Learning not to remember: How predicting the future impairs encoding of the present

Brynn E. Sherman, Nicholas B. Turk-Browne*

Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06520, USA

Abstract

Memory enables reminiscence about past experiences and guides processing of future experiences. However, these two functions are inherently at odds: remembering specific past experiences requires storing idiosyncratic properties, but by definition such properties may not be shared with similar situations in the future and thus are not as useful for prediction. We discovered that, when faced with this conflict, the brain prioritizes prediction over encoding. Behavioral tasks showed that pictures allowing for prediction of what will appear next based on learned regularities are poorly encoded into memory. Brain imaging revealed that predictive representations in the hippocampus may be responsible for this worse episodic encoding and suggested that such interference may result from competition between hippocampal pathways. This tradeoff between statistical learning and episodic memory may be adaptive, focusing encoding on experiences for which we do not yet have a predictive model.

Introduction

Human memory contains two fundamentally different kinds of information — episodic and statistical. Episodic memory refers to the encoding of specific details of individual experiences (e.g., what happened on your last birthday),

*Corresponding author; Lead Contact
Email address: nicholas.turk-browne@yale.edu (Nicholas B. Turk-Browne)

whereas statistical learning refers to extracting what is common across multiple experiences (e.g., what tends to happen at birthday parties). Episodic memory allows for vivid recollection and nostalgia about past events, whereas statistical learning leads to more generalized knowledge and predictions about new situations. Episodic memory occurs rapidly and stores even related experiences distinctly in order to minimize interference, whereas statistical learning occurs more slowly and overlays memories in order to represent their common elements or regularities. Given these behavioral and computational differences, theories of memory have argued that these two kinds of information must be processed serially and stored separately in the brain (McClelland et al., 1995; Squire, 2004): episodic memories are formed first in the hippocampus and these memories in turn provide the input for later statistical learning in the neocortex as a result of consolidation (Frankland and Bontempi, 2005; Richards et al., 2014; Tompary and Davachi, 2017).

Here we reveal a relationship between episodic memory and statistical learning in the reverse direction, whereby learned regularities determine which memories are formed in the first place. Specifically, we examine whether the ability to predict what will appear next — a signature of statistical learning — reduces encoding of the current experience into episodic memory. This hypothesis depends on two theoretical commitments: first, that the adaptive function of memory is to guide future behavior by generating expectations based on prior experience (Schacter et al., 2017); second, that memory resources are limited, because of attentional bottlenecks that constrain encoding (Aly and Turk-Browne, 2017) and/or because new encoding interferes with the storage or retrieval of existing memories (Shiffrin and Atkinson, 1969). Accordingly, in allocating memory resources, we propose that it is less important to encode an ongoing experience when it already generates strong expectations about future states of the world. When the current experience affords no such expectations, however, encoding it into memory provides the opportunity to extract new, unknown regularities that enable more accurate predictions in subsequent encounters. After demon-

strating this role for statistical learning in episodic memory behaviorally, we identify an underlying mechanism in the brain using fMRI, based on the recent discovery that both processes depend on the hippocampus and thus compete to determine its representations and output (Schapiro et al., 2017).

We exposed human participants to a stream of pictures and later tested their memory (Figure 1A). The pictures consisted of outdoor scenes from 12 different categories (e.g., beach, mountain, farm). Some of the categories (type A) were predictive of which category appeared next (type B), whereas other categories (type X) were non-predictive. That is, every time participants saw a picture from an A category, they always saw a picture from a specific B category next; however, when a picture from an X category appeared, it was followed by a picture from one of several other categories (Figure 1B). Participants were not informed about these predictive A $\rightarrow$ B category relationships and learned them incidentally through exposure (Brady and Oliva, 2008). Although each category was shown several times, every individual picture in the stream was a novel exemplar from the category and only shown once. For example, whenever a picture from the beach category appeared, it was a new beach that they had not seen before. After the stream, we tested memory for these individual pictures amongst new exemplars from the same categories. The key question was whether memory for the exemplars from the predictive categories would be remembered worse than those of non-predictive categories. This was first tested in two behavioral experiments.

Results

While viewing the stream, the 30 participants in Experiment 1 performed a cover task in which they judged whether or not there was a manmade object in the scene. Response times (RTs) on this task were used to assess whether participants had learned the predictive relationships between A and B categories. If so, we expected that responses for B categories would become faster over time in the experiment, as participants became able to anticipate which

category would appear. Providing evidence of learning, there was a significant interaction between condition (A, B, X) and time (1st, 2nd, 3rd, 4th quartile of the stream) ($F(6, 174) = 2.27$, $p = 0.039$). This interaction reflected a pattern of growing facilitation for pictures from the B category, relative to both X and A categories whose appearance in the stream could not be anticipated (Figure S1A).

Given our hypothesis, we expected worse encoding of the exemplars from the predictive A categories, leading to greater forgetting in a later memory test. Indeed, there was a main effect of condition ($F(2,58) = 4.75$, $p = 0.012$), with a lower hit rate for pictures from the A categories relative to both B ($t(29) = -2.79$, $p = 0.009$) and X ($t(29) = -2.33$, $p = 0.027$) categories (Figure 1C). There was no difference in hit rate between B and X categories ($t(29) = 1.19$, $p = 0.243$), showing that the memory deficit is selective to whether a category was predictive (A vs. X), not whether it was predictable (B vs. X).

Because the predictive A $\rightarrow$ B relationships had to be learned during the stream, we expected that the memory deficit for pictures from the A category would emerge over time. Indeed, this effect was evident only in the third and fourth quartiles of stream exposure (Figure S1B). Furthermore, the overall memory deficit reflected a failure to encode specific A exemplars rather than a generic impairment for A categories (De Brigard et al., 2017; Smith et al., 2013), as the false alarm rate for new exemplars from each category at test did not differ by condition ($F(2, 58) = 0.29$, $p = 0.751$).

This experiment demonstrated that episodic encoding is worse for predictive vs. non-predictive pictures using a surprise recognition memory test. We interpret this result as evidence of competition between prediction and encoding in the hippocampus. However, recognition tests do not selectively probe aspects of episodic memory that critically depend on the hippocampus. Participants could have relied upon a generic sense of familiarity with the pictures, which can be supported by cortical areas (Brown and Aggleton, 2001; Davachi et al., 2003; Norman and O'Reilly, 2003).

We thus designed Experiment 2 with a different, recall-based memory test.

This test required retrieval of specific spatiotemporal details, which is known to tax the hippocampus (Miller et al., 2013) and is a hallmark function of episodic memory (e.g., remembering who arrived first at a birthday party). A new group of 30 participants was exposed to the same kind of picture stream and performed the same cover task as in Experiment 1. However, during the memory test phase they were unexpectedly asked to indicate at what exact time (on the clock) they had seen each picture during the stream. As before, encoding of the time was incidental as they were not informed in advance that they would be tested. This kind of precise temporal source memory requires the retrieval of details about the context in which each picture was encoded, which depends on the hippocampus (Davachi and DuBrow, 2015; Mitchell and Johnson, 2009).

We analyzed the accuracy of the time reports in terms of absolute deviation from the correct time on the clock at encoding, such that higher values indicate *less* precise memory (Figure 1D). Consistent with the results of Experiment 1, there was a main effect of condition (A, B, X) on temporal deviation ($F(2,58) = 3.17$, $p = 0.049$). Pictures from the A categories had greater deviation (less precision) than those from the X categories ($t(29) = 2.26$, $p = 0.031$); the same pattern was present for A vs. B categories but did not reach significance ($t(29) = 1.45$, $p = 0.157$), nor did B and X categories differ ($t(29) = 1.23$, $p = 0.229$).

What explains the reduced encoding of predictive pictures in Experiments 1 and 2? We propose that this results from the co-dependence of statistical learning and episodic memory on the same brain region — the hippocampus (Schapiro et al., 2017). Specifically, we hypothesized that the appearance of a picture from an A category triggers the retrieval and predictive representation of the corresponding B category in the hippocampus. This in turn prevents the hippocampus from forming and encoding a new representation of the specific details of that particular A picture that would be needed for later recall from episodic memory.

To test this hypothesis, Experiment 3 employed high-resolution fMRI in 36 new participants who viewed the same kind of picture stream as in Experiments 1 and 2. We used a multivariate pattern classification approach from machine

learning (Cohen et al., 2017), which quantified neural prediction of B categories during the encoding of A pictures. Classification models were trained for each category based on patterns of fMRI activity in a separate phase of the experiment where participants were shown pictures from all categories in a random order. These classifiers were then tested during viewing of the stream containing category pairs, providing a continuous readout of neural evidence for each category. We performed this analysis based on fMRI activity patterns from the hippocampus, our primary region of the interest (ROI), as well as from control ROIs in occipital and parahippocampal cortices (Figure 2, bottom). These control ROIs were chosen because we expected them to be sensitive to the category of the current picture being viewed but not to predict the upcoming B category given an A picture.

To validate our approach, we first trained and tested classifiers on the viewing of pictures from the A ("Perception of A") and B ("Perception of B") categories (Figure 2). Both types of perceptual categories could be decoded in occipital and parahippocampal ROIs. Interestingly, only the perception of B categories (not A) could be decoded in the hippocampus. We next tested our main hypothesis that the hippocampus predicts B categories during viewing of the associated A categories. We trained classifiers on pictures from each B category and tested on pictures from the corresponding A category ("Prediction of B"). Crucially, the upcoming B category could be decoded during A in the hippocampus, but this was not possible in occipital or parahippocampal ROIs. Combining these results, there was a trade-off in the hippocampus between the perception of A (train on A, test on A) and prediction of B (train on B, test on A): during viewing of A, participants with above-chance classification for the upcoming B category (vs. other B categories) had lower classification accuracy for the current A category (vs. other A categories) ($t(34) = 2.23$, $p = 0.033$). This shows that prediction can interfere with the ability of the hippocampus to represent the current picture. Control analyses ruled out potential confounds related to the timing of the fMRI signal: training classifiers on A categories and testing on corresponding B pictures ("Lingering of A"), did not yield reliable decoding in

any ROI.

To further assess the specificity of these results to the hippocampus, we ran an exploratory whole-brain searchlight analysis. We again validated our approach by decoding the perception of B categories (train on B, test on B). The resulting regions, which represented scene categories, were largely consistent with our *a priori* ROIs (Figure S4B). Notably, the prediction of B (train on B, test on A) produced no significant clusters across the brain after correcting for multiple comparisons (Figure S4A), consistent with this effect being specific to the hippocampal ROI.

After having demonstrated prediction from statistical learning in the hippocampus, we tested our critical hypothesis that this prediction would impair simultaneous encoding into episodic memory. We quantified this brain-behavior relationship by correlating (1) each participant's decoding accuracy for prediction of B during A in the hippocampus with (2) their difference in hit rate for A vs. X categories (the key behavioral effect in Experiments 1-2) in the memory test (Figure 2, right). We limited this analysis to participants with decoding accuracy above chance (0.5), as variance at or below chance cannot be interpreted (results were robust to this exclusion, see Supplementary Materials). Consistent with our hypothesis, there was a negative correlation: more accurate prediction of B categories in the hippocampus in response to A pictures was associated with a greater deficit in memory for the A pictures ($r(22) = -0.63$, $p < 0.001$).

How is this interaction between prediction and encoding implemented in the circuitry of the hippocampus? A recent biologically-plausible neural network model of the hippocampus (Schapiro et al., 2017) suggests that episodic memory and statistical learning depend on different pathways, the trisynaptic pathway (TSP) and monosynaptic pathway (MSP), respectively (Figure 3B). The TSP consists of a connection between entorhinal cortex (EC) and the CA1 subfield via intermediate connections through the dentate gyrus (DG) and CA3 subfield. DG and CA3 have sparse activity because of high lateral inhibition; this allows for the distinct representation of similar experiences (i.e., pattern separation; (Leutgeb et al., 2007)), which is needed to avoid interference between episodic

memories. The MSP consists of a direct recurrent connection between EC and CA1. CA1 has lower inhibition and thus higher overall activity and less sparsity; this leads to overlap in the representation of similar experiences, which allows their regularities to be reinforced.

Accordingly, both the TSP and MSP converge on CA1, which we propose is the locus of conflict between episodic memory and statistical learning. We aimed to test this theory with the fMRI data from Experiment 3 using a functional connectivity approach known as psychophysiological interaction (PPI). We reasoned CA1 would interact more with TSP (reflected in correlated activity with a combined CA2/3/DG ROI) during episodic memory and with MSP (reflected in correlated activity with an EC ROI) during statistical learning (Figure 3C). Specifically, we hypothesized that during periods of low episodic encoding and high statistical learning, CA1 should be more functionally connected with with EC than CA2/3/DG.

We quantified episodic encoding using a non-parametric measure of recognition memory fidelity (A') across pictures from both A and B categories (together, referred to as "Structured") and for pictures from X categories ("Random"). We combined A and B categories because these pictures provided the opportunity for both the extraction of regularities and the encoding of individual episodes. Higher A' values indicate more episodic encoding and lower A' values indicate less episodic encoding (and perhaps thus more statistical learning). We separately correlated EC $\leftrightarrow$ CA1 and CA2/3/DG $\leftrightarrow$ CA1 connectivity from the PPI analysis with A' for Structured and Random pictures (Figure 3D).

For Structured pictures, EC $\leftrightarrow$ CA1 connectivity was related to lower A' or worse episodic memory performance ($r(34) = -0.41$, $p = 0.014$), whereas CA2/3/DG $\leftrightarrow$ CA1 connectivity had no relationship to A' ($r(34) = 0.062$, $p = 0.72$); these two correlations significantly differed ($z(35) = -2.01$; $p = 0.044$). For Random pictures, CA1 connectivity with both pathways was unrelated to A' (EC: $r(34) = -0.039$, $p = 0.82$; CA2/3/DG: $r(34) = 0.15$, $p = 0.37$). The lack of any positive relationship of CA2/3/DG $\leftrightarrow$ CA1 connectivity with A' suggests that statistical learning is necessary to reveal an impact of CA1 connectivity

(with EC) on episodic memory. Namely, greater MSP engagement in response to regularities reflects a bias away from episodic memory.

Discussion

To summarize the three experiments, our core findings were (1) that prediction from statistical learning interferes with encoding into episodic memory and (2) that this competition may be explained by the multiplexed function of the hippocampus across convergent pathways. These findings are related to two theoretical issues in the learning and memory literature.

First, the hippocampus is necessary for both memory encoding and retrieval, and yet these functions are fundamentally at odds. Given a partial match between the current experience and past experiences, encoding leverages pattern separation to store a new trace of the current experience, whereas retrieval invokes pattern completion to access old traces of those past experiences. To resolve this incompatibility, it has been argued that the hippocampus toggles between encoding and retrieval states (Hasselmo et al., 1996; Duncan et al., 2012; Patil and Duncan, 2018). In the present study, if seeing a picture from an A category triggers the retrieval of its associated B category, the hippocampus may be pushed into a retrieval state that suppresses memory encoding.

Second, after learning predictive relationships in classical conditioning, "blocking" can occur when new cues are introduced. After one conditioned stimulus (CS1) has been paired with an unconditioned stimulus (US), no associative learning occurs when a second conditioned stimulus (CS2) is added (Kamin, 1969). This is interpreted as CS2 being redundant with CS1, that is, not providing additional predictive value given that the US can be fully explained by CS1. In the present study, the A pictures contain two kinds of features: those that are diagnostic of the category (e.g., sand and water for a beach) and those that are idiosyncratic to each exemplar (e.g., particular people, umbrellas, boats, etc.). If categorical features are sufficient to predict the upcoming B category, idiosyncratic features may not be attended or represented (Mackintosh, 1975;

Kruschke, 2001), impeding the formation of episodic memory. Our findings are not fully consistent with this account, however. Blocking might predict that the A pictures are represented more categorically and this enables prediction of the B category; yet, we found a trade-off in the hippocampus between perceptual evidence of the A category and predictive evidence of the B category.

Stepping back, why are the computationally opposing functions of episodic memory and statistical learning housed together in the hippocampus? We propose that this shared reliance allows them to regulate each other. By analogy, using your right foot to operate both the brake and gas pedals in a car serves as an anatomical constraint that forces you to either accelerate or decelerate, but not both at the same time. A similarly adaptive constraint may be present in the hippocampus, reflecting mutual inhibition between episodic memory and statistical learning. When predictive information is available in an environment, it may be redundant to encode new experiences. Moreover, encoding such experiences would risk over-fitting or improperly updating known, predictive regularities with idiosyncratic or noisy details. By focusing on upcoming events, the hippocampus can better serve as a comparator between expectations and inputs (Kumaran and Maguire, 2006), prioritizing the encoding of novel and unexpected events (Greve et al., 2017; Henson and Gagnepain, 2010).

Author Contributions

Conceptualization, B.E.S. and N.B.T-B.; Methodology, B.E.S. and N.B.T-B.; Investigation, B.E.S.; Formal Analysis, B.E.S.; Writing, B.E.S. and N.B.T-B.; Supervision, N.B.T-B.; Funding Acquisition, B.E.S. and N.B.T-B.

Acknowledgments

This work was supported by NSF GRFP to B.E.S., as well as NIH R01 MH069456, NSF CCF 1839308, and the Canadian Institute for Advanced Research to N.B.T-B.

Declaration of Interests

The authors declare no competing interests.

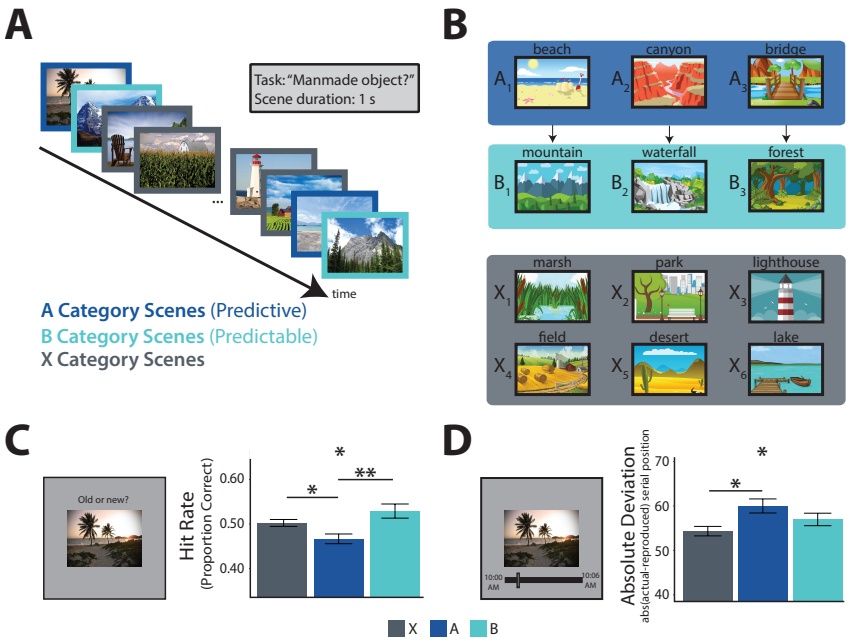

Figure 1: Behavioral Experiments. A) Task design: participants viewed a continuous stream of scene pictures, during which they made a judgment of whether or not there was a manmade object in the scene. B) Example scene category pairings for one participant: 3 of 12 categories were assigned to condition A; each was reliably followed by one of 3 different categories assigned to condition B (illustrated by arrows). The remaining 6 categories were assigned to condition X and were not consistently preceded or followed by any particular category. C) Left: surprise recognition memory test. Right: proportion of old exemplars recognized as a function of condition (higher hit rate is better memory). D) Left: temporal source memory test. Right: absolute difference between reported and actual time of encoding as a function of condition (higher deviation is worse memory). Error bars reflect within-participant standard error of the mean. *p <0.05, **p <0.01.

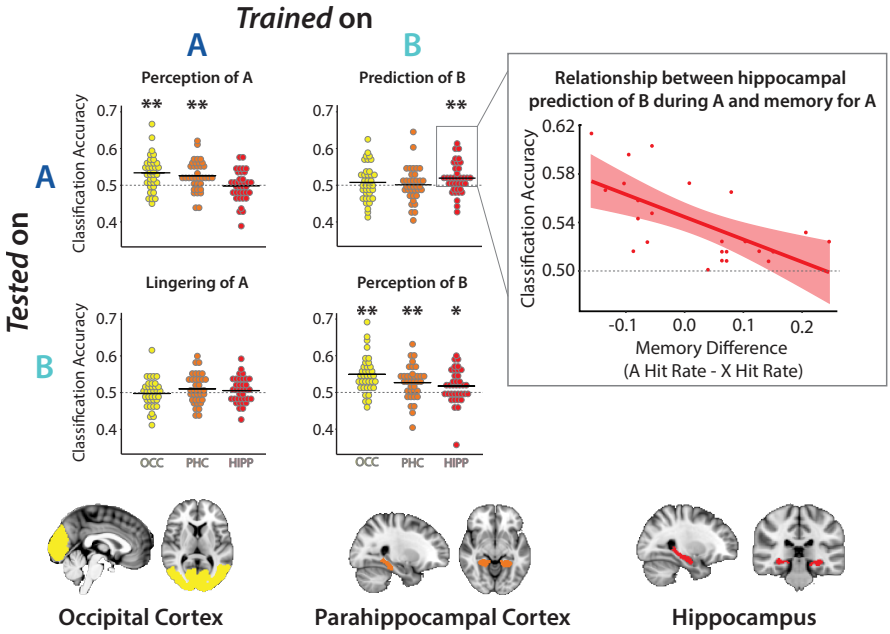

Figure 2: Category Decoding in fMRI Experiment. Upper left: Classification accuracy in occipital cortex (OCC), parahippocampal cortex (PHC), and hippocampus (HIPP) plotted for each of the four combinations of training or testing on A or B categories. For every A/B combination and ROI, each dot is one participant and the black line is the mean across participants. Perception of A: OCC: $t(35) = 4.34$, $p < 0.001$; PHC: $t(35) = 3.82$, $p < 0.001$; HIPP: $t(35) = -0.17$, $p = 0.87$. Prediction of B: OCC: $t(35) = 0.94$, $p = 0.35$; PHC: $t(35) = 0.17$, $p = 0.87$; HIPP: $t(35) = 2.73$, $p = 0.0098$. Lingering of A: OCC: $t(35) = -0.38$, $p = 0.71$; PHC: $t(35) = 1.57$, $p = 0.13$; HIPP: $t(35) = 0.96$, $p = 0.34$. Perception of B: OCC: $t(35) = 5.96$, $p < 0.001$; PHC: $t(35) = 3.52$, $p = 0.0012$; HIPP: $t(35) = 2.26$, $p = 0.030$. Bottom: Regions of interest. HIPP and PHC were manually segmented in native participant space (transformed into standard space for visualization); OCC was defined in standard space and transformed into native participant space. Upper right: Correlation between "Prediction of B" classification accuracy in HIPP and the difference in hit rate between A and X. Only participants with classification higher than chance (>0.5) were included.

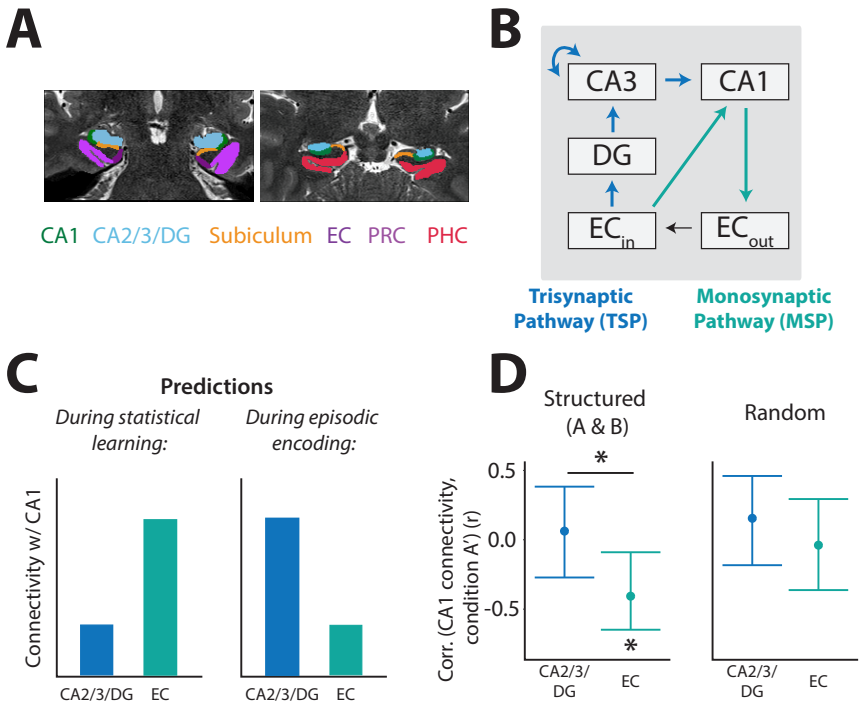

Figure 3: Hippocampal Connectivity in fMRI Experiment. A) Manually segmented hippocampal subfield and MTL cortical ROIs (anterior and posterior slices from representative participant shown): CA1, combined CA2/3/DG, subiculum, entorhinal cortex (EC), perirhinal cortex (PRC), and parahippocampal cortex (PHC). B) Two pathways in the hippocampal circuit: trisynaptic pathway (TSP, blue) and monosynaptic pathway (MSP, teal). C) Pathway-specific predictions about CA1 functional connectivity during statistical learning (left) and episodic memory (right). D) Relationship of CA1 functional connectivity in TSP (CA2/3/DG) and MSP (EC) with episodic memory performance (A′) across participants, separated by whether there were regularities to be extracted by statistical learning (Structured) or not (Random). Error bars represent 95% confidence intervals. *p <0.05.

Methods

Experiment 1

Participants. Thirty individuals (19 female; age range: 18-31, mean age = 21.2) were recruited from the Yale University community for either course credit or $10 compensation. Informed consent was obtained in a manner approved by the Yale University Human Subjects Committee.

Stimuli and Apparatus. Participants were seated approximately 50cm away from a 69cm monitor (1920 x 1080 pixel resolution; 60 Hz refresh rate). Scene stimuli consisted of 300 unique scene images drawn from 12 scene subcategories (25 images/subcategory), collected from Google image searches. Each participant viewed 22 scenes from each subcategory, randomly selected from the set of 25. Sixteen of these images (per each category; 192 total) were used in the learning phase, two for the category pair test, and four as foils in the recognition test. Scene stimuli were presented centrally and subtended 27.8 x 20.8 degrees of visual angle. Stimuli were presented using MATLAB (The MathWorks, Natick, MA) with the Psychophysics Toolbox (Brainard and Vision, 1997; Pelli and Vision, 1997).

Procedure. Participants first completed a learning phase. On each trial, they viewed a photograph of a scene for 1000 ms, during which they had to respond based on whether it contained a manmade object (See Main Text Figure 1A). Participants were instructed to respond as quickly and accurately as possible (response mappings of 'j'/'k' onto 'yes'/'no' were counterbalanced across participants), and we recorded response time (RT) and accuracy. The scene remained on the screen for 1000 ms regardless of button press to equate encoding time, and trials were separated by a 500-ms inter-stimulus interval during which a fixation cross appeared.

Every scene was trial-unique, but was drawn from one of 12 outdoor scene categories (beaches, bridges, canyons, deserts, fields, forests, lakes, lighthouses, marshes, mountains, parks, and waterfalls; see Main Text Figure 1B). Each

scene category appeared 16 times over the course of the learning phase, for a total of 192 trials. The photographs for half of the scene categories always contained a manmade object, and thus all exemplars in a category required the same response, and the responses were balanced overall. Unbeknownst to participants, and orthogonal to the required response, half of the scene categories were assigned to pairs. Given the first scene in a pair (A category scenes), the category of the second scene (B category scenes) was predictable with a transition probability of 1.0. The other half of scene categories were neither predictive nor predictable (X category scenes). Pictures from these categories were inserted on their own randomly. The assignment of scene categories to A/B/X conditions was itself randomized for each participant. The order of the photograph sequence was randomized with the following constraints: category pairs and pairs of category pairs could not repeat back-to-back (i.e., no A1/B1/A1/B1 or A1/B1/A2/B2/A1/B1/A2/B2); repetitions of each category were spread equally across quartiles of the learning phase to minimize differences in study-test lag between categories; and the overall transition probability between "yes" and "no" responses was forced to be statistically indistinguishable from 0.5.

After the learning phase, participants performed five minutes of a distracting math phase to minimize recency effects. Each of 60 math problems consisted of division and subtraction, and the answer to the problem was always 1, 2, 3, or 4. Participants responded using the 1, 2, 3, and 4 keys on the keyboard, with a maximum response window of 5 s. The inter-stimulus interval was adjusted based on the RT (5 s - RT), to ensure that this phase lasted exactly 5 min given the 60 trials. Participants were instructed to respond as accurately as possible.

Participants then underwent two surprise memory tests (category pair test and episodic memory test), the order of which was counterbalanced across participants. The category pair test involved explicit judgments of the category pairings from the learning phase. Participants were presented with two pairs of photographs on every test trial and were asked to indicate which pair felt more familiar based only what they had seen during the learning phase. The

pairs were shown sequentially: the first scene from one pair appeared for 1000 ms, followed by a 500-ms blank interval, followed by the second scene of the pair for 1000 ms; after a 1000 ms gap with a fixation cross, a second pair was presented in the same manner. After both pairs, participants responded using the '1' key to indicate if the first pair felt more familiar or the '2' key if the second pair felt more familiar. Participants had a maximum of 6 s to respond. Each scene in the category pair test was a completely novel exemplar of its category. Half of the test trials contained a true category pair (when it was a trial testing a pair from the learning phase); whether it appeared first or second was counterbalanced. The other half of the trials contained a "dummy coded" pair of the X categories (there was no correct answer on these trials). This was done to equate the frequency of categories, which was important for participants who received the category pair test before the episodic memory test. Each true/dummy-coded pair was tested twice against a scrambled pair of the same categories (e.g., if beach → field, mountain → bridge, canyon → forest were category pairs from the learning phase, the foils might be beach → bridge, mountain → forest, canyon → field).

The episodic memory test was designed to assess episodic memory for the trial-unique scenes from the learning phase. On each trial, one scene was presented and participants had to indicate whether it was "old" (i.e., presented during the learning phase) or "new" (i.e., not previously seen in the experiment). After making an old/new response (using 'j'/'k' keys on the keyboard), participants then rated their confidence in this response ("not confident"/"confident", using 'd'/'f' keys). Participants had 6 s to make each response. All 192 scene photographs from the learning phase were shown, in addition to 48 foils (4 novel exemplars from each category). The order of the scenes was randomized.

Experiment 2

Participants. Thirty individuals (19 female; age range: 18-23, mean age = 19.3) were recruited from the Yale University community for either course credit or $10 compensation. Informed consent was obtained in a manner approved by

the Yale University Human Subjects Committee.

Stimuli and Apparatus. Same as Experiment 1.

Procedure. The procedure was identical to that of Experiment 1, with the following exception: we replaced the confidence judgment of the episodic memory test with a temporal source judgment. In other words, participants were presented with a scene and again asked to judge whether it was "old" or "new" (using the 'd' and 'f' keys). However, instead of then being asked to judge how confident they were in their response, old responses were followed by the presentation of a timeline, bound by the start and end clock times of the learning phase. Participants used the mouse to click along the timeline to indicate when they remember seeing the scene. No temporal source judgments were collected after new responses.

Experiment 3

Participants. Thirty-eight individuals (24 female; age range: 18-35, mean age = 23.1) were recruited from the Yale University community for $30 compensation. Informed consent was obtained in a manner approved by the Yale University Human Investigation Committee. One participant was excluded to due a neurological anomaly, and one participant was excluded for chance-level episodic memory performance ($A' < 0.5$).

Stimuli and Apparatus. Stimuli were presented on a rear-projection screen using a projector (1920 x 1068 pixel resolution; 60 Hz refresh rate). Participants viewed the stimuli through a mirror mounted on the head coil. Scene stimuli consisted of 480 unique images drawn from 12 subcategories (40 images/subcategory), collected from Google image searches (180 additional stimuli were collected for this experiment; the other 300 are identical to those used in Experiments 1 & 2). Each participant viewed 39 scenes from every subcategory, randomly selected from the set of 40: 21 per category (252 total) for the learning phase, 14 (168) for the pre- and post- learning templating phases, and 4 (48) as foils in the episodic memory test.

Procedure. The procedure was identical to that of Experiment 1 with the following changes:

Instead of one continuous block of the learning phase (with 16 repetitions of each scene category), it was divided into three fMRI runs, each with seven repetitions/category (such that were 21 repetitions/category in total across the learning phase). As in Experiments 1 & 2, each image was presented for 1 s, but the ITI in this experiment varied between 2 s (39.3% of trials), 3.5 s (39.3% of trials), and 5 s (21.4% of trials). For the manmade object cover task, participants responded using their right index and middle fingers on an MR-compatible button box.

Before and after the three learning runs, respectively, there were "pre" and "post" templating phases (one fMRI run each). To participants, these phases were identical to the learning phase (e.g., stimulus timing and task were identical). However, there were no category-level regularities in these two runs. Scenes from all categories were presented in a random order. To limit the impact of this random presentation on subsequent learning, participants completed a distracting math task between the "pre" templating run and the first learning run. Each of these five functional runs (three learning runs and pre/post runs) lasted 6.4 minutes.

For the episodic memory test, as in Experiment 1, a scene was presented and participants indicated (with their index and pinky fingers) whether it was "old" (i.e., presented during the learning phase) or "new" (i.e., not previously seen in the experiment). They then rated their confidence in this response ("very unsure","unsure","sure","very sure"), using their index through pinky fingers, respectively. Participants had 6 s to respond to each of these questions. They completed this task while in the scanner, but no fMRI data were collected. No category pair test was administered in this experiment.

MRI Acquisition. Data were acquired on a Siemens Prisma 3T scanner using a 64-channel head coil at the Magnetic Resonance Research Center at Yale University. Functional images were acquired using an EPI sequence with the

following parameters: TR = 1500 ms; TE = 32 ms; 90 axial slices; voxel size = 1.5 x 1.5 x 1.5 mm; flip angle = 64 degrees; multiband factor = 6. Additionally, a pair of opposite phase-encode spin-echo volumes were collected for distortion correction (TR = 11,220 ms; TE = 66 ms). One T1-weighted MPRAGE (TR = 1800 ms; TE = 2.26 ms; voxel size = 1 x 1 x 1 mm; 208 sagittal slices; flip angle = 8 degrees) and two T2-weighted turbo spin echo (TR = 11,390 ms; TE = 90 ms; 54 coronal slices; voxel size = 0.44 x 0.44 x 1.5 mm; distance factor = 20%; flip angle = 150 degrees) anatomical images were collected.

fMRI Preprocessing. fMRI data processing was carried out using FEAT (fMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl) version 5.0.10. EPI and anatomical images were skull-stripped using the Brain Extraction Tool (Smith, 2002). Susceptibility-induced distortions measured via the opposing-phase spin echo volumes were corrected using FSL's topup tool (Andersson et al., 2003). Each functional run was high-pass filtered with a 128 s period cut-off, corrected for head motion using MCFLIRT (Jenkinson et al., 2002), and motion outliers were computed. Slice-timing correction was performed. No spatial smoothing was applied. Lastly, the six motion parameters, as well as motion outliers, were regressed against the BOLD timecourse using a general linear model (GLM). The residuals from this preprocessing model (which contain BOLD responses to the task after controlling for motion) were then used for subsequent analyses.

Functional images were registered to each participant's T1 anatomical scan using boundary-based registration, as well as to a 2 mm MNI standard brain, using 12 degrees of freedom. Lastly, the two T2 anatomical images collected were registered to one another and averaged; the resulting averaged image was registered to the T1 anatomical image using FLIRT (Jenkinson and Smith, 2001).

Regions of Interest. Our primary region of interest (ROI) was the hippocampus. Hippocampal subfields CA1, CA2/3/DG, and subiculum, as well as medial temporal lobe (MTL) cortical regions entorhinal cortex (EC), perirhinal cortex

(PRC), and parahippocampal cortex (PHC) were manually segmented on participants' averaged T2 anatomical scan, using published criteria on anatomical landmarks (Insausti et al., 1998; Pruessner et al., 2002; Duvernoy, 2005; Aly and Turk-Browne, 2015). Corresponding ROIs from each hemisphere were concatenated to create one bilateral ROI per region, as we had no hemisphere-specific hypotheses. Additionally, a bilateral whole hippocampus ROI was created by concatenating the three bilateral subfield ROIs. Individual ROIs were transformed into each participant's functional space for subsequent analyses.

The occipital cortex ROI was defined using the MNI structural atlas, thresholded at 25% probability. This standard space ROI was then transformed into each participant's functional space for subsequent analyses.

Category Decoding Analysis. A multivariate pattern classification approach was used to assess evidence for a particular category during the learning phase. This approach involved training a classifier on fMRI activity patterns for each category from the "pre" templating run (when there were no regularities present and none had been learned) and testing for classifier evidence of these categories during the three (independent) learning runs.

More specifically, for each functional run, the residuals from the preprocessing GLM (with known noise sources removed but still containing task responses) were aligned to the final functional run and z-scored across time. The voxel x time matrices were then masked to only include voxels within an ROI. The timepoints corresponding to the presentation of each of two categories of interest were extracted and shifted by 3 TRs (4.5 seconds) to account for the hemodynamic lag. The voxel activity patterns from these shifted time points were then extracted and used as training or test data for the classifier. Timepoints that included a motion outlier were excluded from the training/test sets.

Linear SVMs were trained on data and labels from the pre-learning templating run, using the SVC function in Python's scikit-learn module, with a penalty parameter of 1.00. Classifiers were then tested with data corresponding to the timepoints of the trained categories in the three learning runs (concatenated)

and made guesses as to the category label of each test example. Accuracy was computed as the proportion of correct guesses.

We ran the following comparisons: Perception of A (training on pre-learning examples of A, testing for evidence of A during the presentation of A in the learning phase), Perception of B (training on pre-learning examples of B, testing for evidence of B during the presentation of B in the learning phase), Lingering of A (training on pre-learning examples of A, testing for evidence of A during the presentation of B in the learning phase), and Prediction of B (training on pre-learning examples of B, testing for evidence of B during the presentation of A in the learning phase).

Each participant encountered three A and three B categories over the course of the experiment. Thus, for each of the four comparisons above, we built three different binary classifiers and then averaged their accuracy. In other words, a classifier was trained to distinguish between two scene categories from the same condition (e.g., B) based on the pre-learning templating run, and tested for evidence of those two categories during the subsequent presentation of two categories (e.g., their corresponding As for Prediction of B) in the learning phase. Accuracy (percent correct) was then computed for each of these three classifiers (A1 vs. A2; A2 vs. A3; A1 vs. A3) and averaged, resulting in one mean accuracy value per comparison, per participant.

To provide an example, if we use the category pairs from earlier beach $\rightarrow$ field, mountain $\rightarrow$ bridge, canyon $\rightarrow$ forest, then B classifiers would be trained for field vs. bridge, bridge vs. forest, and field vs. forest. To calculate evidence for Prediction of B: the field vs. bridge classifier would be applied to the beach and mountain trials — such that the classifier estimated evidence for field and bridge during each beach or mountain trial — and accuracy was computed (such that, for example, accuracy on a beach trial was 1 if the classifier outputted more evidence for field than for mountain). This was repeated for the bridge vs. forest classifier (testing for evidence of these categories during mountains and canyons) and the field vs. forest classifier (testing for evidence of these categories during beaches and canyons). The accuracies of these three classifiers were averaged

into a single accuracy for each participant. This was repeated for the three other comparisons above and for each ROI. To assess reliability at the group level, performance was compared to a chance level of 0.50 across participants using a one-sample t-test.

PPI Analysis. To examine how functional connectivity between hippocampal subfields is affected by statistical learning, we conducted a psychophysiological interaction (PPI) analysis. Because we aimed to assess the impact of learned regularities, we limited this analysis to the second and third runs of the learning phase. Including the first learning run may have weakened the key contrast of connectivity during periods with and without regularities, as participants had no knowledge of the regularities at the beginning of that run.

To perform the PPI, we concatenated the aligned, normalized residual timecourses from the pre-processing GLM across the two included runs. We then averaged the activity of voxels in each ROI to compute a mean timecourse for CA1, CA2/3/DG, and EC. Additionally, we extracted the onsets of Structured (A & B) and Random (X) pictures, and convolved each of these two condition regressors with a double-gamma hemodynamic response function (fmrisim function in BrainIAK, http://brainiak.org). We then ran a GLM in R (https://cran.r-project.org/), in which we predicted the timecourse of CA1 as a linear combination of regressors for the timecourses of CA2/3/DG and EC, task events in the two learning conditions (Structured and Random), and the interaction between ROIs and learning conditions. The interaction regressors were defined as the products of the ROI and condition timecourses (EC*Structured, EC*Random, CA2/3/DG*Structured, CA2/3/DG*Random). Each regressor, as well as the CA1 timecourse, was z-scored and entered simultaneously into the model. A separate model was run for each participant, resulting in one coefficient per regressor per participant. For an interaction regressor of interest, the coefficients were correlated with A' memory performance for the corresponding condition across participants.

Searchlight Analysis. To explore the specificity of our Prediction of B results in the brain, we performed the category decoding analysis within a 27-voxel cube that was moved across all functional voxels (searchlight function in BrainIAK). Each aligned, normalized residual volume from the pre-processing GLM was registered to standard space. These volumes were masked for each searchlight and the retained voxels were subjected to the same decoding pipeline described above for the ROIs. The result was a searchlight map per participant, in which the value at each voxel reflected the average classification accuracy for the cube centered at that voxel. The reliability of these maps was assessed at the group level using nonparametric randomization tests (randomise function in FSL) (Winkler et al., 2014), corrected for multiple comparisons using threshold-free cluster enhancement (Smith and Nichols, 2009). As a control analysis, we ran the same searchlight procedure for the Perception of B comparison.

Data and Code Availability. fMRI data can be downloaded from OpenNeuro and behavioral data can be downloaded from Dryad. Analysis code can be accessed on Github.

References

Aly, M., Turk-Browne, N.B., 2015. Attention stabilizes representations in the human hippocampus. Cerebral Cortex 26, 783–796.

Aly, M., Turk-Browne, N.B., 2017. How hippocampal memory shapes, and is shaped by, attention. Springer. pp. 369–403.

Andersson, J.L., Skare, S., Ashburner, J., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20, 870–888.

Brady, T.F., Oliva, A., 2008. Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. Psychological Science 19, 678–685.

Brainard, D.H., Vision, S., 1997. The psychophysics toolbox. Spatial Vision 10, 433–436.

Brown, M.W., Aggleton, J.P., 2001. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? Nature Reviews Neuroscience 2, 51.

Cohen, J.D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K.A., Pillow, J., Ramadge, P.J., Turk-Browne, N.B., et al., 2017. Computational approaches to fmri analysis. Nature Neuroscience 20, 304.

Davachi, L., DuBrow, S., 2015. How the hippocampus preserves order: the role of prediction and context. Trends in Cognitive Sciences 19, 92–99.

Davachi, L., Mitchell, J.P., Wagner, A.D., 2003. Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. Proceedings of the National Academy of Sciences 100, 2157–2162.

De Brigard, F., Brady, T.F., Ruzic, L., Schacter, D.L., 2017. Tracking the emergence of memories: A category-learning paradigm to explore schema-driven recognition. Memory & Cognition 45, 105–120.

Duncan, K., Sadanand, A., Davachi, L., 2012. Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. Science 337, 485–487.

Duvernoy, H.M., 2005. The human hippocampus: functional anatomy, vascularization and serial sections with MRI. Springer Science & Business Media.

Frankland, P.W., Bontempi, B., 2005. The organization of recent and remote memories. Nature Reviews Neuroscience 6, 119.

Greve, A., Cooper, E., Kaula, A., Anderson, M.C., Henson, R., 2017. Does prediction error drive one-shot declarative learning? Journal of Memory and Language 94, 149–165.

Hasselmo, M.E., Wyble, B.P., Wallenstein, G.V., 1996. Encoding and retrieval of episodic memories: role of cholinergic and gabaergic modulation in the hippocampus. Hippocampus 6, 693–708.

Henson, R.N., Gagnepain, P., 2010. Predictive, interactive multiple memory systems. Hippocampus 20, 1315–1326.

Insausti, R., Juottonen, K., Soininen, H., Insausti, A.M., Partanen, K., Vainio, P., Laakso, M.P., Pitkänen, A., 1998. Mr volumetric analysis of the human entorhinal, perirhinal, and temporopolar cortices. American Journal of Neuroradiology 19, 659–671.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical Image Analysis 5, 143–156.

Kamin, L., 1969. Predictability, surprise, attention, and conditioning. pp. 279–296.

Kruschke, J.K., 2001. Toward a unified model of attention in associative learning. Journal of Mathematical Psychology 45, 812–863.

Kumaran, D., Maguire, E.A., 2006. An unexpected sequence of events: mismatch detection in the human hippocampus. PLoS biology 4, e424.

Leutgeb, J.K., Leutgeb, S., Moser, M.B., Moser, E.I., 2007. Pattern separation in the dentate gyrus and ca3 of the hippocampus. Science 315, 961–966.

Mackintosh, N.J., 1975. A theory of attention: variations in the associability of stimuli with reinforcement. Psychological Review 82, 276.

McClelland, J.L., McNaughton, B.L., O'Reilly, R.C., 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review 102, 419.

Miller, J.F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., Hefft, S., Merkow, M., Polyn, S.M., Jacobs, J., et al., 2013. Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. Science 342, 1111–1114.

Mitchell, K.J., Johnson, M.K., 2009. Source monitoring 15 years later: what have we learned from fmri about the neural mechanisms of source memory? Psychological Bulletin 135, 638.

Norman, K.A., O'Reilly, R.C., 2003. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychological Review 110, 611.

Patil, A., Duncan, K., 2018. Lingering cognitive states shape fundamental mnemonic abilities. Psychological Science 29, 45–55.

Pelli, D.G., Vision, S., 1997. The videotoolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision 10, 437–442.

Pruessner, J.C., Köhler, S., Crane, J., Pruessner, M., Lord, C., Byrne, A., Kabani, N., Collins, D.L., Evans, A.C., 2002. Volumetry of temporopolar, perirhinal, entorhinal and parahippocampal cortex from high-resolution mr images: considering the variability of the collateral sulcus. Cerebral Cortex 12, 1342–1353.

Richards, B.A., Xia, F., Santoro, A., Husse, J., Woodin, M.A., Josselyn, S.A., Frankland, P.W., 2014. Patterns across multiple memories are identified over time. Nature Neuroscience 17, 981.

Schacter, D.L., Benoit, R.G., Szpunar, K.K., 2017. Episodic future thinking: Mechanisms and functions. Current Opinion in Behavioral Sciences 17, 41–50.

Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., Norman, K.A., 2017. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning.

Philosophical Transactions of the Royal Society B: Biological Sciences 372, 20160049.

Shiffrin, R.M., Atkinson, R.C., 1969. Storage and retrieval processes in long-term memory. Psychological Review 76, 179.

Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17, 143–155.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83–98.

Smith, T.A., Hasinski, A.E., Sederberg, P.B., 2013. The context repetition effect: Predicted events are remembered better, even when they don't happen. Journal of Experimental Psychology: General 142, 1298.

Squire, L.R., 2004. Memory systems of the brain: a brief history and current perspective. Neurobiology of Learning and Memory 82, 171–177.

Tompary, A., Davachi, L., 2017. Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. Neuron 96, 228–241.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.