**Title:** Distinct hippocampal mechanisms support concept formation and updating

**Authors:** Michael L. Mack[a], Bradley C. Love[b,c*], Alison R. Preston[d,e,f*]

**Affiliations:**

[a]Department of Psychology, University of Toronto, Toronto, ON, CA

[b]Experimental Psychology, University College London, London, UK

[c]Alan Turing Institute, London, UK

[d]Department of Psychology, The University of Texas at Austin, Austin, TX, USA

[e]Center for Learning and Memory, The University of Texas at Austin, Austin, TX, USA

[f]Department of Neuroscience, The University of Texas at Austin, Austin, TX, USA

*Authors contributed equally

**Corresponding Author:**

Michael L. Mack
Department of Psychology
University of Toronto
100 St. George Street, 4th Floor
Toronto, Ontario M5S 3G3
michael.mack@utoronto.ca

**Keywords:** hippocampus, fMRI, category learning, computational modeling

## Abstract

Learning systems must constantly decide whether to create new representations or update existing ones. For example, a child learning that a bat is a mammal and not a bird would be best served by creating a new representation, whereas updating may be best when encountering a second similar bat. Characterizing the neural dynamics that underlie these complementary memory operations requires identifying the exact moments when each operation occurs. We address this challenge by interrogating fMRI brain activation with a computational learning model that predicts trial-by-trial when memories are created versus updated. We found distinct neural engagement in anterior hippocampus and ventral striatum for model-predicted memory create and update events during early learning. Notably, the degree of this effect in hippocampus, but not ventral striatum, significantly related to learning outcome. Hippocampus additionally showed distinct patterns of functional coactivation with ventromedial prefrontal cortex and angular gyrus during memory creation and premotor cortex during memory updating. These findings suggest that complementary memory functions, as formalized in computational learning models, underlie the rapid formation of novel conceptual knowledge, with the hippocampus and its interactions with frontoparietal circuits playing a crucial role in successful learning.

## Significance statement

How do we reconcile new experiences with existing knowledge? Prominent theories suggest that novel information is either captured by creating new memories or leveraged to update existing memories, yet empirical support of how these distinct memory operations unfold during learning is limited. Here, we combine computational modeling of human learning behaviour with functional neuroimaging to identify moments of memory formation and updating and characterize their neural signatures. We find that both hippocampus and ventral striatum are distinctly engaged when memories are created versus updated; however, it is only hippocampus activation that is associated with learning outcomes. Our findings motivate a key theoretical revision that positions hippocampus is a key player in building organized memories from the earliest moments of learning.

Learning often relies on integrating new experiences with existing knowledge to aggregate information across events and encode regularities in the environment. Not all experiences, however, match prior knowledge. Imagine a child learning for the first time about bats and discovering that, despite having wings and being able to fly, they are not birds but mammals. Classic learning theories posit that such moments are learned no differently than any other (e.g., all experiences are abstracted into a "mammal" prototype[1,2], or all experiences are encoded as new "mammal" exemplars[3]). Alternatively, qualitatively different memory operations may be triggered throughout learning based on the match between the current and prior experiences[4–6]: existing memories may be updated to generalize across related experiences or new memories may be created to distinctly capture novel current experiences. Such a learning mechanism may promote more precise conceptual discrimination; for example, the child can preserve their existing knowledge about mammals while distinctly encoding a detailed representation for bats allowing for more flexible future inferences. However, empirical evidence in support of such flexible learning has proven challenging, especially in humans, given the lack of empirical methods to precisely identify when qualitatively different memory creation and updating operations occur throughout learning. Here, we combine computational modeling with human neuroimaging to quantify, moment-by-moment, the distinct memory mechanisms that support successful concept learning.

Seminal findings identified multiple systems underlying learning[7–9]; whereas hippocampal engagement in the earliest moments of learning were thought critical for capturing initial episodic memories of stimuli, incremental learning processes in basal ganglia were key to forming stimulus-response associations that best reflected behaviour. However, this perspective has been refined and expanded with recent findings demonstrating the more flexible nature of the hippocampus in building structured memories[5,10–16]. Specifically, it is theorized that during learning, hippocampal comparator processes evaluate the overlap between new experiences and current knowledge to trigger either memory integration[17], in which existing memories are updated with new information, or memory differentiation[18], in which new memories are created to be representationally distinct from existing knowledge[15,19]. The transient nature of these hippocampal memory operations has been demonstrated in animal models: moment-by-moment fluctuations in slow and fast gamma oscillations between hippocampal subfields and entorhinal cortex[20–22] as well as distinct sharp-wave ripples in hippocampus[23,24] seem to reflect behaviourally relevant state changes in memory function.

In humans, evidence for such state changes in hippocampal operations is emerging; for example, mismatch between cued memories and current experience lead to heightened hippocampal activation[25] and distinct connectivity patterns[26], and hippocampal representations of associative pairs become markedly differentiated at the moment they have been learned[27]. Moreover, during new category learning, trial-by-trial shifts in decision certainty as predicted by computational learning models[28–30] is associated with

hippocampal activation. The key open question is whether or not learning-related engagement of the hippocampus reflects distinct memory create and update operations as they unfold throughout learning.

We address this important gap with SUSTAIN[4], a computational learning model that uniquely makes explicit predictions about *when* and *how* newly encountered information is encoded into memory. Although borne out of a tradition of cognitive theories, SUSTAIN's latent mechanisms for representation learning parallel theorized hippocampal memory operations[5,31], a conceptual convergence with increasing empirical support[28,32,33]. We also had participants (N=25) learn to classify different sets of visual objects each with three binary features across four learning tasks[34] (**Figure 1**) designed to engage both memory creation and updating. Specifically, the learning tasks were defined by category rules that required generalizing across experiences that share features, but also distinctly encoding other experiences that diverge from expectations.

SUSTAIN accounts for learning in these tasks with an adaptive clustering mechanism (**Figure 2A**) in which clusters (i.e., memory traces) of stimulus features and their association to a category label are created (e.g., a child's first experience with a bat) or updated (e.g., encountering a second bat) depending on the match between the model's existing memories and the current stimulus information. Generalization occurs through incremental updates of existing clusters, whereas differentiation arises via cluster creation. Memory creation in SUSTAIN is infrequent in stationary environments and tends to occur early in learning; as such we included four variants of the learning task to capture enough of these create trials with a balanced number of update trials. SUSTAIN model simulations of the current learning tasks thus provide a means for identifying when memory operations were deployed throughout learning to build internal memory representations that supported participants' behaviour. These model-based predictions also provide an opportunity to identify neural function associated with dynamic memory operations. Namely, brain regions involved in building adaptive memory should be uniquely engaged during SUSTAIN-predicted memory updating and creation.

To test this prediction, we recorded functional magnetic resonance imaging (fMRI) data while participants learned the four tasks, derived trial-by-trial memory operation predictions through participant-specific model simulations, and interrogated brain activation through the lens of these model predictions. We demonstrate that regions of hippocampus and ventral striatum reflect dynamic memory operations during early learning, but only hippocampal activation predicts learning outcomes.

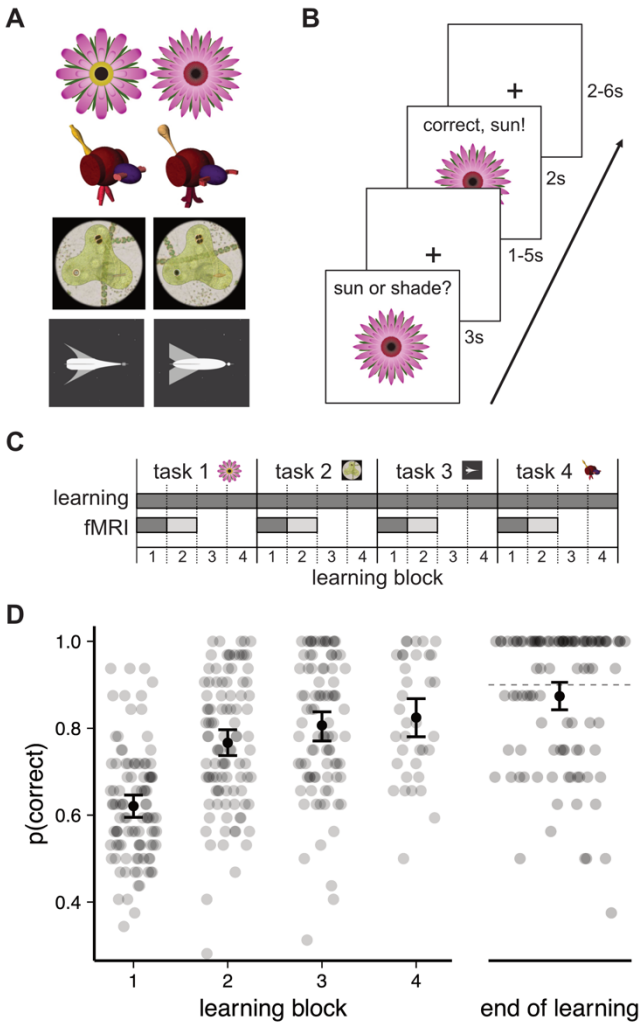


**Figure 1.** Learning tasks stimuli and trial schematic and learning performance. **A)** The four stimulus sets consisted of visual objects composed of three binary feature dimensions (Table 1). **B)** Learning trials followed typical feedback-based learning paradigm: a stimulus was shown for 3s during which participants could make a categorization response. After a variable delay (1-5s), feedback including the stimulus, whether or not the response was correct, and the correct category was presented for 2s followed by a variable delay (2-6s) before the next trial. **C)** Experimental sessions consisted of four learning tasks each with four blocks of feed-back learning trials. fMRI data was collected during the first two learning blocks. **D)** The probability of a correct response is plotted separately for each learning block and end of learning. Group averages and bootstrapped 95% confidence intervals are plotted in black; accuracy for individual participants in each of the four tasks are plotted in the smaller transparent points. A learning task would end after blocks 2 or 3 if the participant reached threshold of 90% correct on last 16 trials; this is reflected in fewer points plotted in blocks 3 and 4. End of learning accuracy (right) is depicted for each participant separately for each task and includes performance for tasks in which participants reached threshold in earlier learning blocks. The dotted line depicts the 90% learning threshold. Data includes N=25 participants each in 4 tasks.

## Results

*Neural indices of learning-guided memory formation*

Participants learned to varying degrees across the four tasks (**Figure 1D**) with end of learning performance splitting roughly equally into tasks in which learning was successful (>90% accuracy, 57/100 tasks) or not (<90% accuracy, 43/100 tasks). To characterize how different memory operations—creation and updating—guided successful decision making performance, model-based predictions of when distinct memory functions occurred throughout early stages of learning were derived from participant and task-specific simulations of SUSTAIN[4]. For each trial, the model output a label indicating whether a new memory was created or an existing knowledge cluster updated (**Figure 2A**). We then identified brain regions distinctly associated with each memory mechanism by conducting a voxel-wise mixed-effects regression on whole-brain trial-by-trial neural

activation, as estimated with the Least-Squares Separate approach[35]. Specifically, at each voxel we estimated how computational model-derived memory create versus update trials were associated with changes in neural activation.

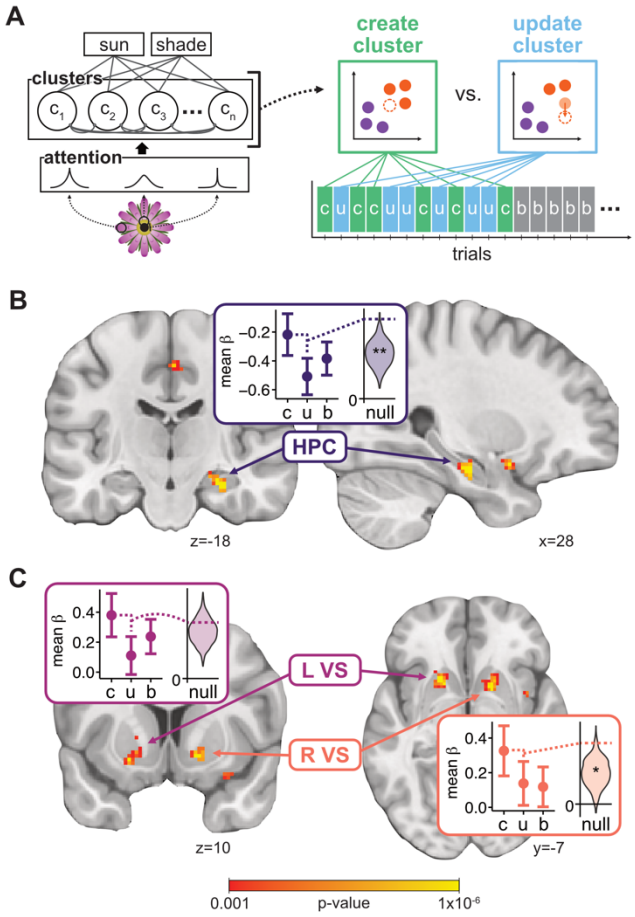


**Figure 2.** Illustration of the SUSTAIN computational model and model-based fMRI analyses of memory formation functions. **A)** SUSTAIN formalizes learning as the interaction between feature-based selective attention and memory representations formed via a clustering mechanism. On each learning trial a new cluster is created (green *c*) or an existing cluster is updated (blue *u*) depending on the match of the current stimulus to existing clusters. As learning progresses clusters become effectively fixed (grey *b*). **B)** A region in right hippocampus (HPC, peak [30, -18, -19]) showed higher activation for create relative to update trials. Inset plot depicts mean β estimates at region peak for create (c), update (u), and baseline (b) trials. Error bars represents 95% confidence intervals. Violin plots depict shuffled null distribution of create-update effect (*z* statistic) with observed effect. **C)** Two regions showing greater create than update engagement were localized to left and right ventral striatum (L VS, peak [-20, 14, -8]; R VS, peak [8, 20, -4]). Clusters were defined by a voxel wise threshold of p=0.001 and cluster-extent threshold of p=0.05 (>42 voxels). N=25. **p<0.01, *p<0.05

Notably, an anterior region of the right hippocampus (HPC, peak MNI coordinate [30, -18, -19], 57 voxels, voxel-wise threshold p = 0.001, cluster-extent threshold p = 0.05) and bilateral regions of ventral striatum (left: L VS, peak [-20, 14, -8], 188 voxels; right: R VS, peak [8, 20, -4], 225 voxels) demonstrated distinct neural engagement for create versus update trials (**Figure 2B/C**); all three regions showed more engagement when new knowledge clusters were created relative to when clusters were updated. Importantly, the effects in both anterior hippocampus (p = 0.005) and right ventral striatum (p = 0.022), but not left ventral striatum (p = 0.232), were robust to within-participant random permutation tests (**Figure 2B/C**, violin plot insets). Additional regions across the brain similarly tracked when new memories were created and existing memories updated, including precuneus, premotor cortex, and superior parietal cortex (**Table 2**). Importantly, these neural signatures of distinct memory operations were independent of trial outcome

or number and were evident across all tasks, each with their own stimulus sets; thus, ruling out explanations based on error-related and stimulus-specific effects. Rather, these findings demonstrate that fundamentally different memory operations, as formalized in computational predictions of SUSTAIN, are associated with unique neural signatures in key learning regions like hippocampus and striatum.

*Hippocampal engagement to memory operations predicts accurate categorization decisions*

Given their theorized role in learning[5,32,36–38], we next assessed that degree that the distinct memory operations supported by anterior hippocampus and ventral striatum were associated with categorization accuracy. We reasoned that the more a participant showed distinct engagement to create versus update trials early in learning, the better their learning outcome. In line with this prediction, we found that for tasks in which participants were highly accurate by the end of learning (>90% accuracy), anterior hippocampus activation during early learning exhibited a larger difference in memory create versus update trials (**Figure 3A**; $\beta$=0.336, CI=[0.056, 0.616], p=0.019). No such effect was found in either ventral striatum cluster (L VS: $\beta$=-0.113, CI=[-0.479, 0.254], p=0.547; R VS: $\beta$=-0.206, CI=[-0.499, 0.088], p=0.169). In a complementary analysis, we performed a logistic regression to predict whether the create versus update activation difference early in learning predicted accurate categorization decisions at the end of learning (**Figure 3B**). Anterior hippocampus activation predicted learning outcome (log odds=2.752, CI=[1.225, 6.181], p=0.014), while ventral striatum did not (L VS: log odds=0.742, CI=[0.429, 1.282], p=0.285; R VS: log odds=0.663, CI=[0.333-1.322], p=0.243). These findings point to a specific role for hippocampus in mediating when individuals exploit existing knowledge and when they create new knowledge. The more distinctly anterior hippocampus was engaged for memory create versus update events in the initial stages of learning, as predicted by SUSTAIN, the more likely that participants reached high categorization accuracy at the end of learning.

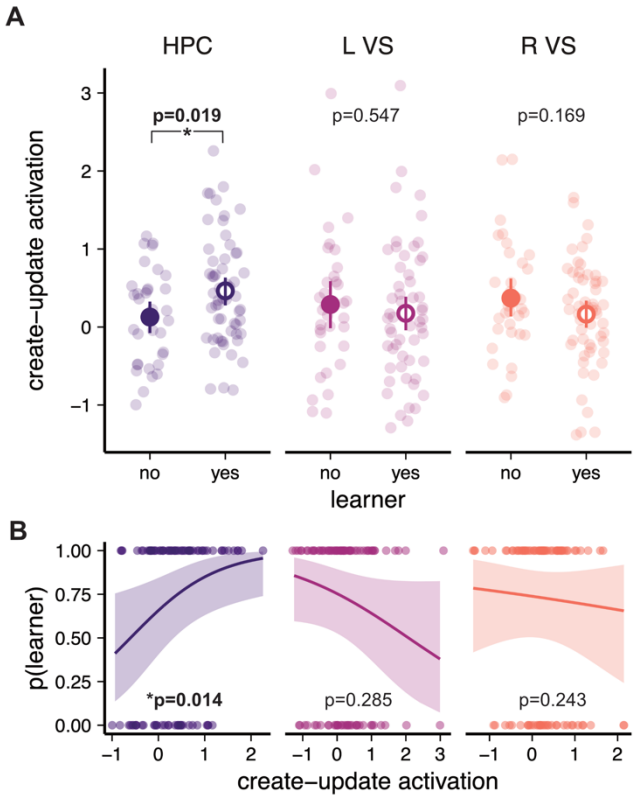


**Figure 3.** Neural indices of the memory computations that lead to successful learning outcomes. **A)** Activation differences that track memory creation and updating are plotted separately for tasks that reached learning criterion and those that did not (learner yes versus no) for hippocampus (HPC) and ventral striatum (VS). Group averages (larger circles) and bootstrapped 95% confidence intervals (dark lines) are depicted along with individual points for each completed task (lighter circles). **B)** Results of a logistic regression analysis assessing the relationship between whether activation reflected memory creation or updating and the likelihood that a task was successfully learned ("p(learner)") are depicted with estimated marginal means of the statistical model (dark lines and confidence bands) and the individual task data (lighter circles). Data includes N=25 participants each in 4 tasks.

*Networks supporting memory creation and updating*

With evidence that anterior hippocampus activation is linked to qualitatively different memory operations and that the degree of this engagement distinction leads to successful learning outcomes, we reasoned that distinct cortical networks may work in concert with anterior hippocampus to support memory creation versus updating early in learning. Indeed, memory creation is triggered by surprising events and results in a detailed encoding of the current experience into conceptual knowledge that influences subsequent category decisions. As such, hippocampal memory create mechanisms may be supported by frontoparietal regions, including lateral parietal cortex areas sensitive to novelty[39,40] and detailed memory encoding[41] and medial prefrontal cortex (mPFC) regions that encode goal-relevant memory models[29,30,33,42–47].

To investigate these possibilities, we performed an interregional functional correlation analysis[48] to identify brain regions that were distinctly coactive with anterior hippocampus during memory create versus update trials. A linear mixed effects regression analysis that related the functional time series of the right anterior hippocampus cluster to the rest of the brain as a function of which memory computation was predicted by SUSTAIN on a given trial revealed two distinct sets of regions (**Figure 4**). Subgenual vmPFC (peak [-8, 24, -10], 78 voxels) and right angular gyrus (peak [48, -42, 22], 47 voxels) were

significantly more correlated with anterior hippocampus during memory create versus update trials. In contrast, three regions in close proximity within premotor cortex (peak [-16, -22, 76], 161 voxels; peak [12 -18, 70], 160 voxels; peak [-12, -8, 78], 57 voxels) exhibited greater correlation with anterior hippocampus during memory update versus create trials. Importantly, these interregional correlation effects were observed independent of response accuracy and trial number. Moreover, none of these regions showed distinct overall engagement to memory formation functions. Thus, that trial-by-trial neural dynamics in anterior hippocampus are distinctly reflected in these brain regions during specific memory formation events supports the proposal that hippocampal learning occurs in coordination with key frontoparietal networks[29,40,46,49–52].

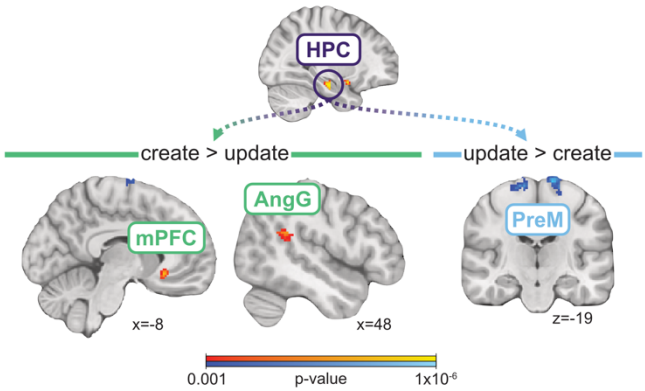


**Figure 4.** Interregional functional correlation with HPC. Regions of mPFC and angular gyrus showed greater functional coactivation with HPC during memory create relative to update trials. Regions of premotor cortex exhibited greater correlation with HPC during update relative to create trials. Clusters were defined by a voxel wise threshold of p=0.001 and cluster-extent threshold of p=0.05, which corresponded with a cluster extent of 42 voxels. N=25.

## Discussion

Here, our neurocomputational methods isolate the moments in time that qualitatively distinct memory operations lead to successful learning outcomes. We show that when anterior hippocampus is engaged during moments of new memory creation, as opposed to existing memory updating, participants are more successful at categorization decisions. In contrast, although the engagement of ventral striatum tracks the qualitative shift between memory creation and updating, such activation does not predict learning outcomes. Together, these findings suggests that anterior hippocampus, together with vmPFC and angular gyrus, plays a key role in forming distinct knowledge clusters that discriminate members of different concepts from one another, supporting the rapid acquisition of category knowledge and accurate decision making[5,13,16].

Classical views posit distinct memory systems involved in concept learning: whereas hippocampal-based declarative memory system is key for encoding initial episodic memories of concept exemplars and explicit knowledge of concepts, a procedural memory system mediated by the basal ganglia underlies implicit learning in which automatic associations between stimuli and responses (i.e., concept labels) are

incrementally learned over repeated experiences[8,9,53]. Our findings suggest a major shift from this classical view of learning. Although we find that bilateral ventral striatum engagement is associated with model-predicted shifts in memory operations during learning, this activation was not associated with individual differences in learning outcomes. The striatal effect may reflect initial learning of stimulus-response associations, as suggested by the classical view. However, the category structure in the current learning tasks were defined by complex multidimensional associations between stimuli and the correct category response; the more rigid striatal learning system may be ill-equipped for such learning. The hippocampus, on the other hand, offers a flexible learning system capable of building multidimensional category representations[16,29,33,54]. Indeed, that hippocampal activation to create and update memory trials early in learning was associated with learning outcome offers compelling for such an account.

By focusing on the initial moments of learning as characterized by model-based predictions of memory operations, our findings significantly extend recent work establishing hippocampus as a key player in novel concept learning. Although it has been demonstrated in both humans[29,33,55,56] and model[16,57] that hippocampal activation patterns reflect the latent structure of multidimensional category spaces, these results largely depend on neural coding at the end of learning. Recent work has found that measures of category evidence throughout learning relate to trial-by-trial activation in hippocampus[28,30], yet the latent mechanisms underlying these effects have remained an open question. Here, we establish that anterior hippocampal engagement during the initial moments of novel concept learning is mediated by shifts between qualitatively different memory create and update operations. More broadly, our findings are consistent with recent human neuroimaging evidence of novelty signaling[25,26,58–60] and rapid remapping in the hippocampus during spatial navigation[61–63], episodic memory[60], associative memory[27], and event segmentation[64,65]. The current results provide unique support for a similar rapid encoding mechanism at play during new category learning in which the engagement of different memory operations is reflected in distinct hippocampal states[25,26]. Coordination of these states are key to building adaptive knowledge structures that code regularities and discriminate novel experience to promote overall better learning outcomes[4,5,16].

Importantly, our findings were possible only by linking learning behaviour, as revealed in the initial moments and final outcome of learning, to neural activation with a computational model. Leveraging SUSTAIN to simulate each participant's behaviour in the learning tasks allowed us to identify key moments in learning associated with the model's memory creation and updating functions. Notably, this participant- and task-specific mapping between behaviour, model predictions, and neural measure was essential—greater hippocampal engagement during create versus update trials effect was significantly stronger than permutation analyses that shuffled the mapping between tasks within participants (see **Figures 2 & 5**). Two additional aspects of our design and analysis

strengthen the results: a) the memory modification effect generalized across the four different learning tasks participants performed, arguing against stimulus-specific factors; and b) most participants successfully learned some but not all tasks (19 participants had a mix, 5 learned all, and 1 learned none), allowing for the stronger test of within-participant differences in neural signatures associated with learning outcome.

That hippocampal engagement during learning reflected the latent dynamics of memory operations in SUSTAIN[4] provides a direct mechanistic interpretation. SUSTAIN proposes that during learning, the degree of match between the current experience and stored knowledge dictates the type of response a learner will make. But importantly, unlike other successful categorization models that depend on fixed representational formats[1–3], SUSTAIN uniquely posits that this degree of match also drives how stored knowledge is modified to accommodate the current experience—a high match may signal an update to an existing memory trace, but a low match may lead to the creation of a new memory trace that captures the current experience. Learning to precisely discriminate any given categorization structure requires an optimal number of clusters that generalize to new experiences to varying degrees. SUSTAIN's memory create and update memory operations, coupled with selective attention to stimulus dimensions, allows for building concept knowledge that reflects regularities while also supporting accurate discrimination given the learning goal. The current findings are also consistent with and extend mounting evidence that draws theoretically relevant links between the formal mechanisms of SUSTAIN and neural activation and representations in humans[5,28,31,33,44,66] and animals[67,68]

The hippocampus, of course, does not act alone and its interactions with cortical structures during critical learning moments are reflected in our findings. Hippocampal interactions with prefrontal cortex are associated with successful formation of structured memory networks[45,69] that support complex associative learning[17,46], schemas[61], and categories[29,30,33,70]. Motivated by these observations, we predicted that during initial learning, experiences that deviated from prior knowledge enough to warrant creating a new memory trace would potentially require coordination between hippocampus and PFC. Indeed, in line with this prediction, we observed heightened HPC-mPFC functional coactivation for create relative to update trials. This result supports the idea that hippocampal signals of mismatch between current and prior experiences may influence schema abstraction in mPFC[71] to highlight information most diagnostic for current task goals[29,44].

We also found that ventral posterior parietal cortex showed higher coactivation with hippocampus during learning trials associated with memory creation. Although the role of parietal cortex in memory behaviour[40,52,72] and representation[41,73] at retrieval is well established, ventral posterior regions of parietal cortex including angular gyrus seem to play an important role during encoding. Specifically, ventral posterior parietal cortex is a

key node in the ventral attention network associated with reorienting attention to unexpected but behaviourally relevant information during memory encoding[39,74,75]. Furthermore, angular gyrus engagement during encoding supports the integration of event elements and more vivid subsequent memory[76]. These encoding-based functions would distinctly support learning in the current paradigm—orienting to new or unexpected feature combinations is key in triggering the creation of new memories and encoding these experiences would be strengthened by rapidly encoding the features into a distinct memory trace. Our observation that ventral posterior parietal cortex reflects hippocampal engagement during memory creation provides unique evidence that these two regions may coordinate rapid learning of information that diverges from prior experiences.

We additionally observed that areas of premotor cortex were functionally more coactive with hippocampus during memory updating than creation. Premotor cortex is theorized to play a key role in COVIS, a prominent multiple systems account of learning[77,78]. Specifically, COVIS includes an implicit procedural learning circuit that connects striatum to premotor cortex within which direct stimulus-response associations are thought to be established and reinforced over learning. A theoretical extension of this model posits an additional more direct link between sensory cortex and premotor cortex to account for automatic categorization decisions[79]. That we found coactivation between premotor cortex and hippocampus rather than striatum runs counter to these theories. However, place coding in hippocampus[80–82], in which behaviourally relevant spatial sequences are stored and activated to support action plans during navigation, suggests the hippocampus is responsible for encoding motor alongside sensory information. The functional coactivation we observed between hippocampus and premotor cortex may reflect the activation of stronger associative memories between the current stimulus and motor response, a situation more likely to occur when existing memories are updated. These results motivate future lines of inquiry into how multiple learning systems may coordinate during early learning.

In summary, our findings propose a key revision to long-held theories of concept learning[8,9,53] such that hippocampus is a key player in building organized memories from the earliest moments of learning. Although we found unique signatures of complementary memory formation functions during early learning in hippocampus and striatum, notably, only the hippocampal signature of memory modification was associated with learning outcome. These results are consistent with neurobiological theories of hippocampal function[13,14,16] and formal cognitive models of learning[4], and provide novel empirical support for the theoretical convergence of these two perspectives[5,31]. By leveraging trial-specific model predictions of latent memory operations, we identified theoretically meaningful moments during learning and characterized the neural mechanisms that support the formation of flexible concept knowledge.

## Acknowledgements

The authors thank Meg Schlichting, Katherine Duncan, and Dasa Zeithamova for helpful discussions in preparation of this manuscript. This work was supported by the National Institute of Mental Health (F32-MH100904 to MLM, R01-MH100121 to ARP), Natural Sciences and Engineering Research Council (Discovery Grant RGPIN-2017-06753 to MLM), Canada Foundation for Innovation and Ontario Research Fund (36601 to MLM), Economic and Social Research Council (ES/W007347/1 to BCL), and a Royal Society Wolfson Fellowship (18302 to BCL).

## Methods

*Participants*

Twenty-five volunteers (13 females, mean age 21.6 years old, ranging from 18 to 29 years) participated in the experiment. All subjects were right-handed, had normal or corrected-to-normal vision, and were compensated $75 for participating.

*Stimuli*

Four stimulus image sets were used in the experiment (**Figure 1A**). Each stimulus set included eight images consisting of all combinations of three binary stimulus features (flower: outer petal shape, inner petal shape, centre colour; fribbles: tail, ears, legs; amoeba: cell structure features in the 3 cell arms; spaceships: wings, body, antennae). There were two versions of each (e.g., flower had pointy or round outer petals). Following procedures from previous work using the amoeba stimuli[83], the eight amoeba stimuli were randomly presented on different backgrounds throughout learning. All feature versions are depicted in Figure 1A. All stimulus images used in the experiment are available on OSF (https://osf.io/jk79v/).

*Procedures for learning tasks*

After an initial screening and consent in accordance with the University of Texas Institutional Review Board, participants were instructed on the classification learning problems. Participants then performed the problems in the MRI scanner by viewing visual stimuli back projected onto a screen through a mirror attached onto the head coil. Foam pads were used to minimize head motion. Stimulus presentation and timing was performed using custom scripts written in Matlab (Mathworks) and Psychtoolbox (www.psychtoolbox.org) on an Apple Mac Pro computer running OS X 10.7.

Participants were instructed to learn to classify the stimuli based on the combination of the features using the feedback displayed on each trial. As part of the initial instructions, participants were made aware of the three features and the two different values of each feature across all four stimulus sets. Before beginning each learning task, additional instructions that described the cover story for the current task and which buttons to press

for the two stimulus categories were presented to the participants. One example of this instruction text is as follows: “Each flower prefers either Sun or Shade to grow best. The environment that each flower prefers depends on one or more of its features. On each trial, you will be shown a flower and you will make a response as to that flower’s preferred environment. Press the ‘1’ button under your index finger for Sun or the ‘2’ button under your middle finger for Shade.” After the instruction screen, the two fMRI scanning runs (described below) for that problem commenced, with no further problem instructions. After the two scanning runs for a problem finished, the participants continued the learning task until criterion was reached or all trials for a task were completed. The next learning task began with the corresponding cover story description. Importantly, the rules that defined the learnings tasks were not included in any of the instructions; rather, participants had to learn these rules through trial and error.

The category structures underlying the learning tasks correspond with Shepard et al.‘s problem types 3, 4, and 5[34] (**Figure 1A, Table 1**). These structures were selected because optimally learning them requires attending to all three feature dimensions and and both generalizing across and distinctly representing specific stimuli[4,34]. In other words, memory creation and updating are key processes for learning these tasks.

| | feature attribute | | | category structure | | |
|---|---|---|---|---|---|---|
| stimulus | 1 | 2 | 3 | type 3 | type 4 | type 5 |
| 1 | 0 | 0 | 0 | B | B | B |
| 2 | 0 | 0 | 1 | B | B | B |
| 3 | 0 | 1 | 0 | B | B | B |
| 4 | 0 | 1 | 1 | A | A | A |
| 5 | 1 | 0 | 0 | A | B | A |
| 6 | 1 | 0 | 1 | B | A | A |
| 7 | 1 | 1 | 0 | A | A | A |
| 8 | 1 | 1 | 1 | A | A | B |

**Table 1:** Stimulus features and class associations for the three learning problems. Each of the eight stimuli are represented by the binary values of the three feature attributes. The stimuli are assigned to different classes (A or B) across the three category structures according to rules that depend on a combination of feature attributes.

The binary values of the feature attributes along with the class association for the three category structures are depicted in Table 1. The stimulus features were randomly mapped onto the attributes for each participant and task. For each participant, the four stimulus

sets were randomly paired with the three category structures with one structure repeated. The order of the tasks was randomly shuffled across participants.

The learning tasks consisted of learning trials (**Figure 1B**) during which a stimulus image was presented for 3s. During stimulus presentation, participants were instructed to respond to the stimulus' category by pressing one of two buttons on an fMRI-compatible button box. Stimulus images subtended 7.3° × 7.3° of visual space. The stimulus presentation period was followed by a 1-5s fixation. A feedback screen consisting of the stimulus image, text of whether the response was correct or incorrect, and the correct category was shown for 2s followed by a 2-6s fixation. The timing of the stimulus and feedback phases of the learning trials was jittered to optimize general linear modeling estimation of the fMRI data. Within one functional run, each of the eight stimulus images for a given task was presented in four learning trials. The order of the learning trials was pseudo randomized in blocks of 16 trials such that the eight stimuli were each presented twice. One functional run was 388s in duration. Each of the learning tasks included two functional runs corresponding to the first 64 trials of the task. After these two functional runs, learning trials continued for two more blocks of 32 trials without fMRI scanning and with no fixation periods after stimulus presentation or feedback. If participants reached 90% accuracy on the final 16 trials of learning blocks 2, 3, or 4, the task ended. Otherwise, participants completed 144 trials for a task. The entire experiment lasted approximately 75 minutes.

*Learning performance*

Performance during the learning tasks was characterized within each learning block as the average accuracy across the block's 32 trials (**Figure 1D**). During data collection, if a participant reached or surpassed 90% accuracy on the final 16 trials of blocks 2, 3, or 4, that task was ended. That is, some participants learned quickly and finished tasks early. As such, end of learning performance was characterized by average accuracy on the final 16 trials that the participants completed (**Figure 1D**). Across all tasks and participants, the average end of learning accuracy was 87.4% (median: 93.75%, SD: 15.8%, range: 37.5-100%) with a majority reaching the 90% threshold (57/100 completed tasks). Given that end of learning performance was highly skewed, we classified task performance by whether or not the 90% accuracy threshold was reached. This "learner" label was subsequently leveraged in relating neural indices of memory formation to end of learning performance.

*Computational learning model*

Participant behavior was modeled with an established mathematical learning model, SUSTAIN[4]. SUSTAIN is a network-based learning model that classifies incoming stimuli by comparing them to memory-based knowledge representations of previously experienced stimuli (**Figure 2A**). Sensory stimuli are encoded by SUSTAIN into perceptual representations based on the value of the stimulus features. The values of

these features are biased according to attention weights operationalized as receptive fields on each feature attribute. During learning, these attention weight receptive fields, which change as a function of the latent model variable $\lambda_i$, are tuned to give more weight to diagnostic features. SUSTAIN represents knowledge as memory clusters of stimulus features and class associations that are built and tuned over the course of learning. Key to the current work is SUSTAIN's mechanism for modifying these clusters throughout learning. Specifically, on each trial, attention-weighted feature values for the current stimulus are compared to existing clusters. If a cluster matches the stimulus well, that cluster drives the decision process and is subsequently *updated* to reflect the attention-weighted feature values of the current stimuli. If there is a poor match between the stimulus and the existing clusters, a new cluster is *created* that reflects the feature values of the current stimulus. Throughout learning, therefore, new clusters are created, and existing clusters updated depending on the trial sequence and category structure of the current learning task. A full mathematical formulization of SUSTAIN is provided below:

*Perceptual encoding.* An input stimulus is presented to SUSTAIN as a pattern of activation on input units that code for the different stimulus features and possible values that these features can take. For each stimulus feature, *i* (e.g., a flower's outer petals), with *k* possible values (e.g., pointy or rounded petals), there are *k* input units. Input units are set to one if the unit represents the feature value or zero otherwise. The entire stimulus is represented by $I^{pos_{ik}}$, with *i* indicating the stimulus feature and *k* indicating the value for feature *i*. "pos" indicates that the stimulus is represented as a point in a multidimensional space. The distance $\mu_{ij}$ between the *i*th stimulus feature and cluster *j*'s position along the *i*th feature is

$$\mu_{ij} = 1/2 \sum_{k=1}^{v_i} | I^{pos_{ik}} - H_j^{pos_{ik}} | \tag{1}$$

such that $v_i$ is the number of possible values that the *i*th stimulus feature can take and is cluster *j*'s position on the *i*th feature for value *k*. Distance $\mu_{ij}$ is always between 0 and 1, inclusive.

*Response selection.* After perceptual encoding, each cluster is activated based on the similarity of the cluster to the input stimulus. Cluster activation is given by:

$$H_j^{act} = \frac{\sum_{i=1}^{n_a} (\lambda_i)^{\gamma} e^{-\lambda_i \mu_{ij}}}{\sum_{i=1}^{n_a} (\lambda_i)^{\gamma}} \tag{2}$$

where $H_j^{act}$ is cluster *j*'s activation, $n_a$ is the number of stimulus features, $\lambda_i$ is the attention weight receptive field tuning for feature *i*, and γ is the attentional parameter (constrained

to be non-negative). Clusters compete to respond to an input stimulus through mutual inhibition. The final output of each cluster *j* is given by:

$$H_j^{out} = \frac{(H_j^{act})^{\beta}}{\sum_{i=1}^{n_c}(H_j^{act})^{\beta}} H_j^{act} \tag{3}$$

where $n_c$ is the current number of clusters and β is a lateral inhibition parameter (constrained to be non-negative) that controls the level of cluster competition. The cluster that wins the competition, $H_m$, passes its output to the *k* output units of the unknown feature dimension *z*:

$$C_{zk}^{out} = w_{m,zk} H_m^{out} \tag{4}$$

where $C_{zk}^{out}$ is the output of the unit representing the *k*th feature value of the *z*th feature, and $w_{m,zk}$ is the weight from the winning cluster, $H_m$, to the output unit $C_{zk}$. In the current simulations, the class label is the only unknown feature dimension. Thus, equation 4 is calculated for each of the two values of the class label. Finally, the probability of making a response *k* for a queried dimension, *z*, on a given trial is:

$$P(k) = \frac{e^{(dC_{zk}^{out})}}{\sum_{j=1}^{v_z} e^{(dC_{zk}^{out})}} \tag{5}$$

*Memory cluster modification.* SUSTAIN was initialized with zero clusters. During learning, clusters are recruited in response to a combination of the order of the stimuli presented in the participant-specific trial orders and the error feedback received on each trial. Two events could lead to the creation of a new cluster: 1) the model predicts the incorrect class label or 2) the winning cluster's activation is below a threshold, τ (constrained to be between 0 and 1). If either of these two criteria are true, a new cluster is created; otherwise, the winning cluster from the cluster competition is updated to reflect current stimulus features and class label according to the learning rules explained next.

*Learning.* SUSTAIN's learning rules determine how clusters are updated during learning. Only the winning clusters are updated. If a new cluster is recruited on a trial, it is considered the winning cluster. Otherwise, the cluster that is most similar to the current stimulus will be the winner. The winning cluster $H_m$, is adjusted by:

$$\Delta H_m^{pos_{ik}} = \eta(I^{pos_{ik}} - H_m^{pos_{ik}}) \tag{6}$$

where η is the learning rate parameter. The result of the updating is that the winning cluster moves toward the current stimulus. Over the course of learning, each cluster will

tend toward the center of its members. Attention weight receptive field tunings for the different feature dimensions are updated according to:

$$\Delta\lambda i = \eta e^{-\lambda_i \mu_{im}}(1 - \lambda_i \mu_{im}) \quad (7)$$

where *m* indexes the winning cluster. The weights from the winning cluster to the output units are adjusted by a one-layer delta learning rule.

$$\Delta w_{m,zk} = \eta(t_{zk} - C_{zk}^{out})H_m^{out} \quad (8)$$

*Simulations.* Stimuli were presented to SUSTAIN using the same trial order for each task as experienced by the participants. Before each task, the attention weight receptive field tunings and clusters were reinitialized. For each participant and task, the free parameters, $\gamma$, $\beta$, $\eta$, *d,* and $\tau$, were optimized to best match the participant's trial-by-trial responses. Specifically, SUSTAIN's predicted probability of a making the same response as the participant (**Eq. 5**) was summarized with log likelihood (*lnL*) and the model parameters were optimized to maximise likelihood using a differential evolution genetic algorithm approach[33,84] (*scipy* version 1.2.1). To ensure that these trial-by-trial model predictions were successful, we compared the resulting model fits to a second model optimization analysis that followed the more traditional approach of fitting to summaries of accuracies across trial blocks[4]. In this approach, the average accuracy for both participant and model is calculated for each block of 16 trials and these block accuracies are used to calculate model fit error. In minimizing block-wise error between model and participant behaviour, this type of model optimization can account for patterns in learning accuracy across blocks (i.e., participant- and task-specific learning curves). We performed this second model optimization for each participant in each task and then used the optimized parameters to calculate the same log likelihood measure based on trial-by-trial responses as was used in the central model optimization. The logic follows that if trial-wise optimization provides a better account of behaviour (i.e., the specific responses made by participants on each trial), the trial-wise likelihoods will be significantly higher than the likelihoods from the block-wise optimization. Indeed, this is exactly what we found in comparing model fits between approaches in a mixed-effects regression: trial-wise optimization far exceeded block-wise optimization (mean $lnL_{trial}$=-49.7, mean $lnL_{block}$=-63.5, $t(174)=4.56$, $p=9.7\times10^{-6}$). These results suggest the trial-wise optimization provided a better account of participants' specific responses on each trial and validates using the optimized parameters to generate trial-specific memory formation function predictions.

Best fitting parameters from the trial-wise optimization (mean and 95% confidence intervals: $\gamma$ = 8.567 ± 1.374, $\beta$ = 3.494 ± 0.327, $\eta$ = 0.209 ± 0.049, *d* = 21.542 ± 3.532, $\tau$ =0.191 ± 0.028) for each participant and task were then leveraged to generate model-based predictions for trials when memory clusters were created versus updated (**Figure 2A**). Trials were labelled as create or update trials according to the latent trial-wise operations of the model until the last cluster for a given task was created. At this point,

memory modification was considered completed and the remaining trials were labeled as baseline. These predictions of create, update, and baseline trials served as model-based regressors for memory modification events in the fMRI analyses described below.

*MRI data acquisition*

Whole-brain imaging data were acquired on a 3.0T Siemens Skyra system at the University of Texas at Austin Imaging Research Center. A high-resolution T1-weighted MPRAGE structural volume (TR = 1.9s, TE = 2.43ms, flip angle = 9°, FOV = 256mm, matrix = 256x256, voxel dimensions = 1mm isotropic) was acquired for coregistration and parcellation. Two oblique coronal T2-weighted structural images were acquired perpendicular to the main axis of the hippocampus (TR = 13,150ms, TE = 82ms, matrix = 384x384, 0.4x0.4mm in-plane resolution, 1.5mm thru-plane resolution, 60 slices, no gap). High-resolution functional images were acquired using a T2*-weighted multiband accelerated EPI pulse sequence (TR = 2s, TE = 31ms, flip angle = 73°, FOV = 220mm, matrix = 128x128, slice thickness = 1.7mm, number of slices = 72, multiband factor = 3) allowing for whole brain coverage with 1.7mm isotropic voxels.

*MRI data preprocessing*

Anatomical and functional volumes were preprocessed with fmriprep version 1.1.1 using default pipelines. Participant T1 volumes were skull-stripped and registered to the MNI 2009c asymmetric template using *ANTs*[85]. Functional volumes were motion corrected (FSL *mcflirt*), corrected for susceptibility distortions using fieldmaps (FSL *fugue*), and registered to the participant's T1 volume (Freesurfer *bbregister*). Functional volumes were resampled to MNI template space before the main analysis. Whole brain automated parcellation on each participant's T1 volume was also conducted with Freesurfer (*recon-all*).

*Beta series estimation*

For each participant, trial-level beta series were estimated for the first fMRI run in each learning task using the LS-S approach[86]. Beta volumes for both stimulus presentation and feedback events were separately modelled for each trial; however, BOLD activity related to stimulus presentation was the focus of the current study. Confound regressors in the trial-level beta estimation included framewise displacement, six motion parameters (three degrees of translation and rotation) and estimates of BOLD signal noise as defined by the first six components of anatomical CompCor[87].

*fMRI analysis of memory formation events*

To characterize neural signatures of memory formation events, whole-brain beta series of trial-by-trial stimulus presentations was analyzed with a multi-level mixed effects linear regression approach (*statsmodels* python library, version 0.8). All participants' beta series from each task were concatenated and analyzed simultaneously with a custom

searchlight kernel implemented in PyMVPA. Specifically, within each searchlight (radius=2 voxels), beta estimates were averaged and evaluated with a regression model that included trial-level fixed factors for: 1) model-predicted memory cluster modification events (create, update, baseline), 2) correct or incorrect responses, and 3) trial number. Participant was also included in the regression as a random intercept. Thus, the regression model was constructed to characterize neural activation that differed between model-predicted memory cluster create and update events relative to baseline while controlling for potentially confounding effects of response correctness and learning time. Applying this custom searchlight to the whole brain resulted in a statistical *t*-map of the contrast between create and update trials.

Cluster-level inferences for the create versus update contrast was performed by first estimating the noise model of the dataset and statistical analysis. The residual volumes from the regression analysis describe above were extracted and analyzed with AFNI[88] *3dFWMHx* using the *acf* option to estimate the intrinsic autocorrelation of the data ($a$=0.659, $b$=3.082, $c$=11.479). These *acf* parameters were then inputted to AFNI *3dClustSim* (version 19.1.14) to estimate a cluster-extent threshold with 2-sided thresholding and third-nearest neighbour clustering. The result was a thresholding scheme in which the create vs. update statistical map was voxel-wise corrected at $p$=0.001 and cluster corrected at $p$=0.05, which corresponded to a cluster-extent threshold of greater than 42 voxels.

| region | Z peak | voxels | x | y | z |
|---|---|---|---|---|---|
| precuneus | 5.08 | 399 | -14 | -70 | 34 |
| right ventral striatum | 5.01 | 225 | 20 | 8 | -4 |
| left ventral striatum | 4.99 | 188 | -20 | 14 | -8 |
| premotor | 5.33 | 97 | -2 | 0 | 34 |
| superior parietal | 4.09 | 65 | -10 | -34 | 46 |
| right hippocampus | 5.26 | 57 | 30 | -18 | -19 |
| posterior cingulate | 4.18 | 54 | 6 | -38 | 26 |
| premotor | 4.71 | 44 | 6 | -14 | 48 |

**Table 2:** Significant clusters showing distinct neural engagement to model-derived memory formation events (i.e., memory create vs. update). All reported clusters showed greater activation for create relative to update events. Cluster information includes the corresponding anatomical label as defined in the Harvard-Oxford Structural or Juelich Histological atlases, the *t* statistic of the peak voxel, the cluster size in number of voxels, and peak voxel location in MNI coordinates (x, y, z).

*Relating neural signatures of memory formation to learning*

Time series were extracted from the trial-by-trial beta estimates for the hippocampal and ventral striatum ROIs. Specifically, the betas within a sphere defined by a radius of 2mm and centred on the voxel with the peak Z value for the contrast defining the cluster (MNI coordinates - HPC: 30, -18, -20; R VS: 8, 20, -4: L VS: -20, 14, -8) were averaged for each trial and ROI. These time series were then used to calculate a neural index of the difference in create vs. update activation (i.e., memory modification effect) for each participant, task, and ROI.

We evaluated the relationship between the neural indices of the memory modification effect and end of learning performance with two complementary analyses. First, in a mixed effects linear regression, we entered end of learning performance (tasks labeled as learners versus not), ROI, and their interaction as predictors of the memory modification effect with participants included as a random effect (**Figure 3A**). We observed a main effect of learner ($\chi^2$=4.340, p=0.037) and an interaction with ROI ($\chi^2$=6.457, p=0.039). Follow-up analyses that estimated the relationship between end of learning performance and create-update activation for each ROI in separate regression models were also conducted to aid interpretation. These analyses showed an effect of learning performance on memory formation specific to HPC (HPC: $\beta$=0.336, CI=[0.056, 0.616], p=0.019) with no effect from ventral striatum (L VS: $\beta$=-0.113, CI=[-0.479, 0.254], p=0.547; R VS: $\beta$=-0.206, CI=[-0.499, 0.088], p=0.169).

We also performed a mixed effects binomial logistic regression analysis with the memory modification effect and ROI as predictors of end of learning performance (learner versus not) for each task with participants as a random effect (**Figure 3B**). With the binary outcome of task learner, this analysis is akin to a classification approach. We observed a main effect of memory modification ($\chi^2$=5.726, p=0.017) and an interaction with ROI ($\chi^2$=8.276, p=0.016). Critically, as revealed by ROI-specific statistical models, these effects were driven by a significant effect of HPC memory modification (log odds=2.752, CI=[1.225, 6.181], p=0.014). There was no significant relationship between memory modification effects in ventral striatum and end of learning performance (ps>0.24).

*Reliability of temporal coupling between model and neural measures*

A key aspect of our approach is linking trial-wise model predictions of memory formation functions, as derived through simulations of learning behaviour, to neural measures. If task-specific model predictions of memory functions are capturing important dynamics in neural engagement, the observed differences between memory create and update trials should be significantly stronger than when the temporal coupling between model and brain is broken. To directly evaluated this, we conducted a permutation analysis in which model-based memory formation predictions were randomly shuffled across tasks within each participant and the primary analysis comparing neural engagement to memory

create and update trials was performed. This procedure was repeated 1000 times to generate a distribution of estimated create-update trial effects (i.e., *z* statistics) separately for the HPC and striatum ROIs. A *p* value was calculated that corresponded to the proportion of shuffled analyses that resulted in a larger create-update effect than the observed effect. The results showed that the temporal coupling of task-specific model predictions and neural engagement in HPC led to a significantly stronger difference between create and update trials relative to the shuffled distribution (p=0.005); this was also true of right VS (p=0.022) but not left VS (p=0.232).

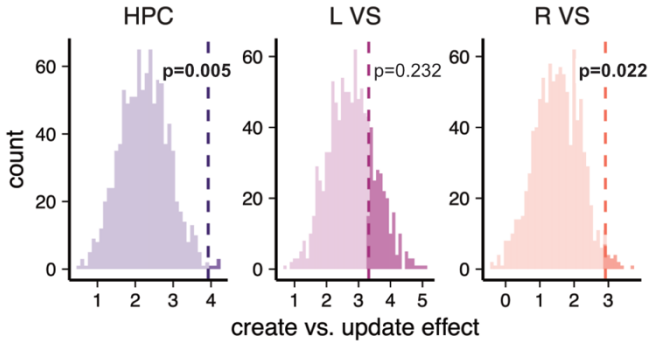


**Figure 5.** Permutation analysis of create-update effect. The results of the create-update difference from 1000 shuffled analyses (histograms) are depicted relative to the observed create-update effect (dashed lines). *P* values correspond to the proportion of shuffled analyses that resulted in create-update effects higher than the observed effect (darker regions). N=25.

*Hippocampal engagement relates to SUSTAIN cluster activation*

For any given learning experience, SUSTAIN's memory operations (i.e., create a new cluster or update an existing cluster) are engaged depending on the similarity of the current stimulus information to the model's most activated (i.e., most similar) stored cluster, a quantity formalized as cluster activation, $H_{act}$ (**eq. 2**). The more that the current stimulus matches one cluster, the more likely that cluster will be updated; on the flip side, if the current stimulus does not match any of the stored clusters, a new cluster will likely be created. Our central analysis focused on the binary outcome representing these distinct memory formation events. However, it follows that the trial-by-trial fluctuations of the model-predicted cluster activation underlying these events may relate to neural activation. Specifically, we predicted that higher cluster activation would be associated with lower neural engagement. To test this prediction, we conducted a mixed-effects regression that linked neural activation within our ROIs (i.e., peak activity) to model-predicted maximum cluster activation on each trial. These models also included fixed effects for trial number and correctness and a random effect of participant. As expected, we found that hippocampal engagement was significantly related to maximum cluster activation ($\beta$=-0.47, CI=[-0.82, -0.11], p=0.01); neither striatum ROIs showed such a relationship (ps>0.33). These results extend the central findings that hippocampal engagement during early learning reflects the key latent variables and operations for building memories that support new categories.

*Functional connectivity*

To characterize potential functional networks associated with the hippocampus during memory create versus update events, we conducted a functional connectivity analysis[48]. Specifically, the timeseries from the peak of the hippocampal cluster was entered into a searchlight-based mixed effects linear regression analysis such that trial-by-trial hippocampal activation interacted with the trial-level create versus update regressor to predict whole-brain trial-by-trial beta estimates. The regression model also included confound regressors for trial correctness and trial number. To additionally account for non-task related temporal noise in the beta series, we also included nuisance regressor timeseries from the ventricles (left: -18, -38, 15; right: 21, -38, 15) and white matter (left: -24, -5, 35; right: 24, -5, 35). The searchlight (radius=2 voxels) regression analysis was conducted on the whole brain to generate statistical *t*-maps that highlighted brain regions that were distinctly functionally coactive with hippocampus for memory create or update trials (**Figure 4**).

| | Label | peak *t* value | voxels | x | y | z |
|---|---|---|---|---|---|---|
| create > update | vmPFC | 4.31 | 78 | -8 | 24 | -10 |
| | right angular gyrus | 4.1 | 47 | 48 | -42 | 22 |
| update > create | premotor cortex | 4.67 | 161 | -16 | -22 | 76 |
| | | 4.21 | 160 | 12 | -18 | 70 |
| | | 3.83 | 57 | -12 | -8 | 78 |

**Table 3:** Significant clusters showing function specific interregional correlation with HPC. Cluster information includes the corresponding anatomical label as defined in the Harvard-Oxford Structural or Juelich Histological atlases, the *t* statistic of the peak voxel, the cluster size in number of voxels, and peak voxel location in MNI coordinates (x, y, z).

## References


1. Homa, D., Cross, J., Cornell, D., Goldman, D. & Schwartz, S. Prototype abstraction and classification of new instances as a function of number of instances defining the prototype (concept formation and learning). *J. Exp. Psychol.* **101**, 116–122 (1973).
2. Smith, J. D. & Minda, J. Prototypes in the Mist: The Early Epochs of Category Learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 1411–1436 (1998).
3. Nosofsky, R. Attention, similarity, and the identification-categorization relationship. *J. Exp. Psychol. Gen.* **115**, 39–61 (1986).
4. Love, B. C., Medin, D. & Gureckis, T. M. SUSTAIN: A network model of category learning. *Psychol. Rev.* **111**, 309–332 (2004).
5. Mack, M. L., Love, B. C. & Preston, A. R. Building concepts one episode at a time: The hippocampus and concept formation. *Neurosci. Lett.* **680**, 31–38 (2018).
6. Morton, N. W. & Preston, A. R. Concept formation as a computational cognitive process. *Curr. Opin. Behav. Sci.* **38**, 83–89 (2021).
7. Poldrack, R. *et al.* Interactive memory systems in the human brain. *Nature* **414**, 546–550 (2001).
8. Knowlton, B., Mangels, J. & Squire, L. A Neostriatal Habit Learning System in Humans. *Science* **273**, 1399–1402 (1996).
9. Shohamy, D., Myers, C. E., Kalanithi, J. & Gluck, M. A. Basal ganglia and dopamine contributions to probabilistic category learning. *Neurosci. Biobehav. Rev.* **32**, 219–236 (2008).
10. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
11. Norman, K. & O'Reilly, R. Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary-Learning-Systems Approach. *Psychol. Rev.* **110**, 611–646 (2003).
12. O'Reilly, R. C. & Rudy, J. W. Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function. *Psychol. Rev.* **108**, 311–345 (2001).
13. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, (2017).
14. Rolls, E. T. The mechanisms for pattern completion and pattern separation in the hippocampus. *Front. Syst. Neurosci.* **7**, 74 (2013).
15. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* **153**, 40–56 (2018).
16. Sučević, J. & Schapiro, A. C. A neural network model of hippocampal contributions to category learning. *eLife* **12**, e77185 (2023).
17. Schlichting, M. L., Mumford, J. A. & Preston, A. R. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* **6**, 8151 (2015).
18. Hulbert, J. C. & Norman, K. a. Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cereb. Cortex* 1–15 (2014) doi:10.1093/cercor/bhu284.
19. Ritvo, V. J. H., Nguyen, A., Turk-Browne, N. B. & Norman, K. A. Differentiation and Integration of Competing Memories: A Neural Network Model. *eLife* **12**, (2023).
20. Bieri, K. W., Bobbitt, K. N. & Colgin, L. L. Slow and Fast Gamma Rhythms Coordinate Different Spatial Coding Modes in Hippocampal Place Cells. *Neuron* **82**, 670–681 (2014).
21. Igarashi, K. M., Lu, L., Colgin, L. L., Moser, M. B. & Moser, E. I. Coordination of entorhinal-hippocampal ensemble activity during associative learning. *Nature* **510**, 143–147 (2014).
22. Colgin, L. L. *et al.* Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature* **462**, 353–357 (2009).
23. Singer, A. C., Carr, M. F., Karlsson, M. P. & Frank, L. M. Hippocampal SWR Activity Predicts Correct Decisions during the Initial Learning of an Alternation Task. *Neuron* **77**, 1163–1173 (2013).
24. Joo, H. R. & Frank, L. M. The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. *Nature Reviews Neuroscience* vol. 19 744–757 Preprint at https://doi.org/10.1038/s41583-018-0077-1 (2018).

25. Duncan, K., Ketz, N., Inati, S. J. & Davachi, L. Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus* **22**, 389–398 (2012).
26. Bein, O., Duncan, K. & Davachi, L. Mnemonic prediction errors bias hippocampal states. *Nat. Commun.* **11**, 3451 (2020).
27. Wanjia, G., Favila, S. E., Kim, G., Molitor, R. J. & Kuhl, B. A. Abrupt hippocampal remapping signals resolution of memory interference. *Nat. Commun.* **12**, 4816 (2021).
28. Davis, T., Love, B. C. & Preston, A. R. Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cereb. Cortex* **22**, 260–273 (2012).
29. Bowman, C. & Zeithamova, D. Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *J. Neurosci.* **28**, 2605–2614 (2018).
30. Theves, S., Neville, D. A., Fernández, G. & Doeller, C. F. Learning and Representation of Hierarchical Concepts in Hippocampus and Prefrontal Cortex. *J. Neurosci.* **41**, 7675–7686 (2021).
31. Love, B. C. & Gureckis, T. M. Models in search of a brain. *Cogn. Affect. Behav. Neurosci.* **7**, 90–108 (2007).
32. Davis, T., Love, B. C. & Preston, A. R. Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *J. Exp. Psychol. Learn. Mem. Cogn.* **38**, 821–839 (2012).
33. Mack, M. L., Love, B. C. & Preston, A. R. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 13203–13208 (2016).
34. Shepard, R. N., Hovland, C. I. & Jenkins, H. M. Learning and memorization of classification. *Psychol. Monogr.* **75**, 517 (1961).
35. Mumford, J., Turner, B., Ashby, F. & Poldrack, R. Deconvoling BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).
36. Antzoulatos, E. G. & Miller, E. K. Differences between Neural Activity in Prefrontal Cortex and Striatum during Learning of Novel Abstract Categories. *Neuron* **71**, 243–249 (2011).
37. Ballard, I., Miller, E. M., Piantadosi, S. T., Goodman, N. D. & McClure, S. M. Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning. *Cereb. Cortex* **28**, 3965–3975 (2018).
38. Calderon, C. B. *et al.* Signed reward prediction errors in the ventral striatum drive episodic memory. *J. Neurosci.* **41**, 1716–1726 (2021).
39. Uncapher, M. R. & Wagner, A. D. Posterior parietal cortex and episodic encoding: Insights from fMRI subsequent memory effects and dual-attention theory. *Neurobiol. Learn. Mem.* **91**, 139–154 (2009).
40. Hutchinson, J. B. *et al.* Functional heterogeneity in posterior parietal cortex across attention and episodic memory retrieval. *Cereb. Cortex* **24**, 49–66 (2014).
41. Kuhl, B. A. & Chun, M. M. Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **34**, 8051–60 (2014).
42. Preston, A. R. & Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Curr. Biol.* **23**, R764-73 (2013).
43. Schlichting, M. L. & Preston, A. R. Memory integration: Neural mechanisms and implications for behavior. *Curr. Opin. Behav. Sci.* **1**, 1–8 (2015).
44. Mack, M. L., Preston, A. R. & Love, B. C. Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* **11**, 1–11 (2020).
45. Gilboa, A. & Marlatte, H. Neurobiology of Schemas and Schema-Mediated Memory. *Trends Cogn. Sci.* **14**, 417–428 (2017).
46. Schlichting, M. L. & Preston, A. R. Hippocampal–medial prefrontal circuit supports memory updating during learning and post-encoding rest. *Neurobiol. Learn. Mem.* **134**, 91–106 (2016).
47. van Kesteren, M. T. R., Fernández, G., Norris, D. G. & Hermans, E. J. Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7550–7555 (2010).
48. Rissman, J., Gazzaley, A. & D'Esposito, M. Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage* **23**, 752–763 (2004).
49. Ashby, S. R. & Zeithamova, D. Category-Biased Neural Representations Form Spontaneously during Learning That Emphasizes Memory for Specific Instances. *J. Neurosci.* **42**, 865–876 (2022).
50. Xue, G. *et al.* Complementary role of frontoparietal activity and cortical pattern similarity in successful episodic memory encoding. *Cereb. Cortex* **23**, (2013).

51. Braunlich, K., Gomez-Lavin, J. & Seger, C. a. Frontoparietal networks involved in categorization and item working memory. *NeuroImage* **107**, 146–162 (2015).
52. Morton, N. W., Schlichting, M. L. & Preston, A. R. Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference. *Proc. Natl. Acad. Sci.* **117**, 29338–29345 (2020).
53. Squire, L. R. & Zola, S. M. Structure and function of declarative and nondeclarative memory systems. *Proc. Natl. Acad. Sci.* **93**, 13515–13522 (1996).
54. Davis, T., Xue, G., Love, B. C., Preston, A. R. & Poldrack, R. A. Global neural pattern similarity as a common basis for categorization and recognition memory. *J. Neurosci. Off. J. Soc. Neurosci.* **34**, 7472–84 (2014).
55. Theves, S., Fernández, G. & Doeller, C. F. The Hippocampus Maps Concept Space, Not Feature Space. *J. Neurosci. Off. J. Soc. Neurosci.* **40**, 7318–7325 (2020).
56. Bowman, C. R., Iwashita, T. & Zeithamova, D. Tracking prototype and exemplar representations in the brain across learning. *eLife* **9**, e59360 (2020).
57. Heffernan, E. M., Schlichting, M. L. & Mack, M. L. Learning exceptions to the rule in human and model via hippocampal encoding. *Sci. Rep.* **11**, 21429 (2021).
58. Kumaran, D. & Maguire, E. A. Novelty signals: a window into hippocampal information processing. *Trends Cogn. Sci.* **13**, 47–54 (2009).
59. Larkin, M. C., Lykken, C., Tye, L. D., Wickelgren, J. G. & Frank, L. M. Hippocampal output area CA1 broadcasts a generalized novelty signal during an object-place recognition task. *Hippocampus* **24**, 773–783 (2014).
60. Sinclair, A. H., Manalili, G. M., Brunec, I. K., Adcock, R. A. & Barense, M. D. Prediction errors disrupt hippocampal representations and update episodic memories. *Proc. Natl. Acad. Sci.* **118**, e2117625118 (2021).
61. Zheng, L., Gao, Z., McAvan, A. S., Isham, E. A. & Ekstrom, A. D. Partially overlapping spatial environments trigger reinstatement in hippocampus and schema representations in prefrontal cortex. *Nat. Commun.* **12**, 6231 (2021).
62. Steemers, B. *et al.* Hippocampal Attractor Dynamics Predict Memory-Based Decision Making. *Curr. Biol.* **26**, 1750–1757 (2016).
63. Julian, J. B. & Doeller, C. F. Remapping and realignment in the human hippocampal formation predict context-dependent spatial behavior. *Nat. Neurosci.* **24**, 863–872 (2021).
64. Baldassano, C. *et al.* Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* **95**, 709-721.e5 (2017).
65. Clewett, D., DuBrow, S. & Davachi, L. Transcending time in the brain: How event memories are constructed from experience. *Hippocampus* **29**, 162–183 (2019).
66. Braunlich, K. & Love, B. C. Occipitotemporal representations reflect individual differences in conceptual knowledge. *J. Exp. Psychol. Gen.* **148**, 1192–1203 (2019).
67. Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A. & Freeman, J. H. Prelimbic cortex maintains attention to category-relevant information and flexibly updates category representations. *Neurobiol. Learn. Mem.* **185**, 107524 (2021).
68. Broschard, M. B., Kim, J., Love, B. C. & Freeman, J. H. Category learning in rodents using touchscreen-based tasks. *Genes Brain Behav.* **20**, e12665 (2021).
69. Morton, N. W., Sherrill, K. R. & Preston, A. R. Memory integration constructs maps of space, time, and concepts. *Curr. Opin. Behav. Sci.* **17**, (2017).
70. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
71. Robin, J. & Moscovitch, M. Details, gist and schema: hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Curr. Opin. Behav. Sci.* **17**, 114–123 (2017).
72. Wagner, A. D., Shannon, B. J., Kahn, I. & Buckner, R. L. Parietal lobe contributions to episodic memory retrieval. *Trends Cogn. Sci.* **9**, 445–453 (2005).
73. Zhao, Y., Chanales, A. J. H. & Kuhl, B. A. Adaptive Memory Distortions Are Predicted by Feature Representations in Parietal Cortex. *J. Neurosci.* **41**, 3014–3024 (2021).
74. Corbetta, M., Patel, G. & Shulman, G. L. The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron* **58**, 306–324 (2008).

75. Turk-Browne, N. B., Golomb, J. D. & Chun, M. M. Complementary attentional components of successful memory encoding. *NeuroImage* **66**, 553–562 (2013).
76. Tibon, R., Fuhrmann, D., Levy, D. A., Simons, J. S. & Henson, R. N. Multimodal Integration and Vividness in the Angular Gyrus During Episodic Encoding and Retrieval. *J. Neurosci.* **39**, 4365–4374 (2019).
77. Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U. & Waldron, E. M. A Neuropsychological Theory of Multiple Systems in Category Learning. *Psychol. Rev.* **105**, 442–481 (1998).
78. Ashby, F. G. & Maddox, W. T. Human category learning. *Annu. Rev. Psychol.* **56**, 149–78 (2005).
79. Ashby, F. G., Ennis, J. M. & Spiering, B. J. A neurobiological theory of automaticity in perceptual categorization. *Psychol. Rev.* **114**, 632–656 (2007).
80. Foster, D. J. & Knierim, J. J. Sequence learning and the role of the hippocampus in rodent navigation. *Curr. Opin. Neurobiol.* **22**, 294–300 (2012).
81. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map*. (Oxford University Press, Oxford, UK, 1978).
82. Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M. & Tanila, H. The Hippocampus, Memory, and Place Cells: Is It Spatial Memory or a Memory Space? *Neuron* **23**, 209–226 (1999).
83. Blair, M. R., Watson, M. R., Walshe, R. C. & Maj, F. Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *J. Exp. Psychol. Learn. Mem. Cogn.* **35**, 1196–1206 (2009).
84. Storn, R. & Price, K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997).
85. Avants, B. B. *et al.* A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* **54**, 2033–2044 (2011).
86. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).
87. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
88. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).