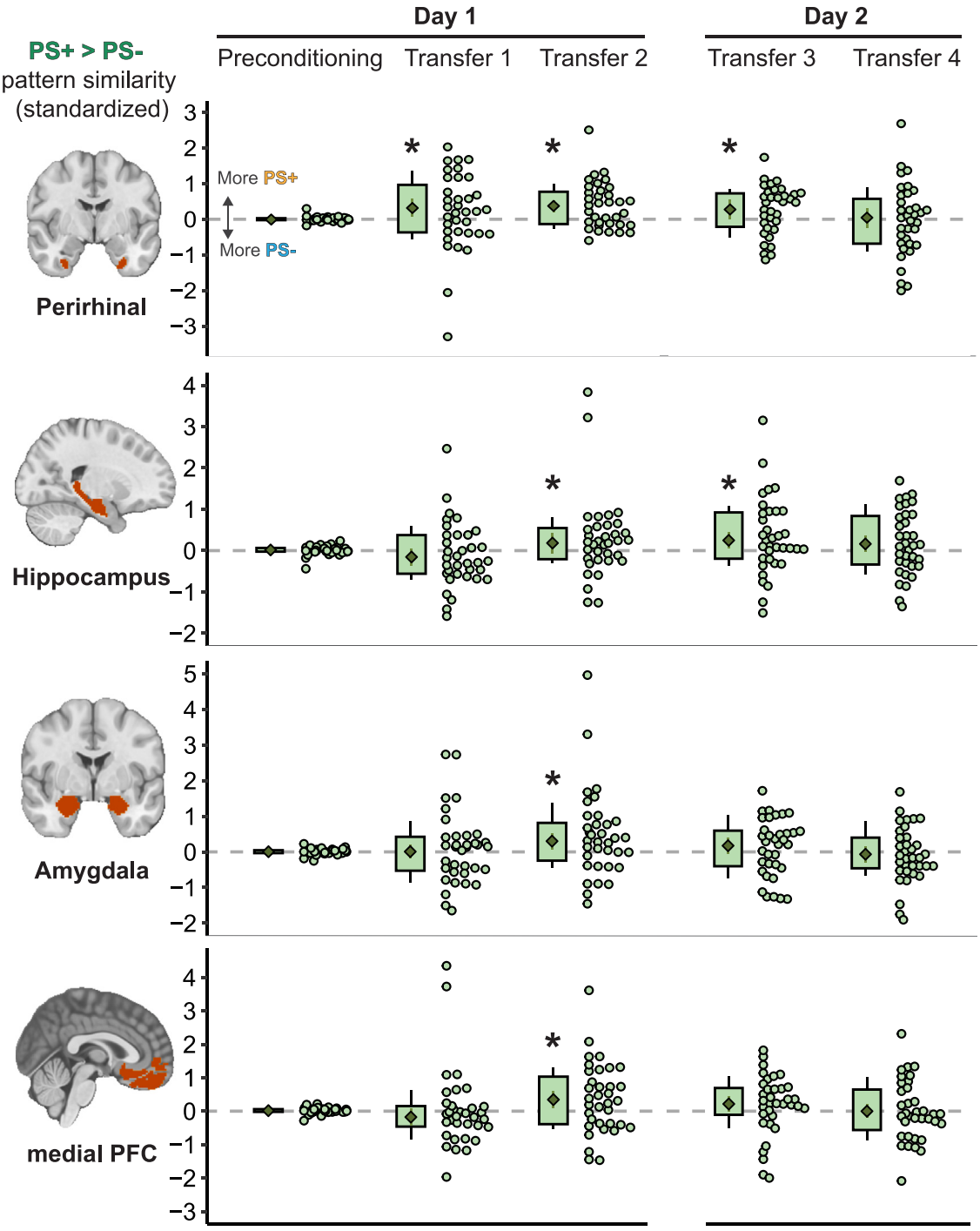

Figure 4. Within-category similarity across *a priori* anatomical ROIs

Perirhinal cortex and hippocampus showed increased PS+ similarity at immediate and 24-h transfer tests, whereas the amygdala and medial PFC only showed this effect at the immediate transfer test. The analytic steps used to produce these results are visualized in Figure 3. Data are represented as PS+ > PS− difference scores for visualization purposes only; all models and estimated marginal means incorporate separate PS+ and PS− values. Larger values indicate stronger similarity for the PS+ vs. PS−. For box-and-whisker plots, the box represents the middle 50% of the individual data points. Shaded point and error bars inside the box represent the mixed-effects regression estimated marginal mean and 95% confidence intervals. Statistical tests are conducted on the estimated marginal means visualized here, which are derived from outlier-resistant robust models that down-weight extreme individual values in tests. Also see Figure S5 for fitted values with separate PS+ and PS− values and Table S3 for full test statistics, as well as Figure S6 for visualization of exploratory analyses in entorhinal cortex. PFC, prefrontal cortex; PS+, preconditioned threat cue; PS−, preconditioned safety cue.

*$p < 0.05$.

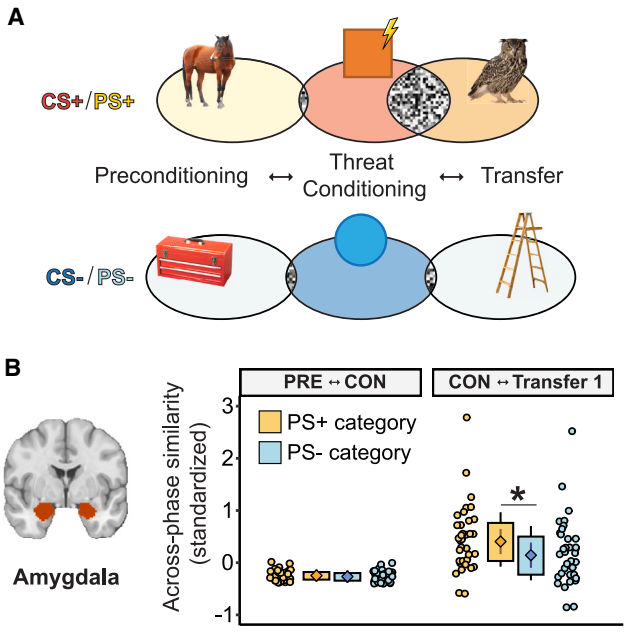

Figure 5. Testing for threat reinstatement via across-phase representational similarity analyses

(A) Overview of across-phase representational similarity analyses used to test for neural threat pattern reinstatement. Multi-voxel threat (CS+) and safety (CS−) patterns were correlated with the corresponding category pattern (CS+ to PS+, CS− to PS−) from either preconditioning or transfer phases. After threat conditioning, we predicted transfer PS+ patterns would more strongly resemble CS+ patterns, whereas preconditioning and PS− patterns would minimally resemble threat conditioning patterns.

(B) Neural threat reinstatement was observed in the amygdala during transfer, as PS+ similarity to CS+ patterns was increased relative to CS−/PS− similarity. As expected, no PS+/PS− difference was observed for preconditioning/threat conditioning patterns. For box-and-whisker plots, the boundaries of the box represent the middle 50% of the plotted individual data points. Inside the box, shaded point and error bars represent the mixed-effects regression estimated marginal mean and 95% confidence intervals. Statistical tests are conducted on the estimated marginal means visualized here, which are derived from outlier-resistant robust models that down-weight extreme individual values in tests.

CON, threat conditioning; PRE, preconditioning; PS+, preconditioned threat cue; PS−, preconditioned safety cue.

*$p < 0.05$.

Across *a priori* ROIs from the medial temporal lobe and mPFC (see Figure 4), there was enhanced pattern similarity for trial-unique items from the PS+ vs. PS− category in the PRC at both day 1 transfer tests (transfer 1: $\beta = 0.29$, $t_{wald}(170) = 3.05$, $p = 0.002$, 95% CI [0.10, 0.48]; transfer 2: $\beta = 0.40$, $t_{wald}(170) = 4.14$, $p < 0.001$, 95% CI [0.21, 0.59]). This selectivity in PS+ pattern similarity in the PRC extended to the 24-h test (transfer 3: $\beta = 0.33$, $t_{wald}(170) = 2.93$, $p < 0.001$, 95% CI [0.14, 0.53]). In the hippocampus, enhanced pattern similarity was observed during the second transfer test on day 1 ($\beta = 0.18$, $t_{wald}(170) = 2.16$, $p = 0.031$, 95% CI [0.01, 0.34]) and extended to the first transfer test on day 2 ($\beta = 0.24$, $t_{wald}(170) = 2.93$, $p = 0.003$, 95% CI [0.08, 0.41]). The amygdala ($\beta = 0.30$, $t_{wald}(170) = 3.29$, $p = 0.001$, 95% CI [0.12, 0.48]) and mPFC ($\beta = 0.34$, $t_{wald}(170) = 3.01$, $p = 0.003$, 95% CI [0.11, 0.57]) exhibited selectively enhanced PS+ pattern similarity during the second transfer test on day 1; however, this selectivity did not extend to day 2 in either region ($ps \geq 0.061$, see Table S3 for full test statistics).

Threat pattern reinstatement during transfer tests

Multivariate analyses have identified the amygdala and mPFC as reflecting neural threat patterns at tests of threat memory retrieval.[32–34] Here, we tested the degree to which multi-voxel threat patterns are reinstated during transfer tests. Multi-voxel activity patterns evoked by CS trials during conditioning were correlated with their corresponding pre-associated PSs during transfer as a form of encoding-retrieval similarity.[32,35,36] There was significant reinstatement of CS+ threat conditioning neural patterns in the amygdala during transfer 1 on PS+ trials (vs. CS−/PS− trials) ($\beta = 0.26$, $t_{wald}(306) = 3.04$, $p = 0.002$, 95% CI [0.09, 0.44]) (see Figure 5B). Threat-specific pattern reinstatement was selective to the first transfer test (other transfer tests, $ps \geq 0.816$). Interestingly, selective threat pattern reinstatement in the mPFC was observed in the second transfer test on day 1 ($\beta = 0.36$, $t_{wald}(306) = 4.72$, $p < 0.001$, 95% CI [0.21, 0.52]). All other mPFC comparisons, as well as the same tests within the PRC, hippocampus, and category-selective visual regions, were nonsignificant ($ps \geq 0.088$).

Evidence of online integration through cortical reactivation during aversive learning

Using our validated classifier, we estimated reactivation of the PS+ category (animals or tools, counterbalanced) during the presentation of the CS+ at the time of threat conditioning in category-selective cortices as an index of online emotional memory integration. According to the online integration account, CS trials will trigger reactivation of the PS representation, which will undergo modification as CS-US learning progresses throughout conditioning, thereby resulting in a modified PS representation at test.[26] Decoded PS− category reinstatement on CS− trials served as a comparison condition. For all analyses, decoded category reactivation is referred to by PS+ or PS− label, as for some participants, the animal category was the PS+, and the tool was PS−, and for others, vice versa. Decoding yields classifier evidence values (probability estimates); larger values indicate greater reactivation likelihood. Aligning with related prior work,[37] classifier evidence for PS+ category reactivation was well distributed ($M = 0.46$, SD = 0.15, interquartile range [IQR] = 0.22) and significantly differed from zero (one-sample t test, $t(33) = 16.8$, $p < 0.001$) (Figure 6B). Evidence values did not significantly differ based on which category was the PS+ or PS− per participant, $t(30) = -0.045$, $p = 0.964$. Similar results were found for the PS− category.

We then used individual participants' classifier evidence for the decoded PS+ (and PS−) on CS+ trials during conditioning to predict their degree of BLA activity during the transfer test on PS+ trials. Individual differences in BLA activity during transfer were the key focus of this analysis, given this region's presumed role in retrieving the modified association of the PS+, which was indirectly altered at the time of threat conditioning (online integration) through reactivation of the PS+ representation.[15] To test for moderation of this relationship, we expanded the interaction with another term in two parallel analyses: one with hippocampus univariate activity to CSs during threat

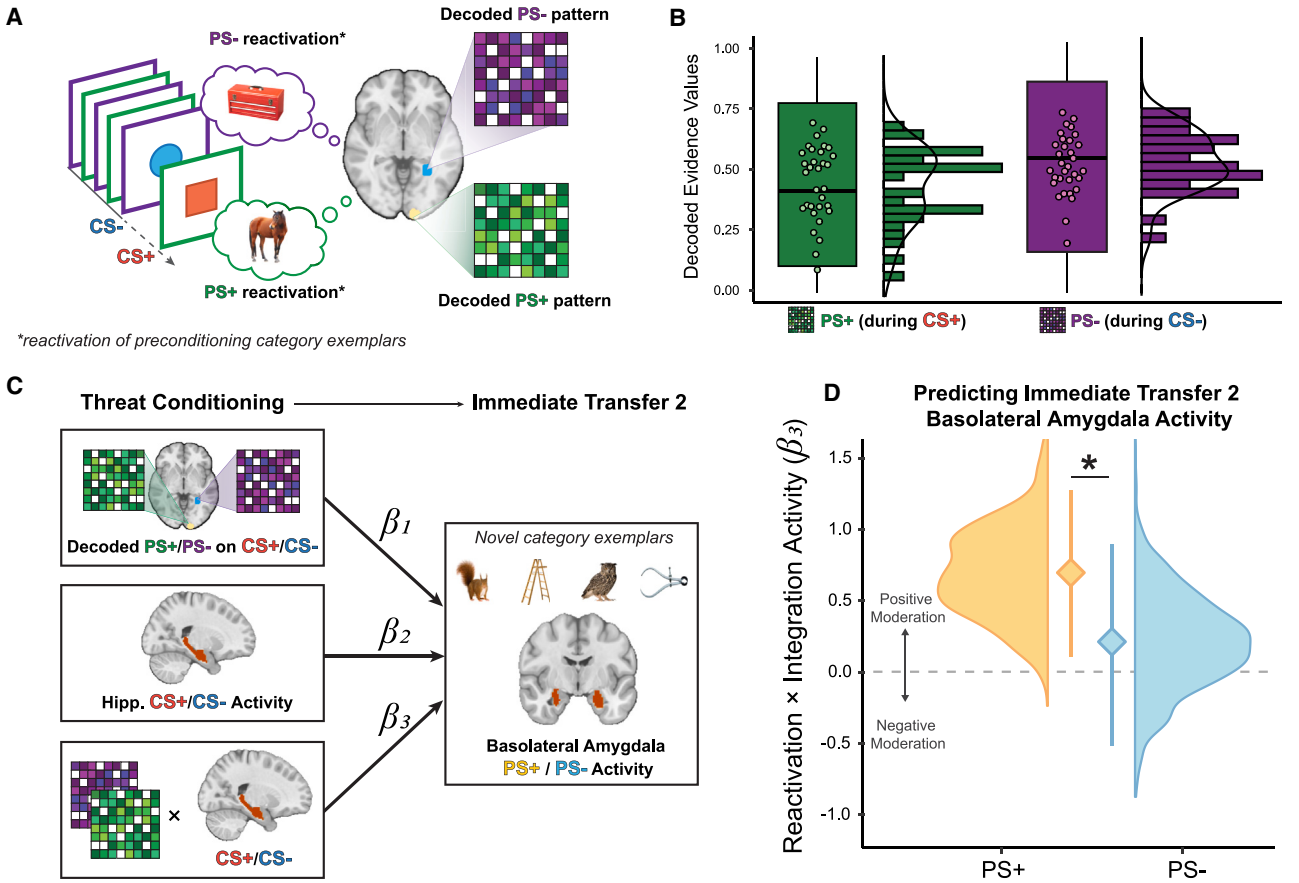

Figure 6. Decoding analyses reveal that category reactivation interacts with memory integration activity to predict generalized BLA activity for novel items from an indirectly threat conditioned category

(A) Schematic of fMRI decoding analyses. During threat conditioning, PS category reactivation is decoded using a classifier validated on localizer data. Decoding is conducted within category-selective occipitotemporal ROIs. Degree of category reactivation is indexed by evidence values reflecting the strength of the PS pattern reactivation.

(B) Distribution of decoded PS+ and PS− reactivation evidence values indicates sufficient variability for individual difference analyses. Error bars show SEM; bolded bands represent the median. Mean evidence values for both PS+ and PS− are significantly different than zero (*p*s < 0.001).

(C) Schematic of decoding moderation analyses. Beta coefficients (β) represent threat conditioning terms predicting the outcome variable (BLA activity during transfer 2). β_1 represents the decoded PS during the CS from category-selective cortex, β_2 represents univariate activity during a CS, and β_3 represents the $\beta_1 \times \beta_2$ interaction. The actual tested model is structured hierarchically via mixed-effects regressions, with all CS+/PS+ and CS−/PS− data in a multilevel (repeated-measures) stimulus term nested within each participant.

(D) Cortical reactivation and online integration results, indicating that the hippocampus significantly moderates the relationship between category cortex reactivation during threat conditioning and BLA activity. For visualization, separate interaction coefficients (decoded PS reactivation × CS activation) were extracted for transfer 2 PS+ and PS− conditions and bootstrapped (*k* = 1,000). Coefficients, 95% confidence intervals, and distributions are plotted against zero to demonstrate interaction significance; confidence intervals not overlapping with zero are significant. We only visualize the hippocampus moderation analysis here; the medial prefrontal cortex was also a significant moderator.

BLA, basolateral amygdala; CS+, conditioned threat cue; CS−, conditioned safety cue; hipp., hippocampus; PS+, preconditioned threat cue; PS−, preconditioned safety cue.

*$p < 0.05$.

conditioning, the other with mPFC activity to CSs during threat conditioning. We focused on these two regions due to their prominence in the memory literature as key hubs for episodic memory integration.[16] To determine whether the addition of a separate hippocampus or mPFC univariate activity term to the model, both as a separate term and then added to the interaction term, significantly improved model fit (as improved fit for the model with the interaction is a requirement for formal moderation analyses), we conducted likelihood ratio tests comparing models with and without the expanded interaction (chi-squared distribution, significance at $p < 0.05$). All models continued to include the repeated-measures stimulus term (CS+/PS+/, CS−/PS−).

Reactivation of the PSs during CS trials did not selectively predict BLA activity during either transfer phase. Addition of the CS univariate activity term (without adding it to the interaction term) resulted in significantly improved model fits for both hippocampal ($\chi^2(2) = 10.868$, $p = 0.004$) and mPFC activity ($\chi^2(2) = 10.672$, $p = 0.004$). Importantly, interacting PS reactivation with CS univariate activity in the hippocampus ($\chi^2(3) = 12.723$, $p = 0.005$)

or mPFC ($\chi^2(3) = 11.6$, $p = 0.008$) during threat conditioning resulted in models that significantly predicted BLA activity during immediate transfer 2 (see Figure 6C for schematic of interaction that includes hippocampal CS activity). Probing these interactions (moderations) revealed that increased BLA PS+ (vs. PS−) during transfer 2 was selectively related to the interaction of increased PS+ (vs. PS−) reactivation and increased univariate activity in the hippocampus ($\beta = 0.52$, $t_{wald}(53) = 2.43$, $p = 0.015$, 95% CI [0.10, 0.95]) (see Figure 6D) or mPFC ($\beta = 0.45$, $t_{wald}(53) = 2.24$, $p = 0.025$, 95% CI [0.06, 0.85]) during CS+ (vs. CS−) trials, confirming this effect as related to integrated aversive conditioning.

DISCUSSION

The complexity of human experience necessitates a flexible memory system that can adapt to a range of novel experiences and efficiently update existing memories to reflect new information. Semantic structures, built up over multiple experiences, help facilitate this process by providing a scaffold for inference and behavioral selection, even without direct experience of potential consequences. Prior work implies that semantic structures could provide ingress points for learned threats to enter and then broadly generalize across semantic networks.[38–40] However, the mechanisms by which emotional experiences integrate with previously acquired knowledge to modify the meaning and salience of different stimuli indirectly related to the experience have not been directly tested. The current study revealed potential neural mechanisms for building integrated memories of threat within a semantic structure, with the majority of our *a priori* predictions supported by the current data.

In accordance with our hypotheses, pre-association of a set of category exemplars with a to-be-conditioned threat stimulus modified the neural representation of unique category exemplars within category-selective occipitotemporal regions, the mPFC, and medial temporal lobe regions that include the amygdala, hippocampus, and PRC. Specifically, pattern similarity among unique exemplars pre-associated with a threat cue became more similar following threat conditioning. Enhanced neural similarity could facilitate the transfer of emotional learning to a diverse set of category exemplars despite their physical distinctions. This finding is consistent with a prior report of increased neural similarity in the occipitotemporal cortex and the amygdala for category-level stimuli directly predictive of an aversive US (direct conditioning)[23] but extends this finding to a higher-order learning paradigm (sensory preconditioning) that necessitates integrating across separate phases of learning. Modulation was transient in the occipitotemporal cortex and the amygdala but persisted beyond 24 h in the hippocampus and PRC, suggesting separation between immediate and longer-term changes in neural organization among these regions.

Prior studies establish the PRC's role in storing semantic information[41,42] and show that representational similarity covaries with semantic and visual dimensions.[43,44] Here, in line with our hypotheses, we observed modulation from emotional learning of stimulus representations in the PRC that persisted beyond the initial test. The hippocampus also maintained increased within-category similarity for the indirectly threat-conditioned category, which is consistent with its central role in threat-related delayed recall and retrieval.[32,45,46] These findings extend reports of PRC and hippocampal involvement in sensory preconditioning[13,16] to demonstrate how the selectively modified representations persist at 24 h following preconditioning and aversive learning.

At the moment of an emotional experience, do we immediately integrate this event with distinct past memories? Or does integration occur upon encountering a new situation that requires retrieval of the emotional memory? Prior research points to neither memory integration process as unilaterally predominant, with both forms of integration involved when memory guides decision-making.[5,24,26] Our results suggest that humans employ both routes when integration involves modulation of pre-established semantic structures.

First, individual difference analysis of category reactivation during conditioning supports an online integration account: selective reactivation of an indirectly conditioned category (PS) in occipitotemporal regions during threat conditioning trials (CS trials) interacted with increased hippocampal or mPFC activity to predict BLA activity during immediate transfer trials. Psychologically, this suggests that retrieval of the pre-associated category representation is integrated into the newly formed threat memory during emotional learning. These results align with prior neuroimaging work showing evidence of online integration of episodic memory[18,47–49] as well as studies showing reactivated category-selective voxels predicting responses during a subsequent retrieval test in aversive learning[50] and associative inference tasks.[49,51] Hippocampal and mPFC involvement during integration that subserves later retrieval is central to prominent episodic memory models.[16,17] Here, interaction between participant-level increases in online integration of reactivation of a semantic category representation and increases in activity in the hippocampus or mPFC during threat conditioning supported increased activity to novel PS+ presentations in the BLA, a prominent region in neural models describing emotion-episodic memory interactions.[52,53] That said, our hypotheses regarding online integration were only partially supported, as decoded reactivation in occipitotemporal regions alone did not predict subsequent BLA activity: it was only through the interaction with hippocampal or mPFC activity at the moment of learning. This suggests an important role for memory formation and retrieval regions that interact with representations in higher-order visual cortex to promote subsequent concept-based generalizations, although further study is needed to clarify the individual differences we observed in this integration process.

Alternatively, there is evidence of retrieval-based integration (i.e., chaining) from selective reinstatement of the threat-specific neural pattern during immediate transfer tests in the amygdala and mPFC. Specifically, overlapping fMRI patterns were selectively correlated with the formation of a threat memory on CS+ trials and retrieval of a threat memory on PS+ trials. One possibility is that the amygdala and mPFC play a more domain-general role in the retrieval of value information at the time of retrieval. Reinstating patterns specific to CS+ on PS+ trials could reflect the general affective salience of the PS+ cues following threat conditioning; the mPFC could support a model-based inference[25] that helps evaluate unique instances of the PS category not directly encountered during pre-conditioning and thus lack a directly learned PS-CS association. Consequently,

chaining in these regions could prioritize reinstatement of the CS-US relation to promote pattern completion and decrease threat discrimination between physically dissimilar category exemplars. Hypotheses here were only partially supported, as threat memory reinstatement was not observed in category-selective visual cortex, hippocampus, or PRC, ventral visual stream components specialized in object recognition[54,55] and regions yielding online integration evidence, as detailed above. In this way, regions tuned toward object recognition and representation might facilitate memory integration through reactivation of previously encountered instances directly associated with the CS. Conversely, regions with a domain-general role regarding value-related information might reinstate the threat memory when the properties of a novel stimulus must be inferred from related, but distinct, past events.

Recognition memory was selectively (PS+ > PS−) and retroactively enhanced for items encoded prior to threat conditioning, in line with tests of direct threat conditioning of categories.[30,56] These results accord with accelerating research on behavioral tagging[57,58] of human episodic memory,[59] which proposes that salient events can rescue weak memories formed minutes-to-hours before (or after) the salient event. Sensory preconditioning, per se, does not require a tag-and-capture mechanism, as there is no evidence to our knowledge that the time window between preconditioning and threat conditioning is a boundary condition for sensory preconditioning to be effective. However, the sensory preconditioning protocol could initiate a behavioral tagging mechanism, given that the design involves "weak" learning (PS-CS pairs) followed by a salient event (CS-US pairs). Here, the timing and overlapping events likely produced the retroactive memory benefit, which aligns with a recent study using a similar preconditioning design that found retroactive enhancement for relational episodic memory.[19] Importantly, selective enhancement in our study was not symmetrical; there was no proactive benefit on memories encoded during the transfer phase (as sometimes seen in prior work[56,60]). The transfer test might not have constituted weak learning (in accordance with behavioral tagging phenomenon), as although participants did not explicitly expect shock, arousal was significantly elevated on typical PS+ items and neuroimaging results showed strong evidence of modulation of PS+ representations during transfer within memory formation areas.

Although physiological arousal during transfer was differentially affected by participants' ratings of category typicality, it is notable that physiological arousal was not maintained throughout the test, and participants did not report that they expected shock to PS+ items. As such, threat transfer via sensory preconditioning was not directly evident from our behavioral measures, a result that aligns with reported difficulty in eliciting robust expression in humans using these types of protocols.[61,62] Desynchrony between threat measures is well-documented,[63,64] particularly between neural and behavioral or subjective outcomes.[65] Here, this might reflect an adaptive desynchrony in which the brain encodes higher-order relationships and their threat salience, as this is relatively efficient and expends minimal resources but does not necessarily evoke behavioral generalization. Another possible explanation is the "strong situation" theory, which describes individual differences in threat learning as a function of experimental threat salience.[66] Strong threat situations refer to contexts with sufficient biological salience to provoke an adaptive and normative response from most, resulting in near-zero response variability. In contrast, weak threat situations are those in which a clearly adaptive response is not required by the encountered threat, leading to natural response variability that can be accounted for by other variation sources. For example, most people who see a honking truck about to hit them will step away to avoid severe harm, whereas for a truck farther away, some might attempt to quickly cross while others will wait. Importantly, without sufficient relevant between-person variation (e.g., differences in psychopathology), a weak situation can become a strong (i.e., zero variability) situation. Our protocol is possibly a weak situation for neural differences, but as the threat threshold for behavioral expression is elevated, it might function as a strong situation only when testing those without threat-related psychopathology (e.g., PTSD).

The limitations of the current effort should be mentioned. Notably, we did not test brain-behavior relationships (e.g., SCR related to neural integration indices) due to increased model complexity (i.e., adding an additional interaction term) resulting in model convergence issues and more general concerns about power and reproducibility.[67] There are also limitations in terms of the ecological validity of the current design. Our CS+ "booster trials" separating transfer phases help reduce arousal habituation during transfer phases,[28] but an explicit, unambiguous reminder of the CS+/US association is unlikely to be encountered in the real world. Additionally, our study does not address how a single instance of threat learning (e.g., a single dog attack) leads to widespread semantic generalization. Our experimental design used repeated instances of category exemplars to generate a strong category-to-CS link, whereas real-life scenarios would likely involve a more isolated category member. The concept learning literature suggests that the basic level concept (e.g., a dog) is the most accessible entry point toward generalization to the superordinate category.[68] But whether higher-order fear generalizes to a broader semantic category based on a more isolated association between a single category exemplar and a CS is an open question.

The integration of emotional memories with existing knowledge is a pillar of human inference in a dynamic and sometimes dangerous world, but this process has received limited empirical attention. We used multivariate fMRI analyses to provide neurobehavioral explanations for the indirect integration of aversive learning with a pre-established semantic structure. Evidence for online and retrieval integration was dependent on neural region and analysis, supporting a "which-when" memory integration account[26] while also suggesting next steps for further delineating the precise neural circuity underlying these processes. However, after ~24 h, canonical threat regions did not maintain threat-related neural representations, and overall behavioral and physiological expression of threat learning was limited, both of which are possibly related to testing a psychiatrically healthy sample. PTSD is empirically related to increased threat-related neural activity at 24-h recall,[69,70] and PTSD symptomology is conceptually consistent with persistent heightened higher-order threat learning.[71,72] As such, this effort provides a potential experimental tool for testing subtle pathogenic processes, as we expect that stronger ~24-h reactivation and behavioral

expression would emerge in relation to PTSD and potentially other forms of anxiety-related psychopathology.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Stimuli
  - Task and procedures
  - Skin conductance response
  - Functional MRI acquisition
  - Image preprocessing
  - General linear models and whole-brain analyses
  - ROI selection and parameter estimate extraction
  - Perceptual localizer
  - Multivariate pattern analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.06.071.

ACKNOWLEDGMENTS

The authors would like to thank Ayesha Nadiadwala, Ryan Webler, Nicole Keller, and Josh Cisler for helpful discussions, as well as Ameera Azar, Andrew Spires, Natasha Muppidi, Raymond Truong, and Rithvik Pakala for assistance with data collection. S.E.C. is funded by the NIH (F32 MH129136). J.A.L.-P. is funded by the NIH (R01 EY028746 and R01 MH129042). J.E.D. is funded by the NIH (R01 MH122387) and the NSF (CAREER award 1844792).

AUTHOR CONTRIBUTIONS

S.E.C.: conceptualization, methodology, software, investigation, project administration, data curation, formal analysis, visualization, writing – original draft, and writing – review & editing. A.C.H.: methodology, software, and writing – review & editing. S.A.B.: investigation, project administration, data curation, and writing – review & editing. J.A.L.-P.: methodology, supervision, and writing – review & editing. J.E.D.: conceptualization, methodology, visualization, supervision, funding acquisition, resources, writing – original draft, and writing – review & editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 27, 2024
Revised: May 20, 2024
Accepted: June 26, 2024
Published: July 25, 2024

REFERENCES

1. Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208. https://doi.org/10.1037/h0061626.
2. Gewirtz, J.C., and Davis, M. (2000). Using Pavlovian Higher-Order Conditioning Paradigms to Investigate the Neural Substrates of Emotional Learning and Memory. Learn. Mem. *7*, 257–266. https://doi.org/10.1101/lm.35200.
3. Gostolupce, D., Lay, B.P.P., Maes, E.J.P., and Iordanova, M.D. (2022). Understanding Associative Learning Through Higher-Order Conditioning. Front. Behav. Neurosci. *16*, 845616.
4. Ghosh, V.E., and Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. Neuropsychologia *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010.
5. Shohamy, D., and Daw, N.D. (2015). Integrating memories to guide decisions. Curr. Opin. Behav. Sci. *5*, 85–90. https://doi.org/10.1016/j.cobeha.2015.08.010.
6. Cooper, S.E., van Dis, E.A.M., Hagenaars, M.A., Krypotos, A.-M., Nemeroff, C.B., Lissek, S., Engelhard, I.M., and Dunsmoor, J.E. (2022). A meta-analysis of conditioned fear generalization in anxiety-related disorders. Neuropsychopharmacology *47*, 1652–1661. https://doi.org/10.1038/s41386-022-01332-2.
7. Cooper, S.E., and Dunsmoor, J.E. (2021). Fear conditioning and extinction in obsessive-compulsive disorder: A systematic review. Neurosci. Biobehav. Rev. *129*, 75–94. https://doi.org/10.1016/j.neubiorev.2021.07.026.
8. Fraunfelter, L., Gerdes, A.B.M., and Alpers, G.W. (2022). Fear one, fear them all: A systematic review and meta-analysis of fear generalization in pathological anxiety. Neurosci. Biobehav. Rev. *139*, 104707. https://doi.org/10.1016/j.neubiorev.2022.104707.
9. Brogden, W.J. (1939). Sensory pre-conditioning. J. Exp. Psychol. *25*, 323–332. https://doi.org/10.1037/h0058944.
10. Brogden, W.J. (1947). Sensory preconditioning of human subjects. J. Exp. Psychol. *37*, 527–539. https://doi.org/10.1037/h0058465.
11. Rescorla, R.A. (1980). Simultaneous and successive associations in sensory preconditioning. J. Exp. Psychol. Anim. Behav. Process. *6*, 207–216. https://doi.org/10.1037/0097-7403.6.3.207.
12. Rizley, R.C., and Rescorla, R.A. (1972). Associations in second-order conditioning and sensory preconditioning. J. Comp. Physiol. Psychol. *81*, 1–11. https://doi.org/10.1037/h0033333.
13. Holmes, N.M., Fam, J.P., Clemens, K.J., Laurent, V., and Westbrook, R.F. (2022). The neural substrates of higher-order conditioning: a review. Neurosci. Biobehav. Rev. *138*, 104687. https://doi.org/10.1016/j.neubiorev.2022.104687.
14. Holmes, N.M., Parkes, S.L., Killcross, A.S., and Westbrook, R.F. (2013). The Basolateral Amygdala Is Critical for Learning about Neutral Stimuli in the Presence of Danger, and the Perirhinal Cortex Is Critical in the Absence of Danger. J. Neurosci. *33*, 13112–13125. https://doi.org/10.1523/JNEUROSCI.1998-13.2013.
15. Wong, F.S., Westbrook, R.F., and Holmes, N.M. (2019). "Online" integration of sensory and fear memories in the rat medial temporal lobe. eLife *8*, e47085. https://doi.org/10.7554/eLife.47085.
16. Schlichting, M.L., and Preston, A.R. (2015). Memory integration: neural mechanisms and implications for behavior. Curr. Opin. Behav. Sci. *1*, 1–8. https://doi.org/10.1016/j.cobeha.2014.07.005.
17. Zeithamova, D., Schlichting, M.L., and Preston, A.R. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. Front. Hum. Neurosci. *6*, 70. https://doi.org/10.3389/fnhum.2012.00070.
18. Wimmer, G.E., and Shohamy, D. (2012). Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. Science *338*, 270–273. https://doi.org/10.1126/science.1223252.
19. Zhu, Y., Zeng, Y., Ren, J., Zhang, L., Chen, C., Fernandez, G., and Qin, S. (2022). Emotional learning retroactively promotes memory integration through rapid neural reactivation and reorganization. eLife *11*, e60190. https://doi.org/10.7554/eLife.60190.
20. Dunsmoor, J.E., and Murphy, G.L. (2015). Categories, concepts, and conditioning: how humans generalize fear. Trends Cogn. Sci. *19*, 73–77. https://doi.org/10.1016/j.tics.2014.12.003.

21. Newell, F.N., McKenna, E., Seveso, M.A., Devine, I., Alahmad, F., Hirst, R.J., and O'Dowd, A. (2023). Multisensory perception constrains the formation of object categories: a review of evidence from sensory-driven and predictive processes on categorical decisions. Philos. Trans. R. Soc. Lond. B Biol. Sci. *378*, 20220342. https://doi.org/10.1098/rstb.2022.0342.

22. de Voogd, L.D., Fernández, G., and Hermans, E.J. (2016). Disentangling the roles of arousal and amygdala activation in emotional declarative memory. Soc. Cogn. Affect. Neurosci. *11*, 1471–1480. https://doi.org/10.1093/scan/nsw055.

23. Dunsmoor, J.E., Kragel, P.A., Martin, A., and LaBar, K.S. (2014). Aversive learning modulates cortical representations of object categories. Cereb. Cortex *24*, 2859–2872. https://doi.org/10.1093/cercor/bht138.

24. Biderman, N., Bakkour, A., and Shohamy, D. (2020). What Are Memories For? The Hippocampus Bridges Past Experience with Future Decisions. Trends Cogn. Sci. *24*, 542–556. https://doi.org/10.1016/j.tics.2020.04.004.

25. Wang, F., Schoenbaum, G., and Kahnt, T. (2020). Interactions between human orbitofrontal cortex and hippocampus support model-based inference. PLoS Biol. *18*, e3000578. https://doi.org/10.1371/journal.pbio.3000578.

26. Holmes, N.M., Wong, F.S., Bouchekioua, Y., and Westbrook, R.F. (2022). Not "either-or" but "which-when": A review of the evidence for integration in sensory preconditioning. Neurosci. Biobehav. Rev. *132*, 1197–1204. https://doi.org/10.1016/j.neubiorev.2021.10.032.

27. Sadacca, B.F., Wied, H.M., Lopatina, N., Saini, G.K., Nemirovsky, D., and Schoenbaum, G. (2018). Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. eLife *7*, e30373. https://doi.org/10.7554/eLife.30373.

28. Dunsmoor, J.E., White, A.J., and LaBar, K.S. (2011). Conceptual similarity promotes generalization of higher order fear learning. Learn. Mem. *18*, 156–160. https://doi.org/10.1101/lm.2016411.

29. Dunsmoor, J.E., and Murphy, G.L. (2014). Stimulus typicality determines how broadly fear is generalized. Psychol. Sci. *25*, 1816–1821. https://doi.org/10.1177/0956797614535401.

30. Hennings, A.C., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2021). Emotional learning retroactively enhances item memory but distorts source attribution. Learn. Mem. *28*, 178–186. https://doi.org/10.1101/lm.053371.120.

31. Fullana, M.A., Harrison, B.J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., and Radua, J. (2016). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. Mol. Psychiatry *21*, 500–508. https://doi.org/10.1038/mp.2015.88.

32. Hennings, A.C., McClay, M., Drew, M.R., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2022). Neural reinstatement reveals divided organization of fear and extinction memories in the human brain. Curr. Biol. *32*, 304–314.e5. https://doi.org/10.1016/j.cub.2021.11.004.

33. Keller, N.E., Hennings, A.C., Leiker, E.K., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2022). Rewarded Extinction Increases Amygdalar Connectivity and Stabilizes Long-Term Memory Traces in the vmPFC. J. Neurosci. *42*, 5717–5729. https://doi.org/10.1523/JNEUROSCI.0075-22.2022.

34. Reddan, M.C., Wager, T.D., and Schiller, D. (2018). Attenuating neural threat expression with imagination. Neuron *100*, 994–1005.e4. https://doi.org/10.1016/j.neuron.2018.10.047.

35. Ritchey, M., Wing, E.A., LaBar, K.S., and Cabeza, R. (2013). Neural Similarity Between Encoding and Retrieval is Related to Memory Via Hippocampal Interactions. Cereb. Cortex *23*, 2818–2828. https://doi.org/10.1093/cercor/bhs258.

36. Tompary, A., and Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. Neuron *96*, 228–241.e5. https://doi.org/10.1016/j.neuron.2017.09.005.

37. Hennings, A.C., McClay, M., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2020). Contextual reinstatement promotes extinction generalization in healthy adults but not PTSD. Neuropsychologia *147*, 107573. https://doi.org/10.1016/j.neuropsychologia.2020.107573.

38. Bower, G.H. (1992). How might emotions affect learning. In The Handbook of Emotion and Memory: Research and Theory, *3* (Erlbaum), p. 31.

39. Foa, E.B., and Kozak, M.J. (1986). Emotional processing of fear: Exposure to corrective information. Psychol. Bull. *99*, 20–35. https://doi.org/10.1037/0033-2909.99.1.20.

40. Lang, P.J. (1977). Imagery in therapy: an information processing analysis of fear. Behav. Ther. *8*, 862–886. https://doi.org/10.1016/S0005-7894(77)80157-3.

41. Charest, I., Allen, E., Wu, Y., Naselaris, T., and Kay, K. (2020). Precise identification of semantic representations in the human brain. J. Vision *20*, 539. https://doi.org/10.1167/jov.20.11.539.

42. Clarke, A. (2020). Dynamic activity patterns in the anterior temporal lobe represents object semantics. Cogn. Neurosci. *11*, 111–121. https://doi.org/10.1080/17588928.2020.1742678.

43. Ferko, K.M., Blumenthal, A., Martin, C.B., Proklova, D., Minos, A.N., Saksida, L.M., Bussey, T.J., Khan, A.R., and Köhler, S. (2022). Activity in perirhinal and entorhinal cortex predicts perceived visual similarities among category exemplars with highest precision. eLife *11*, e66884. https://doi.org/10.7554/eLife.66884.

44. Martin, C.B., Douglas, D., Newsome, R.N., Man, L.L., and Barense, M.D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. eLife *7*, e31873. https://doi.org/10.7554/eLife.31873.

45. Clewett, D., Dunsmoor, J., Bachman, S.L., Phelps, E.A., and Davachi, L. (2022). Survival of the salient: Aversive learning rescues otherwise forgettable memories via neural reactivation and post-encoding hippocampal connectivity. Neurobiol. Learn. Mem. *187*, 107572. https://doi.org/10.1016/j.nlm.2021.107572.

46. Fullana, M.A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., and Harrison, B.J. (2018). Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. Neurosci. Biobehav. Rev. *88*, 16–25. https://doi.org/10.1016/j.neubiorev.2018.03.002.

47. Richter, F.R., Chanales, A.J.H., and Kuhl, B.A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. NeuroImage *124*, 323–335. https://doi.org/10.1016/j.neuroimage.2015.08.051.

48. Shohamy, D., and Wagner, A.D. (2008). Integrating Memories in the Human Brain: Hippocampal–Midbrain Encoding of Overlapping Events. Neuron *60*, 378–389. https://doi.org/10.1016/j.neuron.2008.09.023.

49. Zeithamova, D., and Preston, A.R. (2017). Temporal Proximity Promotes Integration of Overlapping Events. J. Cogn. Neurosci. *29*, 1311–1323. https://doi.org/10.1162/jocn_a_01116.

50. de Voogd, L.D., Fernández, G., and Hermans, E.J. (2016). Awake reactivation of emotional memory traces through hippocampal–neocortical interactions. NeuroImage *134*, 563–572. https://doi.org/10.1016/j.neuroimage.2016.04.026.

51. Mack, M.L., and Preston, A.R. (2016). Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. NeuroImage *127*, 144–157. https://doi.org/10.1016/j.neuroimage.2015.12.015.

52. Dunsmoor, J.E., and Kroes, M.C.W. (2019). Episodic memory and Pavlovian conditioning: ships passing in the night. Curr. Opin. Behav. Sci. *26*, 32–39. https://doi.org/10.1016/j.cobeha.2018.09.019.

53. Gagnon, S.A., and Wagner, A.D. (2016). Acute stress and episodic memory retrieval: neurobiological mechanisms and behavioral consequences. Ann. N. Y. Acad. Sci. *1369*, 55–75. https://doi.org/10.1111/nyas.12996.

54. Kanwisher, N. (2001). Neural events and perceptual awareness. Cognition *79*, 89–113. https://doi.org/10.1016/S0010-0277(00)00125-6.

55. Turk-Browne, N.B. (2019). The hippocampus as a visual area organized by space and time: A spatiotemporal similarity hypothesis. Vision Res. *165*, 123–130. https://doi.org/10.1016/j.visres.2019.10.007.

56. Dunsmoor, J.E., Murty, V.P., Davachi, L., and Phelps, E.A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. Nature *520*, 345–348. https://doi.org/10.1038/nature14106.

57. Ballarini, F., Moncada, D., Martinez, M.C., Alen, N., and Viola, H. (2009). Behavioral tagging is a general mechanism of long-term memory formation. Proc. Natl. Acad. Sci. USA *106*, 14599–14604. https://doi.org/10.1073/pnas.0907078106.

58. de Carvalho Myskiw, J., Benetti, F., and Izquierdo, I. (2013). Behavioral tagging of extinction learning. Proc. Natl. Acad. Sci. USA *110*, 1071–1076. https://doi.org/10.1073/pnas.1220875110.

59. Dunsmoor, J.E., Murty, V.P., Clewett, D., Phelps, E.A., and Davachi, L. (2022). Tag and capture: how salient experiences target and rescue nearby events in memory. Trends Cogn. Sci. *26*, 782–795. https://doi.org/10.1016/j.tics.2022.06.009.

60. Laing, P.A.F., and Dunsmoor, J.E. (2023). Pattern separation of fear extinction memory. Learn. Mem. *30*, 110–115. https://doi.org/10.1101/lm.053760.123.

61. Busquets-Garcia, A., and Holmes, N.M. (2022). Editorial: Higher-Order Conditioning: Beyond Classical Conditioning. Front. Behav. Neurosci. *16*, 928769. https://doi.org/10.3389/fnbeh.2022.928769.

62. Wang, J., Smeets, T., Otgaar, H., and Howe, M.L. (2021). Manipulating Memory Associations Minimizes Avoidance Behavior. Front. Behav. Neurosci. *15*, 746161. https://doi.org/10.3389/fnbeh.2021.746161.

63. Boddez, Y., Baeyens, F., Luyten, L., Vansteenwegen, D., Hermans, D., and Beckers, T. (2013). Rating data are underrated: Validity of US expectancy in human fear conditioning. J. Behav. Ther. Exp. Psychiatry *44*, 201–206. https://doi.org/10.1016/j.jbtep.2012.08.003.

64. Rachman, S., and Hodgson, R. (1974). I. Synchrony and desynchrony in fear and avoidance. Behav. Res. Ther. *12*, 311–318. https://doi.org/10.1016/0005-7967(74)90005-9.

65. LeDoux, J.E., and Pine, D.S. (2016). Using Neuroscience to Help Understand Fear and Anxiety: A Two-System Framework. Am. J. Psychiatry *173*, 1083–1093. https://doi.org/10.1176/appi.ajp.2016.16030353.

66. Lissek, S., Pine, D.S., and Grillon, C. (2006). The strong situation: A potential impediment to studying the psychobiology and pharmacology of anxiety disorders. Biol. Psychol. *72*, 265–270.

67. Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. Nature *603*, 654–660. https://doi.org/10.1038/s41586-022-04492-9.

68. Murphy, G. (2004). The Big Book of Concepts (MIT Press).

69. Lissek, S., and van Meurs, B. (2015). Learning models of PTSD: Theoretical accounts and psychobiological evidence. Int. J. Psychophysiol. *98*, 594–605. https://doi.org/10.1016/j.ijpsycho.2014.11.006.

70. Suarez-Jimenez, B., Albajes-Eizagirre, A., Lazarov, A., Zhu, X., Harrison, B.J., Radua, J., Neria, Y., and Fullana, M.A. (2020). Neural signatures of conditioning, extinction learning, and extinction recall in posttraumatic stress disorder: a meta-analysis of functional magnetic resonance imaging studies. Psychol. Med. *50*, 1442–1451. https://doi.org/10.1017/S0033291719001387.

71. Dunsmoor, J.E., Cisler, J.M., Fonzo, G.A., Creech, S.K., and Nemeroff, C.B. (2022). Laboratory models of post-traumatic stress disorder: The elusive bridge to translation. Neuron *110*, 1754–1776. https://doi.org/10.1016/j.neuron.2022.03.001.

72. Keane, T.M., Zimering, R.T., Caddell, J.M., et al. (1985). A behavioral formulation of posttraumatic stress disorder in Vietnam veterans. Behav. Therapist *8*, 9–12.

73. Tolin, D.F., Gilliam, C., Wootton, B.M., Bowe, W., Bragdon, L.B., Davis, E., Hannan, S.E., Steinman, S.A., Worden, B., and Hallion, L.S. (2018). Psychometric Properties of a Structured Diagnostic Interview for DSM-5 Anxiety, Mood, and Obsessive-Compulsive and Related Disorders. Assessment *25*, 3–13. https://doi.org/10.1177/1073191116638410.

74. Weathers, F.W., Bovin, M.J., Lee, D.J., Sloan, D.M., Schnurr, P.P., Kaloupek, D.G., Keane, T.M., and Marx, B.P. (2018). The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): Development and Initial Psychometric Evaluation in Military Veterans. Psychol. Assess. *30*, 383–395. https://doi.org/10.1037/pas0000486.

75. Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J.K. (2019). PsychoPy2: Experiments in behavior made easy. Behav. Res. Methods *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y.

76. Bach, D.R., Sporrer, J., Abend, R., Beckers, T., Dunsmoor, J.E., Fullana, M.A., Gamer, M., Gee, D.G., Hamm, A., Hartley, C.A., et al. (2023). Consensus design of a calibration experiment for human fear conditioning. Neurosci. Biobehav. Rev. *148*, 105146. https://doi.org/10.1016/j.neubiorev.2023.105146.

77. Hennings, A.C., Cooper, S.E., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2022). Pattern analysis of neuroimaging data reveals novel insights on threat learning and extinction in humans. Neurosci. Biobehav. Rev. *142*, 104918. https://doi.org/10.1016/j.neubiorev.2022.104918.

78. Hennings, A.C., Bibb, S.A., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2021). Thought suppression inhibits the generalization of fear extinction. Behav. Brain Res. *398*, 112931. https://doi.org/10.1016/j.bbr.2020.112931.

79. Cooper, S.E., Dunsmoor, J.E., Koval, K.A., Pino, E.R., and Steinman, S.A. (2023). Test–retest reliability of human threat conditioning and generalization across a 1-to-2-week interval. Psychophysiology *60*, e14242. https://doi.org/10.1111/psyp.14242.

80. Green, S.R., Kragel, P.A., Fecteau, M.E., and LaBar, K.S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. Int. J. Psychophysiol. *1*, 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015.

81. Lykken, D.T., and Venables, P.H. (1971). Direct Measurement of Skin Conductance: A Proposal for Standardization. Psychophysiology *8*, 656–672. https://doi.org/10.1111/j.1469-8986.1971.tb00501.x.

82. Lonsdorf, T.B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., Jentsch, V.L., Meir Drexler, S., Mertens, G., Richter, J., et al. (2019). Navigating the garden of forking paths for data exclusions in fear conditioning research. eLife *8*, e52465. https://doi.org/10.7554/eLife.52465.

83. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4.

84. Esteban, O., Blair, R., Markiewicz, C.J., Berleant, S.L., Moodie, C., Ma, F., and Isik, A.I. (2018). FMRIPrep: a robust preprocessing pipeline for functional MRI. Zenodo. https://doi.org/10.5281/zenodo.852659.

85. Gorgolewski, K.J., Esteban, O., Markiewicz, C.J., Ziegler, E., Ellis, D.G., Notter, M.P., and Jarecka, D. (2018). nipy/nipype: 1.8.3. Zenodo. https://doi.org/10.5281/zenodo.596855.

86. Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., and Ghosh, S.S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. Front. Neuroinform. *5*, 13. https://doi.org/10.3389/fninf.2011.00013.

87. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., and Gee, J.C. (2010). N4ITK: Improved N3 Bias Correction. IEEE Trans. Med. Imaging *29*, 1310–1320. https://doi.org/10.1109/tmi.2010.2046908.

88. Avants, B.B., Epstein, C.L., Grossman, M., and Gee, J.C. (2008). Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative

Brain. Med. Image Anal. *12*, 26–41. https://doi.org/10.1016/j.media.2007.06.004.

89. Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm. IEEE Trans. Med. Imaging *20*, 45–57. https://doi.org/10.1109/42.906424.
90. Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. NeuroImage *9*, 179–194. https://doi.org/10.1006/nimg.1998.0395.
91. Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., and Keshavan, A. (2017). Mindboggling Morphometry of Human Brains. PLoS Comput. Biol. *13*, e1005350. https://doi.org/10.1371/journal.pcbi.1005350.
92. Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., and Collins, D.L. (2009). Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood. NeuroImage *47* (*Suppl 1*), 102. https://doi.org/10.1016/s1053-8119(09)70884-5.
93. Cox, R.W., and Hyde, J.S. (1997). Software Tools for Analysis and Visualization of fMRI Data. NMR Biomed. *10*, 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5.
94. Greve, D.N., and Fischl, B. (2009). Accurate and Robust Brain Image Alignment Using Boundary-Based Registration. NeuroImage *48*, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.
95. Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage *17*, 825–841. https://doi.org/10.1006/nimg.2002.1132.
96. Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2014). Methods to Detect, Characterize, and Remove Motion Artifact in Resting State fMRI. Neuroimage *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048.
97. Behzadi, Y., Restom, K., Liau, J., and Liu, T.T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage *37*, 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.
98. Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052.
99. Lanczos, C. (1964). Evaluation of Noisy Data. J. Soc. Ind. Appl. Math. Ser. B Numer. Anal. *1*, 76–85. https://doi.org/10.1137/0701007.
100. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. *8*, 14. https://doi.org/10.3389/fninf.2014.00014.
101. Mumford, J.A., Turner, B.O., Ashby, F.G., and Poldrack, R.A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage *59*, 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076.
102. Mumford, J.A., Davis, T., and Poldrack, R.A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. NeuroImage *103*, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026.
103. Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. *29*, 162–173. https://doi.org/10.1006/cbmr.1996.0014.
104. Yu, T., Lang, S., Birbaumer, N., and Kotchoubey, B. (2014). Neural correlates of sensory preconditioning: A preliminary fMRI investigation. Hum. Brain Mapp. *35*, 1297–1304. https://doi.org/10.1002/hbm.22253.
105. Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N.J., Habel, U., Schneider, F., and Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anat. Embryol. (Berl) *210*, 343–352. https://doi.org/10.1007/s00429-005-0025-5.
106. Ritchey, M., Montchal, M.E., Yonelinas, A.P., and Ranganath, C. (2015). Delay-dependent contributions of medial temporal lobe regions to episodic memory retrieval. eLife *4*, e05025. https://doi.org/10.7554/eLife.05025.
107. Constantinescu, A.O., O'Reilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science *352*, 1464–1468. https://doi.org/10.1126/science.aaf0941.
108. Garvert, M.M., Dolan, R.J., and Behrens, T.E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. eLife *6*, e17086. https://doi.org/10.7554/eLife.17086.
109. Kim, H., Smolker, H.R., Smith, L.L., Banich, M.T., and Lewis-Peacock, J.A. (2020). Changes to information in working memory depend on distinct removal operations. Nat. Commun. *11*, 6239. https://doi.org/10.1038/s41467-020-20085-4.
110. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
111. Martin, A. (2007). The Representation of Object Concepts in the Brain. Annu. Rev. Psychol. *58*, 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143.
112. R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
113. Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., and Cox, R.W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage *73*, 176–190. https://doi.org/10.1016/j.neuroimage.2013.01.047.
114. Chen, G., Taylor, P.A., Stoddard, J., Cox, R.W., Bandettini, P.A., and Pessoa, L. (2022). Sources of Information Waste in Neuroimaging: Mishandling Structures, Thinking Dichotomously, and Over-Reducing Data. Aperture Neuro *2*, 1–22. https://doi.org/10.52294/2e179dbf-5e37-4338-a639-9ceb92b055ea.
115. Field, A.P., and Wilcox, R.R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. Behav. Res. Ther. *98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013.
116. Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. J. Stat. Softw. *75*, 1–24. https://doi.org/10.18637/jss.v075.i06.
117. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67*, 1–51. https://doi.org/10.18637/jss.v067.i01.
118. Barr, D.J., Levy, R., Scheepers, C., and Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001.
119. Lenth, R.V. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://cran.r-project.org/web/packages/emmeans/emmeans.pdf.
120. Luke, S.G. (2017). Evaluating significance in linear mixed-effects models in R. Behav. Res. Methods *49*, 1494–1502. https://doi.org/10.3758/s13428-016-0809-y.
121. Arnqvist, G. (2020). Mixed Models Offer No Freedom from Degrees of Freedom. Trends Ecol. Evol. *35*, 329–335. https://doi.org/10.1016/j.tree.2019.12.004.
122. Lüdecke, D., Ben-Shachar, M.S., Patil, I., and Makowski, D. (2020). Extracting, Computing and Exploring the Parameters of Statistical Models using R. J. Open Source Softw. *5*, 2445. https://doi.org/10.21105/joss.02445.
123. Ben-Shachar, M.S., Lüdecke, D., and Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. J. Open Source Softw. *5*, 2815. https://doi.org/10.21105/joss.02815.

124. Kassambara, A. (2020). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://cran.r-project.org/web/packages/rstatix/index.html.

125. Lüdecke, D., Waggoner, P., and Makowski, D. (2019). insight: A Unified Interface to Access Information from Model Objects in R. J. Open Source Softw. *4*, 1412. https://doi.org/10.21105/joss.01412.

126. Lüdecke, D., Makowski, D., Waggoner, P., and Patil, I. (2020). performance: Assessment of Regression Models Performance. https://cran.r-project.org/web/packages/performance/performance.pdf.

127. Patil, I., Makowski, D., Ben-Shachar, M.S., Wiernik, B.M., Bacher, E., and Lüdecke, D. (2022). datawizard: An R Package for Easy Data Preparation and Statistical Transformations. J. Open Source Softw. *7*, 4684. https://doi.org/10.21105/joss.04684.

128. Canty, A., and Ripley, B.D. (2022). boot: Bootstrap R (S-Plus) Functions. https://cran.r-project.org/web/packages/boot/boot.pdf.

129. Davison, A.C., and Hinkley, D.V. (1997). Bootstrap Methods and Their Applications (Cambridge University Press).

130. Rights, J.D., and Sterba, S.K. (2023). On the Common but Problematic Specification of Conflated Random Slopes in Multilevel Models. Multivariate Behav. Res. *58*, 1106–1133. https://doi.org/10.1080/00273171.2023.2174490.

131. Jiang, J., Wand, M.P., and Bhaskaran, A. (2022). Usable and Precise Asymptotics for Generalized Linear Mixed Model Analysis and Design. J. R. Stat. Soc. B *84*, 55–82. https://doi.org/10.1111/rssb.12473.

132. Li, P., and Redden, D.T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. BMC Med. Res. Methodol. *15*, 38. https://doi.org/10.1186/s12874-015-0026-x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Deidentified neuroimaging and behavioral data	This paper	NIMH Data Archive (https://nda.nih.gov/); ID: C3797
Software and algorithms		
Custom Python and R analysis code	This paper	OSF: https://osf.io/bpv97/; DOI https://doi.org/10.17605/OSF.IO/BPV97

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Joseph E. Dunsmoor (joseph.dunsmoor@austin.utexas.edu).

Materials availability

This study did not generate any unique reagents.

Data and code availability

- All de-identified neuroimaging and behavioral data have been deposited at the NIMH NDA and are publicly available as of the date of the publication. The data ID is listed in the key resources table.
- All custom Python and R code used for analysis has been deposited at OSF and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- The lead contact can provide any additional information required to reanalyze the data reported in this paper upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We recruited 37 participants (M_{age} = 21.5, SD_{age} = 3.07; 19 identified as women, one as nonbinary, 17 as men) from the local community to complete all measures. All participants completed standardized clinical interviews[73,74] with a trained clinical psychologist or technician and were determined to be free of any psychopathology, neurological disorder, or interfering medical conditions. Two participants did not return for the second testing session; therefore, analyses were conducted on a final sample of N = 35. All study procedures described herein received approval from the University of Texas at Austin Institutional Review Board (IRB #2020020157-MOD1). All participants provided written informed consent prior to participation.

METHOD DETAILS

Stimuli

For PSs, we used 180 non-repeating images of either animals (N = 90) or tools (N = 90) against a white background obtained from public online resources and used in prior studies from our group.[32,33] Each PS presentation was a different basic-level exemplar (e.g., there were not two different pictures of a dog at any point). We did not include threatening/phobia-related stimuli (e.g., spiders, knives). CSs were either an image of an orange square or a blue circle against a white background. Both shapes shared the same width, height, and luminance and were approximately the same size as category exemplar images. Stimulus presentation was controlled by Psychopy.[75]

The US was a 5-ms electrical shock, delivered to the left index and middle finger. Shock intensity was determined through a brief calibration sequence prior to the experiment, in which participants reached a level described as "highly annoying/unpleasant, but not painful" (5-6 on a 10-point scale) through a stepwise procedure.[76] The shock was controlled using the STMEPM-MRI stimulation system (BIOPAC Systems, Goleta, CA).

Task and procedures

The current task, based on similar human work and optimized for MVPA,[77] consisted of seven phases across two days (see Figure 1B). Day 1 consisted of a perceptual localizer, preconditioning, threat conditioning, and two immediate transfer tests. After shock and skin conductance response (SCR) electrodes were attached, participants completed the perceptual localizer (described below)

and then a preconditioning phase in which a PS stimulus was presented for 4-6s, followed immediately by a .5s CS presentation. Participants were instructed prior to the task that their goal was to learn which category is associated with each shape, and that they will need to remember this association. PS category (PS+ and PS-) and CS shape (CS+ and CS-) pairings were counterbalanced across participants. For each PS presentation during preconditioning (30 animals, 30 tools), participants indicated which shape they thought would next appear using a 4-item scale (''Definitely square'', ''Maybe square'', ''Maybe circle'', ''Definitely circle''). Next, participants received instructions that they were now at risk for shock and received a single reminder shock during a blank screen to ensure the perceived intensity had not changed. Participants then completed the threat conditioning phase. Only CSs were presented. One CS co-terminated with shock (CS+, 12 trials, 66% reinforcement) and one was never paired with shock (CS-, 12 trials); this was counterbalanced across participants. CSs were presented for 5-7s. Participants provided shock expectancy ratings for each CS, again using a 4-item scale (''Definitely shock'', ''maybe shock'', ''maybe no shock'', ''definitely no shock''). Threat conditioning was immediately followed by two transfer tests. In these tests, participants viewed novel PSs from each category (30 animals and 30 tools; 15 of each category in each test) for 4-6s and continued to provide shock expectancy ratings as they did during the threat conditioning phase. No US was ever administered following a PS. Between the two transfer tests there was a seamless presentation of two reinforced CS+ trials to remind participants of the CS-US association (i.e., ''booster trials'').[28] We refer to these transfer tests as immediate Transfer 1 and Transfer 2 throughout this report.

On Day 2, approximately 24-hours later, participants completed identical transfer tests as on Day 1 (including the two reminder CS-US trials in between the tests), referred to as 24-hour Transfer 3 and 4. Exemplars presented on Day 2 were also novel category members and not seen on Day 1. Intertrial intervals were jittered (5-9s) for all phases across both days.

Finally, participants completed a recognition memory test, in which they viewed all Day 1 category stimuli (120 in total) in addition to foil stimuli (30 novel tools, 30 novel animals). We explicitly informed participants they would only see Day 1 or new stimuli and never Day 2 stimuli. Image order was pseudo-randomized and balanced such an equal number of items from each category appeared in each third of the phase. Each image was presented for 3s and required a response on a 4-item scale (''definitely new'', ''maybe new'', ''maybe old'', ''definitely old''). Outside of the scanner, participants viewed each Day 1 stimulus and provided self-paced 1-7 typicality ratings for each (1 = not typical at all of its category; to 7 = very typical of its category).

Skin conductance response

SCRs were acquired from the hypothenar eminence of the left palmar surface using disposable pre-gelled snap electrodes connected to the MP-160 BIOPAC System (BIOPAC Systems). In line with previously described procedures,[78,79] an SCR was considered related to CS or PS presentation if the trough-to-peak deflection occurred 0.5–3s following stimulus onset, lasted between 0.5 and 5.0s, and was greater than 0.02 microsiemens (μS), with responses not fitting these criteria scored as a zero. We obtained SCR values using a custom MATLAB (The Mathworks Inc., Natick, MA) script that extracts SCRs for each trial using the above criteria.[80] Raw SCR scores were square root transformed prior to analyses to normalize the distribution.[81] We did not exclude participants based on performance-based rules.[82]

Functional MRI acquisition

Scanning was completed using a Siemens Vida 3T MRI scanner at the University of Texas at Austin with the support of the Biomedical Imaging Center (RRID:SCR_021898). We acquired functional data with a 64-channel head coil at 2.5mm isotropic resolution. Using the scanner software, we automatically oriented slices parallel to the anterior-posterior commissure. We measured BOLD using T2* EPI sequences (TR = 1000ms, TR = 86ms, FOV = 86 x 86, multiband acceleration factor = 6). Before each functional series, we collected T2* field maps with opposite encoding phase to assist with distortion correction. We also collected anatomical images to assist in image registration: a T1w anatomical image (MPRAGE, .8mm isotropic) was collected during each session, and one T2w anatomical image (.8mm isotropic) was collected on Day 2.

Image preprocessing

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.2.6[83,84] (RRID:SCR_016216), which is based on Nipype 1.7.0[85,86] (RRID:SCR_002502).

Anatomical data preprocessing

A total of 2 T1-weighted (T1w) images were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection,[87] distributed with ANTs 2.3.3[88] (RRID:SCR_004757). The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9,[89] RRID:SCR_002823). A T1w-reference map was computed after registration of 2 T1w images (after INU-correction) using mri_robust_template (FreeSurfer 6.0.1[1]). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1,[90] RRID:SCR_001847), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle[91] (RRID:SCR_002438). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c[92] [RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym]

Functional data preprocessing

For each of the 9 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A B0-nonuniformity map (or fieldmap) was estimated based on two (or more) echo-planar imaging (EPI) references with opposing phase-encoding directions, with 3dQwarp[93] (AFNI 20160207). Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration.[94] Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9[95]). BOLD runs were slice-time corrected to 0.003s (0.5 of slice acquisition range 0s-0.00539s) using 3dTshift from AFNI 20160207[93] (RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions[96]) and Jenkinson (relative root mean square displacement between affines[95]). FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by ref Power et al.[96]). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor[97]). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's aseg segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each.[98] Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels.[99] Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.6.2[100] (RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

Copyright waiver

The above boilerplate text was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts unchanged. It is released under the CC0 license.

General linear models and whole-brain analyses

GLMs were computed using the *nilearn* 0.9.2 package[100] in Python 3.8.1, with standard boxcar modeling of trials convolved with a Glover hemodynamic response function and a lag-1 autoregressive model to account for serial correlations. For standard GLM univariate analyses, we included separate regressors for each stimulus across all trials of a given phase (e.g., PS+ and PS- in a transfer phase) and applied smoothing with a Gaussian kernel (FWHM = 6mm). All images were masked to only include likely grey matter voxels (MNI152 template, thresholded at 20% grey matter probability) and values were converted to effect sizes (beta). We also generated Least Squares-Separate (LSS) style betaseries images for each phase,[101,102] in which we iteratively estimated activity for a given trial while including all other trials from the same category in the same phase as regressors of no interest. To facilitate MVPA, no spatial smoothing was applied to LSS betaseries images.

Whole-brain statistical analyses were conducted using AFNI 22.3.04[103] and a family-wise error approach. Two-tailed t-tests were conducted on each voxel within participant-level whole-brain maps using 3dttest++ (with -Clustsim option). Significant cluster size was determined via $k = 10{,}000$ random permutations of null t-test results, which were then run through 3dClustSim to identify cluster-thresholds for each p-value. Using these cluster-thresholds, 3dClusterize was used to identify clusters in group-level masks and

extract cluster statistics (significant at cluster $p < .05$, voxel $p < .001$, third-nearest neighbor clustering) for category-selective activity during a perceptual localizer (see Figure S2; Table S1) and CS+ > CS- activity during the threat conditioning task (see Figure S3; Table S2).

ROI selection and parameter estimate extraction

ROIs were chosen in accordance with *a priori* hypotheses based on extensive rodent aversive preconditioning work[13] and human neuroimaging of threat conditioning and episodic memory integration.[31,104] For smaller regions in which higher spatial specificity was required, bilateral BLA were defined using the Julich probabilistic atlas,[105] whereas bilateral PRC was extracted from a probabilistic map of manually segmented hippocampal and parahippocampus subregions.[106] Both ROIs were defined at the group level. All probabilistic maps except the BLA were thresholded at 50% or greater likelihood of a given voxel belonging to that anatomical region. The BLA mask used a more stringent threshold of 75% due to increased noise in this region and to increase the chance voxels from anatomically adjacent subregions were not included in this mask. For larger ROIs, including the mPFC, full amygdala, and hippocampus, we defined ROIs for each participant anatomically using the relevant FreeSurfer parcellations of the Desikan-Killiany atlas. Although not part of our *a priori* ROIs, the entorhinal cortex serves as the primary input/output hub for the hippocampus and subserves many aspects of conceptual representations.[107,108] Accordingly, for exploratory analyses we also created a bilateral entorhinal cortex ROI that was defined using the Julich probabilistic atlas.

To extract parameter estimates for each ROI, we masked subject-level contrast maps (see above for GLM details) using a given ROI mask and extracted the average of the effect size values within this mask. These subject-level averages were then used as univariate activity variables in further analyses.

Perceptual localizer

Participants viewed images from categories that included animals, tools, shapes, and phase-scrambled animals and tools. Each image was shown for 1s and separated by 1s. Participants were given an irrelevant task (identifying a single repeated image with a button press, i.e., perceptual N-back) to keep their attention during the localizer. Images were presented in two runs. Each run included four blocks that consisted of 8 images each. Each category was presented for two blocks each (total 16 images per category). There were 16s of rest between blocks. Each image was distinct from any other localizer images and from stimuli used during experimental phases. LSS betaseries for each localizer block were computed.

Multivariate pattern analysis

MVPA included representational similarity analyses (RSA) and multivariate decoding. For RSA, we applied the *nilearn* library and custom Python code to LSS betaseries for trial-by-trial data from each phase and each stimulus (weighted by mean univariate estimates from the same phase and stimulus type to reduce noise[32,37,109]) to create a representational similarity matrix for each *a priori* ROI. In these matrices, each cell is a Pearson-correlation between all pairs of PS+ and PS- images across all phases. We then Fisher z transformed matrices and extracted the mean value of the cells that corresponded to within-category similarity (e.g., all cells with PS+ to PS+ correlations). We also extracted the mean value for the cells that corresponded to the similarity of each transfer phase to either threat conditioning or preconditioning for each stimulus type (e.g., all cells with correlations between PS+s in generalization and CS+s in threat conditioning or in preconditioning).

Classification was conducted using the *scikit-learn* library[110] and custom Python scripts. In line with similar prior work[22,23] and knowledge on canonical object regions,[111] we focused classification on occipital and temporal regions with voxels that uniquely code for animals or tool object categories. These category-selective cortices were functionally identified at the group-level through whole-brain analyses (voxel-wise $p \leq .001$, cluster-corrected $p < .05$, see Table S1; Figure S2) of univariate animal>tool and tool>animal contrasts of perceptual localizer fMRI data. For each set of contrasts, a 5mm sphere was drawn around the peak intensity voxel to focus analyses on the most selective voxels and to ensure a similar number of features (voxels) were submitted to classification for each category (animal-cortex = 64 voxels, tool-cortex = 72 voxels). We then trained an L2 weighted logistic regression classifier (''liblinear'' solver) to decode category-specific activity (with animals, tools, and scrambled images submitted to classification) in localizer functional data in each of the two identified ROIs. We assessed classifier sensitivity by cross-validating performance from one localizer block with the other for each classifier (mean animal cortex ROC area under the curve [AUC] = 67.5%, SD = 10%; mean tool cortex ROC AUC 70.5%, SD = 13.2%). Three participants were removed from further analyses due to unreliable classification (AUC values < 0.5), leaving N = 32 for further decoding analyses.

QUANTIFICATION AND STATISTICAL ANALYSIS

All ROI and behavioral/physiological group-level statistical analyses were conducted using R 4.3.[112] We used robust linear mixed-effects regression for all analyses, which is consistent with current recommendations for task-based fMRI data with multiple repeated factors.[113–115] All models were specified using the robustlmm[116] and lme4[117] libraries, reducing bias in clustered (repeated-measures) data by minimizing the influence of extreme outlier observations.[115] All models contained, at minimum, within-subject fixed effect of stimulus (e.g., CS+/CS- or PS+/PS-) and between-subject fixed effect of counterbalance order. Random effect structure was confirmed through likelihood ratio tests of model fit and using a ''keep it maximal'' approach, per standard recommendations.[118] From this model, we tested differences in stimuli estimated marginal means (using the *emmeans*[119] library) within each phase (e.g.,

PS+ vs. PS- during each transfer test). Per recommendations, all models were fit with restricted maximum likelihood and tests used Kenward-Rogers degrees of freedom.[120,121] We provide standardized estimates (β), parametric 95% confidence intervals, and Wald test *t*-values (t_{wald}) and *p*-values for all tests unless otherwise noted.[122] Model diagnostics, tests, and parameter extraction were all conducted using functions from the *easystats, emmeans*, and *rstatix* R libraries.[119,122–126] Standardization of model parameters was completed through post-hoc refit method.[127] When used, bootstrapping was done via a standard workflow using the *boot* library.[128,129]

For recognition memory data, a generalized linear mixed-effects model (binomial/logistic) was used to analyze high-confidence recognition responses (0 = "sure new", 1 = "sure old"). To correct memory recognition scores for false alarms, we included within-subject trial-by-trial false alarms and between-subject mean false alarm rate as fixed-effects covariates in our models to disambiguate within- and between-subject false alarm effects.[130] Asymptotic *z*-tests ($z_{asymp.}$) were conducted for these models.[131] One participant reported falling asleep during recognition and was excluded from recognition analyses (N = 34). For Day 2 behavioral tests, we combined immediate Transfer 1 and Transfer 2 stimuli for all analyses. For behavioral/physiological tests involving PS data (i.e., data with trial-unique stimuli), we modeled data at the trial-level with trial random and fixed effects to account for additional observations and restrict our degrees-of-freedom.[121,132]