Prediction Errors Disrupt Hippocampal Representations and Update Episodic Memories

*Alyssa H. Sinclair[1,2] Grace M. Manalili[2], Iva K. Brunec[3,4],
R. Alison Adcock[1], & Morgan D. Barense[2,5]

[1] Duke University, *Department of Psychology & Neuroscience*, Durham, NC, 27705, USA
[2] University of Toronto, *Department of Psychology*, Toronto, ON, M5S 3G3, Canada
[3] Temple University, *Department of Psychology*, Philadelphia, PA, 19122, USA
[4] University of Pennsylvania, *Department of Psychology*, Philadelphia, PA, 19104, USA
[5] Baycrest Hospital, *Rotman Research Institute*, Toronto, ON, M64 1W1, Canada

Keywords: prediction error, memory, hippocampus, basal forebrain, acetylcholine, reconsolidation, cognitive neuroscience, fMRI

* Correspondence: allie.sinclair@duke.edu

Significance

Our brains draw on memories to predict the future; when our predictions are incorrect, we must modify our memories to improve future predictions. Past studies have demonstrated that the hippocampus signals *prediction error*, or surprise, but have not linked this neural signal to memory updating. Here, we uncover this missing connection: We show that prediction errors change the role of the hippocampus, reversing the relationship between hippocampal activation and memory outcomes. We examine the mechanisms of this shift in neural processing, showing that prediction errors disrupt the stability of hippocampal patterns. We propose that prediction errors shift attention away from internal predictions toward external stimuli. Our findings bear implications for improving education, understanding eyewitness memory distortion, and treating pathological memories.

Abstract

The brain supports adaptive behavior by generating predictions, learning from errors, and updating memories to accommodate new information. *Prediction error,* or surprise, triggers learning when reality contradicts expectations. Prior studies have shown that the hippocampus signals prediction errors, but have never linked this neural signal to memory updating. Here, we uncover new mechanisms that link hippocampal prediction error signals to memory updating. In a human fMRI study, we elicited mnemonic prediction errors by interrupting familiar narrative videos immediately before the expected endings. We found that the same amount of hippocampal activation could exert opposing effects on memory: hippocampal activation preserved memories after expected endings, but updated memories after prediction errors. In contrast to previous studies, we showed that univariate activation was insufficient for understanding hippocampal prediction error signals. We explained this surprising finding by tracking the evolution of hippocampal activation patterns, and connectivity between the hippocampus and neuromodulatory regions. We found that hippocampal activation patterns stabilized as each narrative episode unfolded, sustaining episodic representations. Prediction errors disrupted these sustained representations, and the degree of disruption predicted memory updating. The relationship between hippocampal activation and subsequent memory depended on concurrent basal forebrain activation, providing new evidence about how cholinergic modulation may regulate attention and memory. We conclude that prediction errors create conditions that favor memory updating, prompting the hippocampus to shift attention externally and encode new information.

Introduction

In daily life, we continuously draw on past experiences to predict the future. Expectation and surprise shape learning across many situations, such as when we encounter misinformation in the news, receive feedback on an exam, or make decisions based on past outcomes. When our predictions are incorrect, we must update our mnemonic models of the world to support adaptive behavior. *Prediction error* is a measure of the discrepancy between expectation and reality; this surprise signal is both evident in brain activity and related to learning success (1–6). The brain dynamically constructs memories during recall, recreating and revising past experiences based on new information (7). The intuitive idea that surprise governs learning has long shaped our understanding of memory, reward learning, perception, action, and social behavior (2, 8–14). Yet, the neural mechanisms that allow prediction error to update memories remain unknown.

Past research has implicated the hippocampus in the mnemonic functions required for learning from prediction errors: retrieving memories to make predictions, identifying discrepancies between past and present, and encoding new information (2, 15–20). Functional MRI (fMRI) studies have shown that hippocampal activation increases after predictions are violated; this surprise response has been termed *mismatch detection* (18, 19, 21, 22), or *mnemonic prediction error* (20). Several theoretical frameworks have hypothesized that this hippocampal prediction error signal could update memories (17, 20, 23–26). Although past studies have shown that the hippocampus *detects* prediction errors, the impact of this surprise signal on memory—a crucial link for understanding how we learn from error—has been implied, but not yet demonstrated.

What mechanisms could link hippocampal prediction errors to memory updating? A leading hypothesis is that prediction errors shift the focus of attention by switching cognitive

processing modes (20, 27–31). After prediction errors, we should reset our expectations and shift towards an *externally-oriented* processing mode that supports updating memories with relevant new information. After episodes that align with expectations, we should generate ongoing predictions and shift towards an *internally-oriented* processing mode that sustains and reinforces existing memories. Consistent with this idea, mnemonic prediction errors have been shown to enhance the perceptual input pathway in the hippocampus (32), but suppress the output pathway (20). Surprising events may upregulate external attention (the input pathway), thus preparing the brain to learn from new information. However, the idea that updating memories involves shifting the balance between internal and external processing modes has yet to be tested.

Neuromodulation may be a critical factor that regulates internal and external processing modes, thus allowing the hippocampus to update memories. Currently, there is mixed evidence supporting two hypotheses: acetylcholine and/or dopamine could act upon the hippocampus to regulate inputs after surprise (23–26, 28, 30, 33). Several models have proposed that acetylcholine from the medial septum (within the basal forebrain) regulates the balance between input and output pathways in the hippocampus (26–28, 34–37), thus allowing stored memories to be compared with perceptual input (30, 37, 38). After prediction errors, acetylcholine release could bias the hippocampus towards external processing and enhance encoding of new information (25, 28, 33, 36, 38). On the other hand, dopamine released from the ventral tegmental area (VTA), if transmitted to the hippocampus, could also modulate hippocampal plasticity after prediction errors. Past studies have shown that dopamine upregulates the hippocampal input pathway (39), the hippocampus and VTA are co-activated after surprising events (40, 41), and this co-activation enhances memory encoding and integration (42–45). Overall, the basal forebrain and the VTA are both candidate mechanisms for regulating

hippocampal processing modes after prediction errors, but no past studies have directly tested these predictions. Understanding specific neuromodulatory mechanisms is important for developing interventions, such as for counteracting pathological fear memories.

A separate body of research has also investigated how memories can be updated. Animal and human research on *reconsolidation* (46–49), the process by which memory traces can be reactivated and temporarily destabilized, has shown that surprising reminders enable memory updating (1, 3, 50). Paralleling research on internal and external processing modes, both acetylcholine and dopamine may contribute to reconsolidation (50, 51). Reconsolidation paradigms have generated surprise by imperfectly replicating encoding experiences (e.g., presenting a conditioned stimulus without the expected outcome) (1, 3, 52, 53). However, past reconsolidation studies have not measured neural prediction error signals, such as hippocampal responses. Reconsolidation paradigms are optimized to update memories; by drawing on these behavioral methods to study brain function, we sought to identify the missing link between hippocampal prediction error signals and memory updating.

Our approach diverged from past studies in several ways, allowing us to test previously unanswered questions. First, to link hippocampal prediction error signals to memory updating, we used a reconsolidation-inspired paradigm to transform naturalistic episodic memories. Second, we investigated hippocampal processing modes by tracking how episodic representations are sustained or disrupted over time, going beyond univariate measures of activation. Third, we tested hypotheses about neuromodulatory mechanisms by relating activation in the basal forebrain and VTA to hippocampal processing modes.

In our human fMRI study, we examined trial-wise hippocampal responses to prediction errors during narrative videos. During the Day 1 encoding session, participants viewed 70 full-

length videos that featured narrative events with salient endings (e.g., a baseball batter hitting a home run) (Figure 1A). During the Day 2 reactivation session, participants watched the videos again (Figure 1B). We elicited mnemonic prediction errors by interrupting half of the videos immediately before the expected narrative ending (e.g., the video ends while the baseball batter is mid-swing). These surprising interruptions were comparable to the prediction errors employed in reconsolidation studies (1). Half of the videos were presented in Full-length form (identical to the encoding session), and half were presented in Interrupted form (eliciting prediction error). Reconsolidation group participants (n = 24) completed the Day 2 session while undergoing an fMRI scan, whereas Immediate control group participants (n = 24) completed the study in a behavioral testing room and were not scanned. Our primary fMRI analyses examined the period immediately following the offset of Full and Interrupted videos (Post-Video fixation period) during the Day 2 session in the Reconsolidation group. Importantly, this design compares neural responses to surprising and expected video endings while controlling for visual and auditory input.

Lastly, participants completed a memory test in the form of a structured interview (Figure 1C). On each trial, participants were cued with the name of the video and recalled the narrative. The experimenter then probed for further details with pre-determined questions (e.g., "Can you describe the baseball batter's ethnicity, age range, or clothing?"). Our critical measure of memory updating was *false memories*. Although it can be adaptive to update real-world memories by incorporating relevant new information, we expected that our laboratory paradigm would induce false memories because participants would integrate interfering details across similar episodes (1, 7). Because we were interested in false memories as a measure of memory updating, we instructed participants not to guess and permitted them to skip details they could

not recall. In the Reconsolidation group, participants completed the memory test on Day 3, 24 hours after the reactivation session. In the Immediate control group, participants completed the memory test on Day 2, immediately after the reactivation session (Figure 1D). Reconsolidation theory states that updating memories requires a delay, because re-stabilizing a memory trace involves hours of protein synthesis (49, 54). Therefore, the Immediate control group should not exhibit any memory effects that require protein synthesis-dependent reconsolidation.

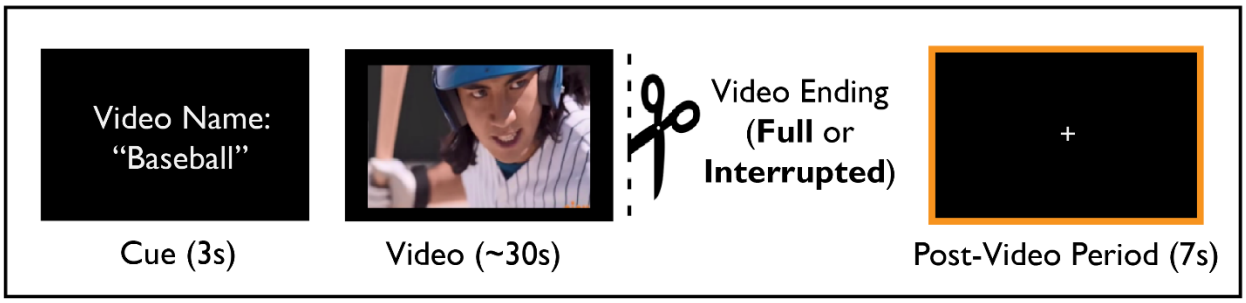
Using these novel methods, we found the following: 1) Prediction errors selectively updated memories in the Reconsolidation group. 2) In contrast to past studies, we found that prediction errors reversed the effect of hippocampal activation on memory: After surprising endings, hippocampal activation was associated with memory updating, but after expected endings, hippocampal activation was associated with memory preservation. 3) Hippocampal activation patterns stabilized during and after videos, but prediction errors disrupted these sustained representations. 4) After prediction errors, disrupting hippocampal patterns led to memory updating. 5) The effect of prediction error on memory depended on co-activation of the hippocampus and basal forebrain, supporting the idea that acetylcholine regulates attention and memory.

A

Encoding Phase: Example Stimulus Video

B

Reactivation Phase: Example Trial



C

Test Phase: Example Memory Test

Experimenter: The next video is "Baseball." Can you describe the main event of the video?
Participant: Okay, so they're **in a stadium,** and there are **lots of people watching.** The **pitcher throws the ball** and the batter **hits it out of the park**.
Experimenter: Can you describe the baseball batter? Age range, hair color, ethnicity, or clothing?
Participant: He looked **East Asian,** in his **mid-40s.** He was **wearing a red uniform**.
Experimenter: Do you remember hair color?
Participant: No, I don't remember.

Legend: **Correct Details** **False Memories**

D

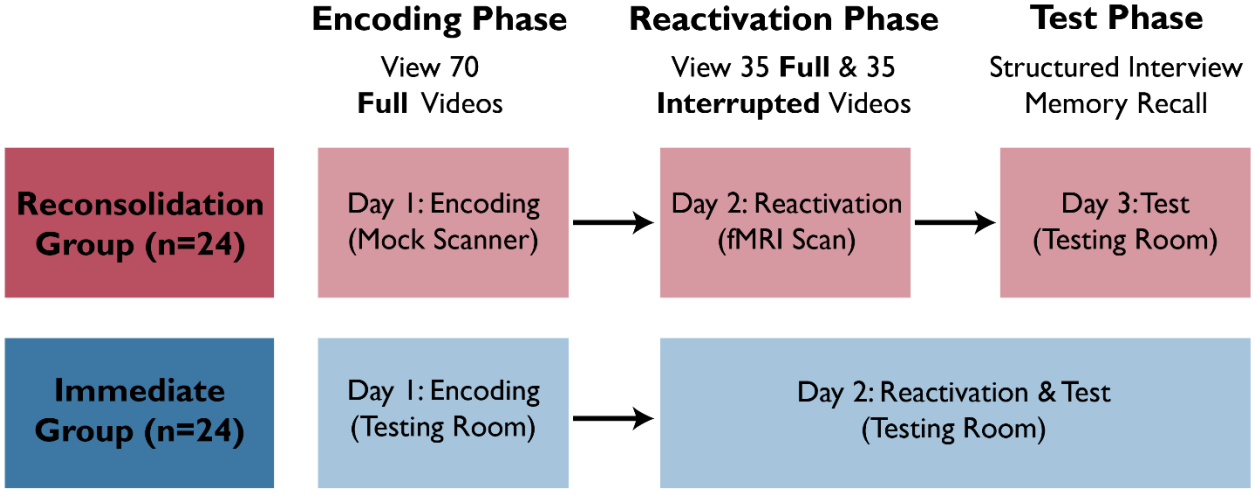


Figure 1. Overview of experimental paradigm. A) During the Day 1 Encoding session, all videos were presented in Full-length form. Here we show frames from a stimulus video named "Baseball", depicting a home run. B) During the Day 2 Reactivation session, participants viewed the videos again, but half were interrupted to elicit prediction error. Participants were cued with the video name, watched the video (Full or Interrupted), and then viewed a fixation screen. The "Baseball" video was interrupted when the batter was mid-swing. fMRI analyses focused on the Post-Video fixation periods after each video (highlighted box). Thus, visual and auditory stimulation were matched across Full and Interrupted conditions, allowing us to compare Post-Video neural activation while controlling for perceptual input. C) During the Test session, participants answered structured interview questions about all 70 videos, and were instructed to answer based on their memory of the Full video originally shown during encoding. Here we show example text illustrating the memory test format and scoring of correct details and false memories. The void response ("I don't remember") is not counted as a false memory. D) Overview of the experiment. All participants completed Encoding, Reactivation, and Test Phases of the study. The Reconsolidation group did the Test Phase 24 hours after Reactivation, whereas the Immediate control group did the Test Phase immediately after Reactivation, in order to investigate whether memory modification required a delay. Only the Reconsolidation group was scanned.

Results

Behavioral Results

We transcribed and scored memory tests for two key measures: number of unique *correct details* (Figure 2A) and *false memories* (Figure 2B). We also collected *confidence ratings* and scored the number of *forgotten videos* (Supplemental Material, Confidence and Forgetting) (Supplementary Figure 1). We defined false memories as distorted details that the participant recalled from the episode (e.g., "The pitcher wore a green hat"). Void responses (e.g., "I don't remember") were not counted as false memories, but were missed opportunities to earn points for correct details. We conducted linear mixed-effects regression to predict memory outcomes from the fixed factors *group* (Reconsolidation and Immediate) and *reactivation type* (Full and Interrupted). In all models, we included random effects to account for by-subject and by-video variability (Methods, Linear Mixed Effects Regression).

Correct Details

We found that prediction errors during memory reactivation enhanced recall of correct details (Figure 2A), such that participants in both groups reported more correct details for Interrupted videos than Full videos, $F_{(1,69)} = 7.59$, $p = .007$, 95% CI [-0.12, -0.02] (Supplementary Table 1). Even though the video endings were omitted, prediction errors strengthened and preserved existing memories. Participants in the Reconsolidation group recalled fewer correct details than participants in the Immediate group, $F_{(1,46)} = 4.69$, $p = .036$, 95% CI: [0.02, 0.31], likely because the Reconsolidation group completed the memory test after a 24-hour delay. There was no interaction between group and reactivation type, $F_{(1,248)} = 0.48$, $p = .488$, 95% CI [-0.04, 0.02], indicating that the effect of prediction error enhancing correct details did not require a delay.

False Memories

We found that prediction errors selectively increased false memories after a 24-hour delay in the Reconsolidation group, replicating our past behavioral results (53) (significant interaction between reactivation type and group, $F_{(1,1067)} = 6.76$, $p = .009$, 95% CI [0.01, 0.07], Figure 2B, Supplementary Table 1). In other words, Interrupted videos increased false memories in the Reconsolidation group ($t(23) = -4.84$, $p < .001$), but not the Immediate group ($t(23) = -0.88$, $p = .387$). We also found main effects of group, $F_{(1,46)} = 105.07$, $p < .0001$, 95% CI [-0.43, -0.29], and reactivation type, $F_{(1,341)} = 10.80$, $p = .001$, 95% CI [-0.08, -0.02], both driven by the effect of prediction error increasing false memories in the Reconsolidation group.

In sum, our behavioral results showed a novel dissociation between reinforcing and updating memories: Prediction errors during memory reactivation strengthened memories

immediately, but updating memories to add new information required a delay, as predicted by reconsolidation theory.

Item Analysis: Surprise Ratings and Semantic Similarity

Expanding on the results reported above, we recruited an independent sample to watch the videos and rate (on a 5-point Likert scale) the degree of surprise elicited by the narrative interruptions (Methods, Online Video Ratings). We found that surprise ratings were unrelated to correct details (Supplementary Table 2), but there was a significant interaction between surprise ratings and group, such that more surprising videos were associated with more false memories selectively in the Reconsolidation group, $F_{(1,2994)} = 4.28$, $p = .039$, 95% CI [-0.06, -0.01].

Here, we use false memories as an index of memory updating; however, incorporating relevant new information into memory can be an adaptive function. We hypothesized that our paradigm would induce false memories because information would be integrated across semantically-related episodes. To test this hypothesis, we quantified semantic similarity among the 70 videos with a text-based analysis (Methods, Memory Tests) (Supplementary Figure 2). Videos that were more semantically similar to other videos in the stimulus set yielded more false memories, $F_{(1,68)} = 7.03$, $p = .010$, 95% CI [0.03, 0.17] (Supplementary Table 3). Semantic similarity did not predict correct details overall, $F_{(1,67)} = 0.09$, $p = .769$, 95% CI [-0.05, 0.04], but showed a significant interaction with reactivation type, $F_{(1,68)} = 8.22$, $p = .006$, 95% CI [0.02, 0.11]. For Full videos, semantic similarity was positively associated with correct details, consistent with schema-based memory benefits (55). For Interrupted videos, semantic similarity was negatively related to correct details, suggesting a trade-off with false memories. Overall, these results suggest that memories were updated with relevant new information, exactly as required for adaptive behavior.

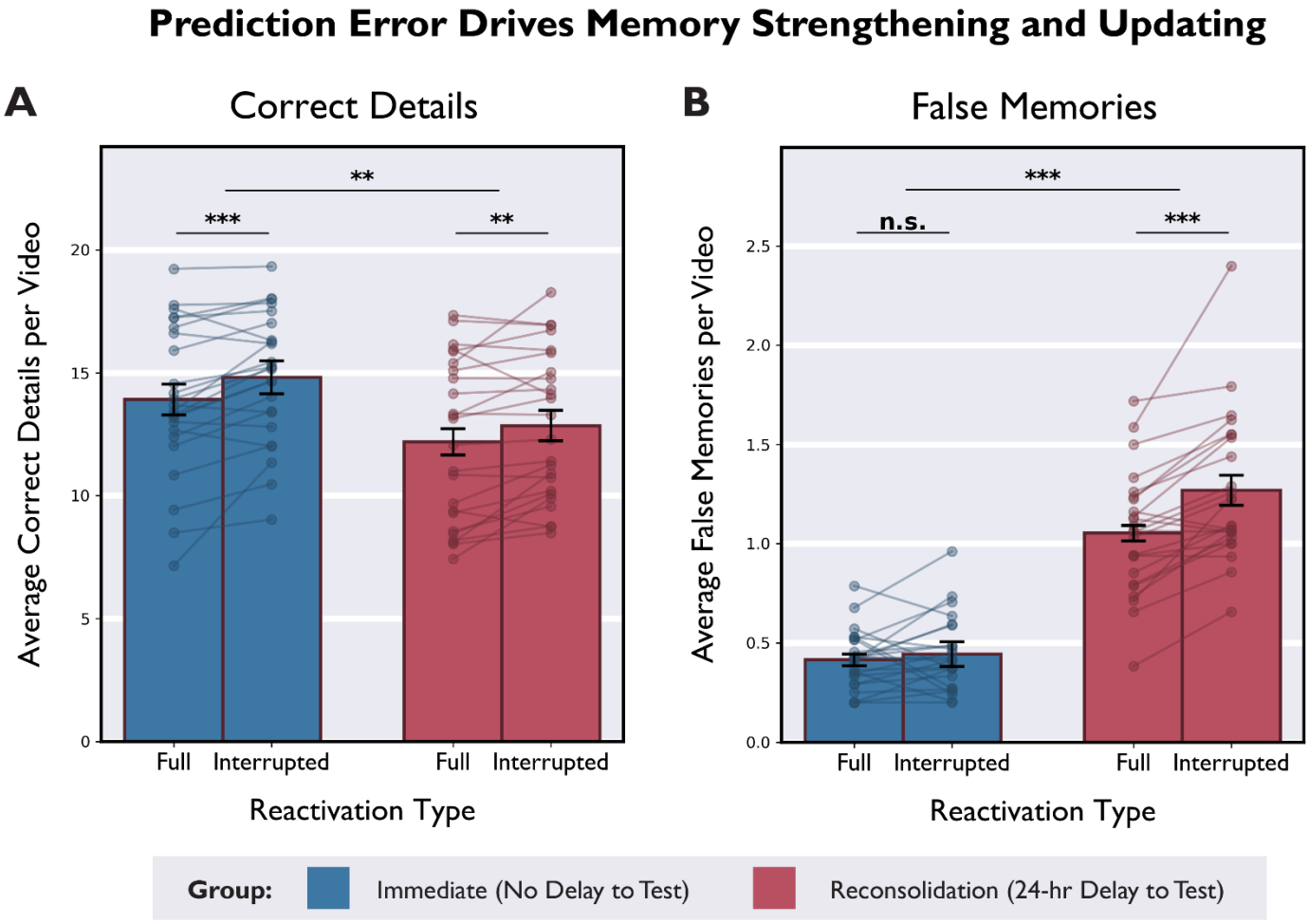


Figure 2. Prediction errors strengthened and updated memories over distinct time-courses. A) In both groups, average Correct Details were higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error can strengthen memory recall both immediately and after a delay. B) Only in the Reconsolidation group (24-hour delay-to-test), average False Memories were higher for videos that were Interrupted during memory reactivation. This interaction demonstrates that prediction error enabled memory updating, but only after a delay that permitted reconsolidation. Dots indicate average scores by-participant, and lines connect within-subjects measures. Error bars depict 95% confidence intervals. * $p < .05$, ** $p < .01$, *** $p < .001$.

Univariate fMRI Results

The primary aim of our univariate fMRI analyses was to test the following questions:

1. Is hippocampal activation related to *reactivation type* (Full vs. Interrupted) and memory updating as indexed by subsequent *false memories?*
2. If so, does activation in the *basal forebrain* or the *VTA* moderate the relationship between hippocampal activation and memory updating?

We analyzed the blood oxygen level-dependent (BOLD) signal from the 24 subjects in the Reconsolidation group (the Immediate group was not scanned). Our analyses focused on the fixation screen presented during the Post-Video period immediately after each video offset. The narrative ending of each video was either as-expected (Full) or a surprising prediction error (Interrupted). We controlled for visual and auditory input across conditions by analyzing neural activation during the Post-Video fixation period (Figure 2B). Whole-brain mass univariate results are provided in the Supplemental Material (Whole-Brain Analysis, Supplementary Table 4, Supplementary Figure 3).

Some past studies have shown that prediction error signals are stronger in left hippocampus and anterior hippocampus (18, 20, 21, 56), whereas posterior hippocampus is more sensitive to video offsets (57). Other studies have shown that anterior and posterior hippocampus parse continuous experience at different timescales (58, 59). On the basis of these findings, we tested separate ROIs for left, right, anterior, and posterior hippocampus (Methods, ROI Masks), but found that our effects were generally very consistent across hippocampal ROIs (Supplemental Material, ROI Differences). Main text results are averaged across bilateral hippocampus, but results from individual ROIs are depicted in Supplementary Figures 4-9.

Relating Hippocampal Activation to Memory Updating

We used single-trial modelling to relate post-video hippocampal activation to subsequent false memories. For our univariate analyses, we modelled a 2s impulse during the Post-Video period (fixation screen), convolved with the canonical double-gamma hemodynamic response function and phase-shifted 2s after video offset. This 2s shift targets the peak Post-Video hippocampal response identified in previous studies (60, 61). We isolated fMRI activation during the Post-Video period on each trial and averaged parameter estimates across all hippocampal voxels (Methods, fMRI Data Analysis). Using linear mixed-effects regression, we predicted trial-wise hippocampal activation from the following variables: *reactivation type* (Full vs. Interrupted), *false memories* (continuous measure), and their interaction.

We found a significant interaction between reactivation type and subsequent false memories associated with hippocampal activation, $F_{(1,1271)} = 8.54$, $p = .004$, 95% CI [-0.13, -0.03] (Figure 3A) (Supplementary Table 5A). After Full videos, greater hippocampal activation was associated with fewer subsequent false memories (Figure 3A, blue). This protective effect is consistent with the idea that Post-Video hippocampal activation reinforces memory for the episode that just concluded (60–62). After an episode that aligns with expectations, the hippocampus should remain in an internal processing mode that favors ongoing predictions and memory retrieval (63). However, when the ending of the video was surprising, we observed exactly the opposite effect. After Interrupted videos, greater hippocampal activation was associated with *more* false memories, consistent with the idea that surprise updates memories by triggering a switch to an external processing mode (Figure 3A, orange).

Investigating the Role of the Basal Forebrain and Ventral Tegmental Area

Next, we tested hypotheses about neuromodulatory mechanisms by examining activation in the basal forebrain (which contains the medial septal nucleus, the primary source of ACh in the hippocampus) (28, 30, 36) and the VTA (which contains dopaminergic neurons that project to the hippocampus) (16, 23). We added two parameters to the model reported above to examine average basal forebrain and VTA activation during the Post-Video period (i.e., single-trial activation estimates in the basal forebrain and VTA for the 2-second target period of the fixation, consistent with modelling of hippocampal activation), as well as interactions with hippocampal activation, reactivation type, and subsequent false memories. All model parameters are listed in Supplementary Table 5B.

We found that basal forebrain activation was significantly positively related to hippocampal activation during the Post-Video period, $F_{(1,1380)} = 9.80$, $p = .002$, 95% CI [0.03, 0.14]. There was also a significant three-way interaction among basal forebrain activation, reactivation type, and false memories predicting hippocampal activation, $F_{(1,1392)} = 7.68$, $p = .006$, 95% CI [-0.13, -0.02] (Figure 3B). This interaction demonstrated that the relationship between hippocampal activation and subsequent memory (Figure 3A) was only evident when the basal forebrain was also activated (Figure 3B, right). When basal forebrain activation was weak, hippocampal activation was unrelated to memory (Figure 3B, left). VTA activation was also positively related to hippocampal activation during the Post-Video period ($F_{(1,1376)} = 21.74$, $p < .001$, 95% CI [0.07, 0.18]), but there was no interaction among VTA activation, reactivation type, and false memories ($F_{(1,1375)} = 0.02$, $p = .876$, 95% CI [-0.06, 0.05]), demonstrating that the VTA did not moderate the effect of hippocampal activation on memory.

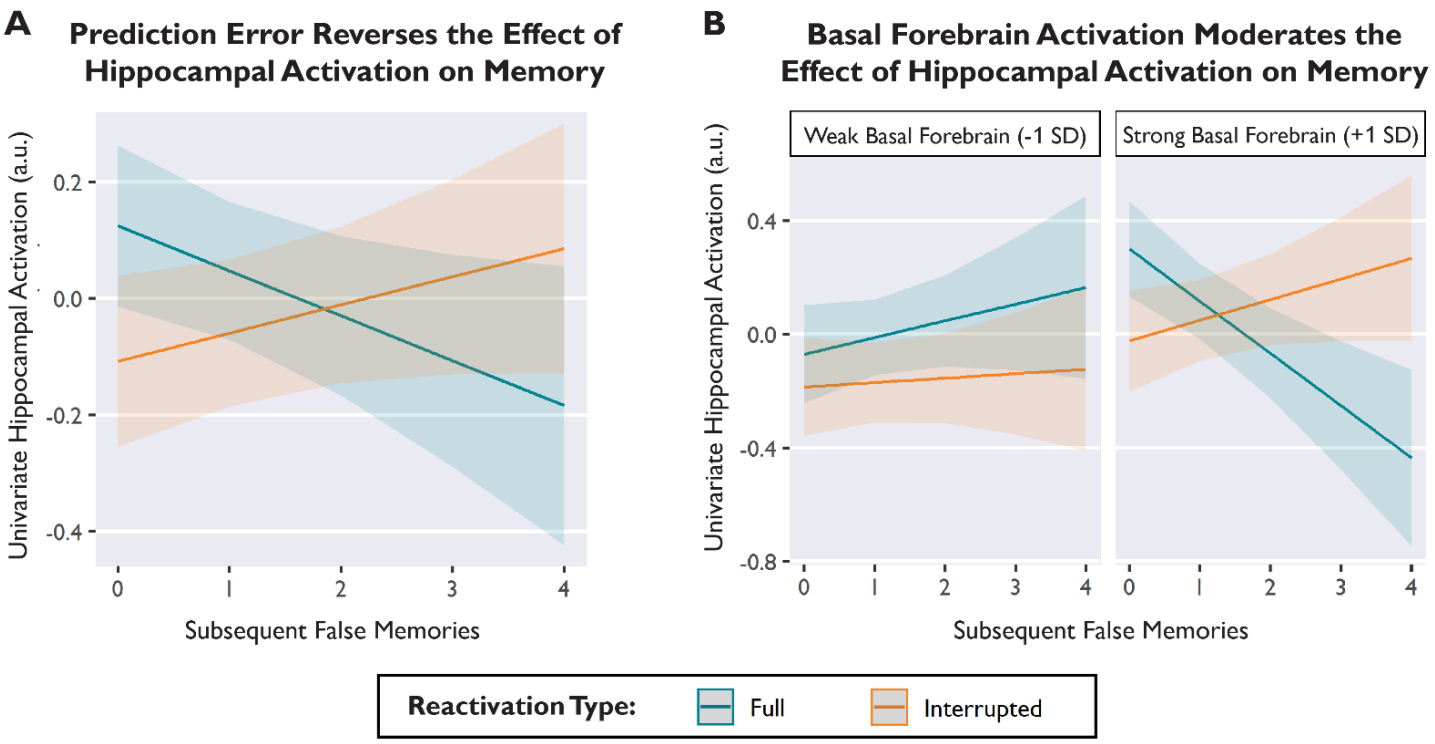


Figure 3. Prediction error reverses the relationship between hippocampal activation and subsequent memory, and this effect depends on concurrent basal forebrain activation. A) After Full-length videos, greater hippocampal activation had a protective effect, predicting fewer false memories (blue). After Interrupted videos, greater hippocampal activation predicted more false memories (orange). B) The effect of prediction error on hippocampal activation and memory was only observed when basal forebrain activation was strong (right). When basal forebrain activation was weak, hippocampal activation was unrelated to memory (left). Basal forebrain activation is binned (weak vs. strong) for visualization, but statistical models used a continuous variable. Lines depict model-predicted estimates, and shaded bands depict the 95% confidence interval. Model-predicted estimates are shown instead of individual data points in order to show within-subjects effects, while controlling for subject and stimulus variance.

Prediction Error Alters the Role of the Hippocampus

Overall, our results suggest that prediction error altered the function of the hippocampus and thus determined the fate of episodic memories. Past studies only showed that hippocampal activation *increased* after prediction errors (18–21, 64), and did not examine the consequences for subsequent memory. Crucially, we found that the same amount of hippocampal activation

could exert opposing effects on memory depending on whether video endings aligned with expectations. We investigated the mechanisms underlying this surprising finding by examining functional connectivity with neuromodulatory regions; we found that the relationship between hippocampal activation and memory depended on concurrent basal forebrain activation, supporting the idea that acetylcholine regulates hippocampal processing.

How does prediction error change hippocampal processing? We propose that during memory reactivation, the hippocampus retrieves a past memory, generates predictions, and checks for a mnemonic prediction error. If no prediction error is detected, then the hippocampus switches to an *internal* processing mode, reinforcing details from within the episode and thus protecting a memory from change (i.e., preventing false memories). If a prediction error is detected, then the hippocampus switches to an *external* processing mode, abandoning ongoing predictions and updating memories with new information (i.e., here increasing false memories). Cholinergic modulation from the basal forebrain could regulate switching between internal and external processing modes, thus influencing memory. In the following section, we test mechanistic predictions of this account.

Multivariate fMRI Results

On the basis of our univariate findings, we proposed that the hippocampus continually generates predictions and sustains representations during ongoing episodes (65, 66), but that prediction errors can trigger the hippocampus to abandon these sustained representations and switch to an external processing mode (27, 28). Therefore, we hypothesized that prediction errors would disrupt sustained representations in the hippocampus; that these effects would be specific to the hippocampus; and that disrupting hippocampal representations would lead to memory updating. Mechanistically, we predicted that activation in the basal forebrain and/or VTA would

link hippocampal representations to memory outcomes, via neuromodulation of hippocampal processing modes (23–28).

The goal of our multivariate analyses was to test the following questions:

1. Are hippocampal representations sustained during and after narrative episodes?
2. Do prediction errors disrupt sustained hippocampal representations?
3. Does univariate activation in the basal forebrain or VTA link hippocampal representations to memory updating?

Autocorrelation During and After Videos

Past studies in rodents and humans have used *autocorrelation* measures, which quantify similarity across neural patterns, to investigate hippocampal representations during naturalistic tasks (58, 59). *Temporal autocorrelation* is an index of multivariate information that is preserved over time; this measures moment-to-moment overlap of activation patterns (58, 66, 67). Intracranial recordings in humans have shown that temporal autocorrelation in the hippocampus ramps up over the course of familiar episodes (66). Ramping autocorrelation reflects sustained neural representations, here consistent with the hippocampus generating predictions and anticipating upcoming stimuli (65, 66). To test whether hippocampal representations were sustained or disrupted over time, we calculated temporal autocorrelation by correlating the activation of all voxels within the hippocampus at timepoint *T* with the activation pattern at timepoint *T+1 sec* (Methods, Multivariate fMRI Analyses). We predicted that hippocampal representations would be sustained after Full videos (increasing autocorrelation), but disrupted after Interrupted videos.

Using linear mixed effects regression, we found that hippocampal autocorrelation increased linearly as videos progressed, ($F_{(1,7379)} = 8.41$, $p = .004$, 95% CI [0.01, 0.04], Figure

4A, Supplementary Table 6A), suggesting that episodic representations were sustained and stabilized during video playback (66). Plotting second-by-second autocorrelation values revealed that autocorrelation trajectories for Full and Interrupted videos diverged at the moment of video offset (Figure 4B): after Full videos, hippocampal representations were sustained during the Post-Video period, whereas after Interrupted videos, hippocampal representations were disrupted. To quantify this Post-Video shift in hippocampal representations, we analyzed the average change in autocorrelation from the factors *segment* (before or after video offset), *reactivation type* (Full or Interrupted), and their interaction term (Supplementary Table 6B). There was a significant interaction between segment and reactivation type predicting change in autocorrelation ($F_{(1,2818)} = 13.24$, $p < .001$, 95% CI [0.03, 0.11]), such that autocorrelation increased more for Full than Interrupted videos during the Post-Video period (Figure 4C). In other words, prediction errors disrupted the continuity of hippocampal representations.

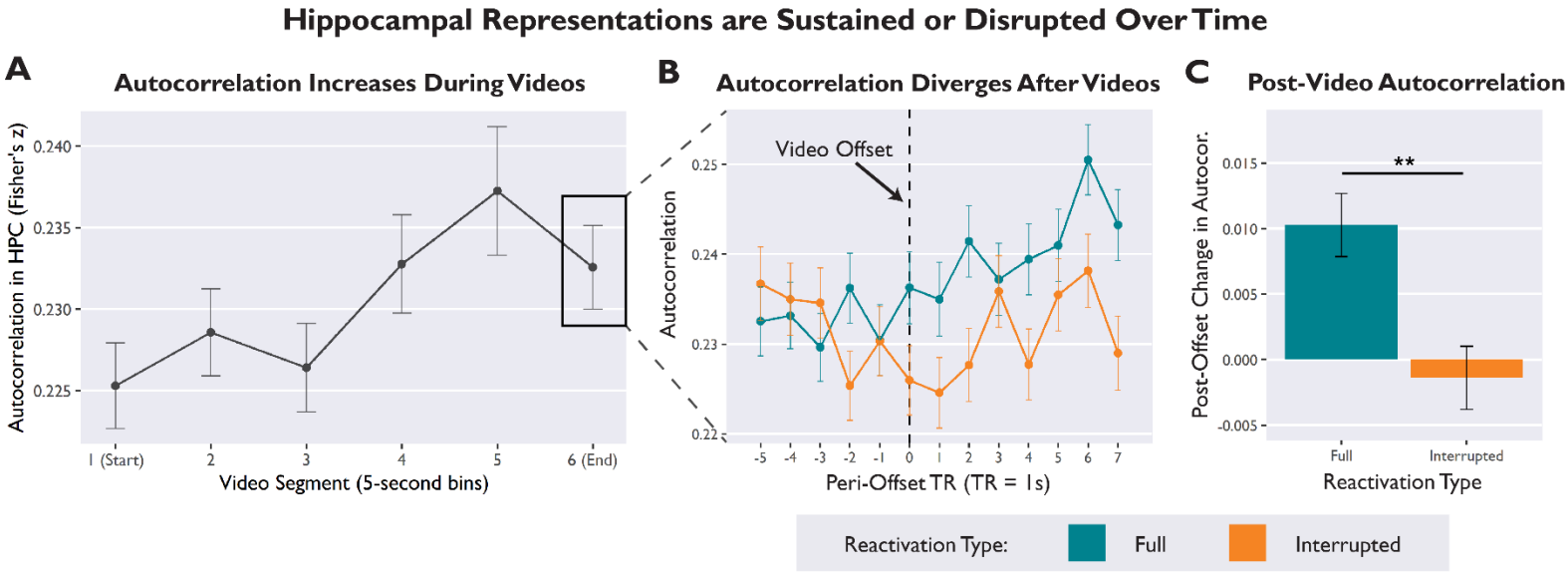


Figure 4. Hippocampal representations are sustained or disrupted over time. A) Temporal autocorrelation in the hippocampus gradually increased over the course of a video, suggesting that episodic representations were sustained over time. Autocorrelation values were averaged over 5-second bins of video playback. B) Autocorrelation trajectories for Full and Interrupted

videos diverged during the Post-Video period. Plot visualizes second-by-second autocorrelation values in the hippocampus, time-locked to the moment of video offset (black dotted line). C) Average Post-Video change in autocorrelation (average autocorrelation scores for the 5-sec bin immediately *after* video offset, minus average autocorrelation for the bin immediately *before* offset). Hippocampal representations were sustained after Full videos, but disrupted after Interrupted videos. Error bars depict SEM.

Relating Autocorrelation to Subsequent Memory

Next, we tested whether disruption of hippocampal representations predicted memory updating. Using linear mixed effects regression, we predicted false memories from the interaction between reactivation type and Post-Video change in autocorrelation, including univariate hippocampal activation as a continuous covariate (thus controlling for any autocorrelation effects that are a consequence of univariate activation). We found a significant interaction between reactivation type and change in autocorrelation predicting false memories, $F_{(1,1349)} = 5.84$, $p = .016$, 95% CI [0.01, 0.11] (Figure 5A, Supplementary Table 7A). After Interrupted videos, disrupting hippocampal representations led to memory updating. Conversely, after Full videos, hippocampal autocorrelation was unrelated to memory.

What determines whether hippocampal representations are sustained or disrupted? To investigate candidate neuromodulatory mechanisms, we extended the model described above (predicting false memories from reactivation type, change in autocorrelation, and univariate activation in the hippocampus) to include univariate activation in the basal forebrain and the VTA (continuous variables). The model included all relevant interaction terms, reported in full in Supplementary Table 7B. Paralleling our univariate findings, we found a significant three-way interaction among basal forebrain activation, reactivation type, and hippocampal autocorrelation that predicted subsequent false memories, $F_{(1,1338)} = 4.56$, $p = .033$, 95% CI [0.01, 0.09] (Figure

5B). In other words, prediction errors disrupted hippocampal representations and led to memory updating, but only when the basal forebrain was also activated (Figure 5B, right). In contrast, the three-way interaction among VTA activation, reactivation type, and hippocampal autocorrelation was not significant, $F_{(1,1352)} = 0.02$, $p = .894$, 95% CI [-0.04, 0.05]. Overall, we found that our autocorrelation results paralleled our univariate findings; basal forebrain activation was crucial for connecting hippocampal representations to memory outcomes.

Lastly, to determine the anatomical specificity of our autocorrelation findings, we tested two control regions: inferior lateral occipital cortex (LOC) and white matter. We predicted that autocorrelation in LOC would be sensitive to all video offsets because of the change in visual input, but *not* sensitive to prediction error. In contrast, physiological noise from white matter should not be sensitive to either video offsets or prediction errors. Autocorrelation in LOC significantly increased after videos ($t(23) = 9.47$, $p < .001$, Cohen's $d = 1.37$), but did not differ by reactivation type ($t(23) = -0.05$, $p < .96$, Cohen's $d = 0.01$). Autocorrelation in white matter did not change post-offset ($t(23) = 0.99$, $p = .329$, Cohen's $d = 0.14$) and did not differ by reactivation type ($t(23) = 0.86$, $p = .40$, Cohen's $d = 0.18$). In summary, these control analyses indicated that our autocorrelation findings were not a brain-wide phenomenon; prediction error selectively disrupted sustained representations in the hippocampus.

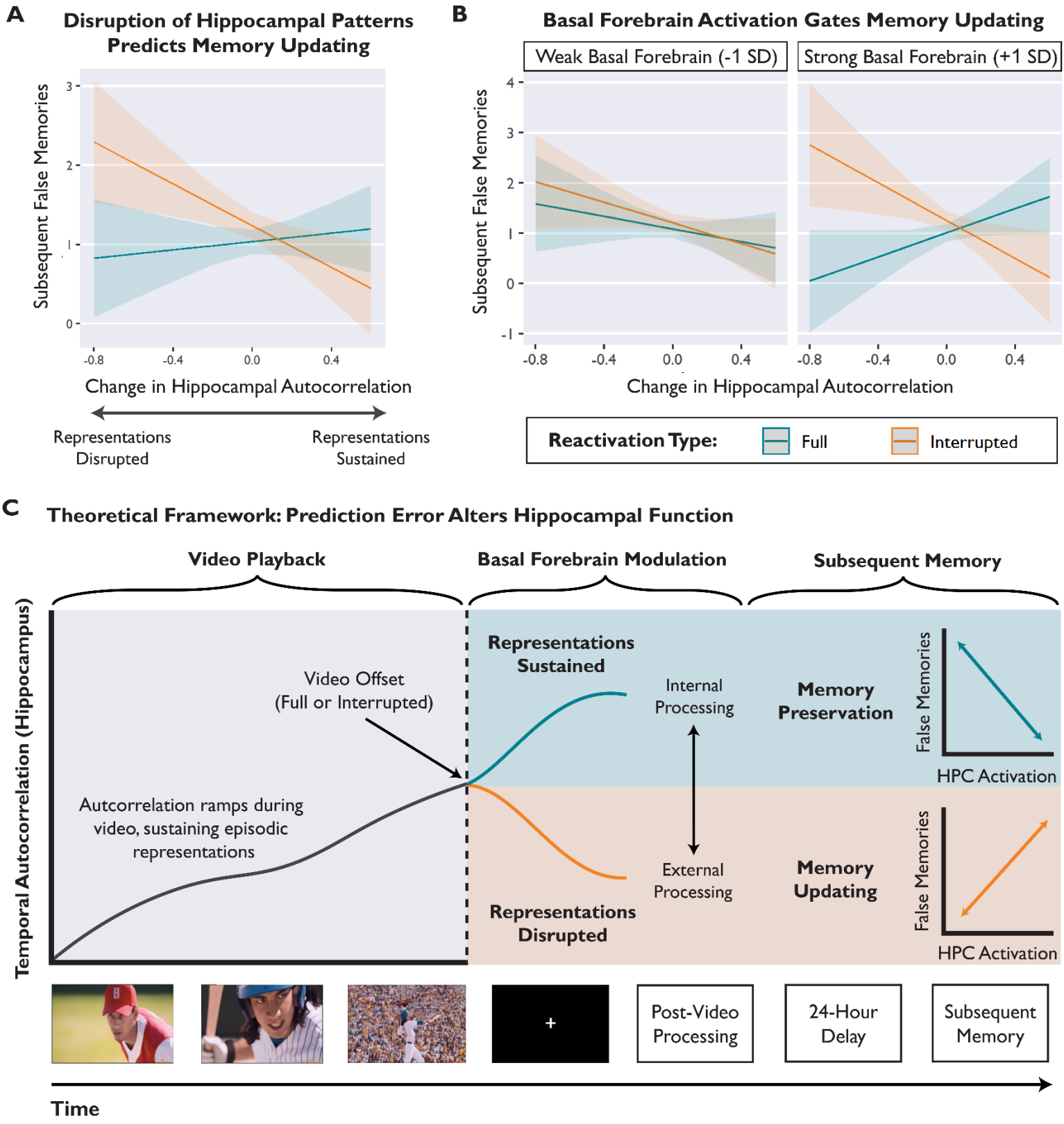

A
Disruption of Hippocampal Patterns Predicts Memory Updating
Subsequent False Memories
Change in Hippocampal Autocorrelation
Representations Disrupted
Representations Sustained
B
Basal Forebrain Activation Gates Memory Updating
Weak Basal Forebrain (-1 SD)
Strong Basal Forebrain (+1 SD)
Subsequent False Memories
Change in Hippocampal Autocorrelation
Reactivation Type:
Full
Interrupted
C
Theoretical Framework: Prediction Error Alters Hippocampal Function
Video Playback
Basal Forebrain Modulation
Subsequent Memory
Temporal Autocorrelation (Hippocampus)
Video Offset (Full or Interrupted)
Autcorrelation ramps during video, sustaining episodic representations
Representations Sustained
Representations Disrupted
Internal Processing
External Processing
Memory Preservation
Memory Updating
False Memories
HPC Activation
False Memories
HPC Activation
+
Post-Video Processing
24-Hour Delay
Subsequent Memory
Time

Figure 5. Prediction errors elicited by Interrupted videos disrupt sustained hippocampal representations, and this disruption predicts memory updating. A) Estimated values from a linear regression model predicting subsequent false memories from the interaction of reactivation type and change in autocorrelation. After Interrupted videos, decreases in autocorrelation were related to increased memory updating. B) The effect of prediction error on hippocampal autocorrelation and subsequent memory depended on concurrent basal forebrain activation. Basal forebrain activation was binned (weak vs. strong) for visualization, but statistical models used a continuous variable. Shaded bands depict 95% confidence intervals around the regression line. In panels A and B, model-predicted estimates are depicted instead of individual data points in order to show within-subjects effects, while controlling for subject and stimulus variability. C) Conceptual schematic depicting the effect of prediction error on hippocampal representations and subsequent memory. During a video, the hippocampus sustains episodic representations over time, consistent with generating ongoing predictions. After video offset, preservation or disruption of these representations indicates whether the hippocampus switches to an internal processing mode, whereby univariate activation preserves memories, or an external processing mode, whereby univariate activation updates memories. Mode switching and memory outcomes depend on co-activation of the hippocampus and basal forebrain during the post-video period.

Discussion

Here, we show that prediction errors modulate the function of the hippocampus and allow memories to be updated with relevant new information. In our fMRI paradigm, we elicited mnemonic prediction errors by interrupting familiar narrative videos immediately before the expected conclusions. Prediction errors reversed the relationship between univariate hippocampal activation and subsequent memory: After expected video endings, hippocampal activation was associated with memory preservation, but after prediction errors, hippocampal activation was associated with memory updating. Tracking the stability of hippocampal representations revealed that prediction errors disrupted activation patterns; this pattern disruption predicted memory updating. Crucially, the association between hippocampal activation (both univariate and multivariate) and memory outcomes depended on concurrent basal forebrain activation during the Post-Video period. We conclude that prediction error,

coupled with basal forebrain modulation, shifts the hippocampus into an external processing mode that supports updating memories with new information (Figure 5B).

Prediction Errors Disrupt Hippocampal Representations and Update Memories

Past studies of mnemonic prediction errors have reported an increase in univariate hippocampal activation, but have not examined whether this neural signal affects memory (17–19, 64). For the first time, we show that after prediction errors, hippocampal activation leads to memory updating. Crucially, we demonstrate that univariate measures are *insufficient* for understanding the effect of prediction error on the hippocampus, because the same amount of hippocampal activation can exert opposing effects on memory. Prediction error reversed the relationship between hippocampal activation and subsequent memory, consistent with a switch between internal and external processing modes (Figure 3). After expected endings (Full videos), hippocampal activation protected against false memories, consistent with an internal processing mode that supports ongoing predictions. After surprising endings (Interrupted videos), hippocampal activation predicted *more* false memories, consistent with an external processing mode that supports updating memories with new information (20, 27, 28).

To test the idea that prediction errors trigger mode switching, we tracked hippocampal activation patterns to examine how episodic representations were sustained or disrupted over time. We used *temporal autocorrelation* (the moment-to-moment overlap of activation patterns) as a measure of continuity in hippocampal representations (58, 66, 67). As narratives progressed, autocorrelation increased, reflecting stability and continuity; this increase in autocorrelation suggested that the hippocampus generates predictions (65, 66) and sustains episodic representations over time (60, 68) (Figure 4A). Crucially, prediction errors disrupted the stability of hippocampal representations (Figure 4B, 4C), and this disruption predicted the degree of

memory updating (Figure 5A). Overall, we propose that disruption of hippocampal patterns reflects switching to an external processing mode that allows memories to be updated with new information (Figure 5B). Our findings substantially advance past research (17, 18, 64): We link hippocampal prediction error signals to memory updating, show that prediction error reverses the relationship between hippocampal activation and memory outcomes, and uncover mechanisms of this shift in hippocampal processing.

Basal Forebrain Activation Relates to Hippocampal Mode Switching

Past studies have argued that either cholinergic (26–28, 34–36) or dopaminergic (16, 23) modulation of the hippocampus could enhance plasticity after prediction error. However, mixed evidence supporting both hypotheses has left the question unresolved (24, 25, 30, 33). Here, we investigated whether activation of the basal forebrain or the VTA could explain the relationship between hippocampal activation and subsequent memory. We found that the effect of prediction error on memory depended on co-activation of the hippocampus and basal forebrain, suggesting that cholinergic modulation is important for mode switching. Hippocampal activation was associated with memory updating after prediction errors, but only when the basal forebrain was also activated. Likewise, resetting hippocampal patterns led to memory updating after prediction errors, but only when the basal forebrain was also activated. Overall, our results support the idea that cholinergic modulation gates memory updating and regulates the balance between internal and external processing modes.

Our findings also offer new insight into the functional connections between the VTA and hippocampus. Although we found a robust positive correlation between VTA and hippocampal activation during the Post-Video period, VTA activation was unrelated to prediction error and did not link hippocampal activation with memory outcomes. These findings are consistent with

the idea that connectivity between the VTA and hippocampus regulates ongoing, tonic modulation of learning states, rather than phasic prediction error signals (16, 25, 69–72). Instead, phasic dopaminergic prediction error signals could be transmitted from the locus coeruleus to the hippocampus (73, 74). Future research could disambiguate the roles of the basal forebrain, VTA, and locus coeruleus by examining connectivity with the hippocampus (both event-related and ongoing) and consequences for memory.

Prediction Error Both Strengthens and Updates Memories

Our behavioral results demonstrated a novel dissociation: prediction error both strengthened and updated memories, but over distinct timecourses (Figure 2). Prediction error increased the number of correct details recalled, both immediately and after a one-day delay. In contrast, prediction error increased false memories only after a delay, supporting that idea that reconsolidation enables memory updating (1, 54, 75). The finding that prediction error also increased correct details demonstrated that the false memories arose from an adaptive updating mechanism, not forgetting.

In the present study, we used false memories as an index of memory updating; in the real world, it is adaptive to update memories with relevant new information. In our paradigm, interference from other stimulus videos likely produced false memories because information was integrated across videos. Previously, we found that prediction errors selectively updated memories with semantically-related information from interference videos (53). Here, we showed that videos that shared greater semantic similarity with the rest of the stimulus set produced more false memories (Supplementary Table 3). Memories could be updated with semantically-related details that arise from reactivation of related memories, or visual input from subsequent videos during the task. This finding accords with reconsolidation research (1, 54, 76) and computational

models of event segmentation (77, 78), which have both shown that interference among related episodes can produce false memories after prediction error. However, memory updating is beneficial in other situations that require integrating old and new knowledge, or correcting erroneous information.

Limitations and Future Directions

Our experimental design was inspired by reconsolidation theory, but evidence for cellular reconsolidation processes in humans is lacking. Numerous behavioral studies have used reconsolidation-like paradigms to demonstrate memory malleability (4, 53, 54, 76), but it remains unknown whether the synaptic mechanisms of reconsolidation are consistent across animals and humans (1). We found that the effect of prediction error on memory updating required a delay, consistent with reconsolidation processes that rely on protein synthesis. Overall, our findings are broadly relevant to research on prediction error and memory even though the synaptic mechanisms remain unknown; reconsolidation theory offers one plausible framework for our results. Another limitation is that the present data lacked the spatial resolution required to segment hippocampal subfields, because we prioritized temporal resolution over spatial resolution in order to more accurately track rapid changes in hippocampal representations during and after naturalistic episodes.

We elicited prediction error by interrupting videos before their expected narrative endings, comparable to the incomplete reminders (e.g., a conditioned stimulus without the expected outcome) that have been previously used in animal and human reconsolidation studies (1, 3, 76). However, it remains unknown whether a reminder must be *incomplete* in order to initiate reconsolidation, or whether other surprising or novel stimuli (e.g., sounds, alternate endings) may also induce memory malleability. Incomplete reminders may be particularly

effective because memory reactivation supports plasticity (79–81). Future studies could investigate memory reactivation by testing encoding-retrieval pattern similarity.

Conclusion

The brain continually generates predictions based on past experiences. When expectations do not align with reality, memories should be updated with relevant new information. We propose that prediction errors prompt the hippocampus to abandon ongoing predictions and switch to an externally-oriented processing mode that allows memories to be updated. In this way, surprise modulates hippocampal states and determines the fate of episodic memories. This theoretical framework bears implications for understanding how memories can be modified in eyewitness testimony, education, and conditions like Post-Traumatic Stress Disorder. Beyond memory research, our results offer new insights for theories on the whole-brain predictive processes that govern attention, perception, action, and decision-making.

Acknowledgements: This research was funded by grants awarded to MDB from the *James S. McDonnell Foundation* (Scholar Award in Understanding Human Cognition) and the *Natural Sciences and Engineering Research Council* of Canada (Discovery Grant and Accelerator Supplement, RGPIN-2014-05959 and RGPIN-2020-05747). AHS has been supported by awards from the *National Science Foundation* (Graduate Research Fellowship) and *Natural Sciences and Engineering Research Council* of Canada (Postgraduate Doctoral Scholarship, Undergraduate Student Research Award). We also give thanks to Carolyn Chung, Tolulemi Gbile, and Aria Fallah for their invaluable contributions to data collection, transcription, and scoring.

Author Contributions: AHS and MDB developed the study design. AHS programmed the study, collected data, conducted analyses, and drafted the manuscript. GMM contributed substantially to data collection and IKB contributed to autocorrelation analyses. MDB and RAA contributed to the analysis approach and interpretation of results. All authors contributed to revising the manuscript and approved the final version.

Declaration of Interests: The authors have no competing interests to declare.

Methods

Data, Code, and Materials

Brief descriptions of the stimulus videos are provided in Supplementary Table 8. The full set of stimulus videos, along with derivative data and code necessary to reproduce results, are provided online in the project repository hosted by the Open Science Framework (https://osf.io/xb7sq/).

Participants

We recruited 55 participants from the University of Toronto community to participate in the study for monetary compensation (Reconsolidation group: $70, Immediate control group: $40). Of these participants, 7 were excluded (reasons stated below), leaving a final sample of 48 participants. The sample size was determined *a priori* to satisfy the following conditions: (a) achieve at least 90% power to detect the interaction effect previously found with a variant of this paradigm ($\eta_p^2 = 0.17$) (82), (b) reproduce the sample size previously used with a variant of this paradigm (53), and (c) evenly allocate participants to 6 pseudorandomized trial order lists. Inclusion criteria were as follows: between the ages of 18-30, normal or corrected-to-normal vision and hearing, no history of neurological or psychiatric disorders, and fluency in English. All participants provided informed consent prior to beginning the study. The study was approved by the University of Toronto Institutional Review Board, Protocol #00035787.

Participants were healthy young adults (age: $M = 22.42$, $SD = 2.41$, range [18, 30]; gender: 75% female, 25% male) with fluency in English, normal or corrected-to-normal vision and hearing, and no history of neurological or psychiatric disorders. fMRI participants were all right-handed. In consideration of the effects of sleep on consolidation, we also asked participants to report approximate hours of sleep over the course of the study. Participants slept an average of

7.28 hours (*SD* = 1.31) between the Day 1 and Day 2 sessions, and Reconsolidation group participants slept an average of 7.02 hours (*SD* = 1) between the Day 2 and Day 3 sessions.

Stimuli

Stimulus videos were sourced from movies, TV, and YouTube clips. We chose 70 videos that featured distinct narrative events (duration *M* = 30 sec, *SD* = 7 sec). Semantic similarity varied across videos (e.g., several videos featured sporting events), but there were no overlapping sources, settings, or characters. The stimulus set included 18 videos that were previously used in a behavioral version of the paradigm. During pilot testing, we ensured that the videos would be infrequently recognized by our participants. The 70 videos used in the experiment are described in Supplementary Table 8 and publicly available on the Open Science Framework (https://osf.io/xb7sq/). The Interrupted version of each video ended abruptly at the narrative climax, omitting the salient ending and violating expectations (duration *M* = 25 sec, *SD* = 4 sec).

For the fMRI version of the task (Reconsolidation group), stimuli were presented with EyeLink Experiment Builder (SR-Research) on a BOLDscreen display monitor (32”, 1920x1090, 100Hz refresh rate), viewed through a mirror attached to the head coil. Auditory stimulation was presented with in-ear MRI-compatible headphones (Sensimetrics, model S14). During the initial scout scan, we performed a sound test by playing the soundtrack of a movie trailer (not included in the stimulus videos) and adjusting the volume. For the behavioral version of the task (Immediate control group), videos were presented on a desktop computer and audio was presented with over-ear headphones.

Procedure

During the **Encoding session**, participants viewed all 70 videos in full-length form (randomized order). Each video was presented twice in a row to ensure that participants had strong expectations about the narrative outcomes for each video, a prerequisite for eliciting prediction error later.

During the **Reactivation session**, participants viewed each video again a single time. Half of the videos were Full and half of the videos were Interrupted. Videos were played in a pseudorandom order such that there were never more than two consecutive Interrupted videos. This pseudorandom presentation prevented participants from anticipating which videos would be interrupted. Participants were counterbalanced and sequentially assigned to one of six pseudorandom trial orders. We also performed eye-tracking during the Encoding and Reactivation sessions for participants in both the Reconsolidation and Immediate groups (EyeLink v.1000+, SR-Research). Eye-tracking was used to monitor alertness during the task, but these data are not discussed further.

Lastly, the **Test session** involved a structured interview-style recall test about details from each of the videos. Participants were cued with the name of each video and prompted to recall the narrative. The experimenter then probed the participant for more information with a pre-determined list of open-ended questions (e.g., “Can you describe the setting or context of the video?”, “Can you describe what the character looked like? Do you remember gender, age range, hair color, or clothing?”). Participants were instructed to answer based on their memory of the full-length videos that had been originally presented during encoding. Because we were interested in false memories as a measure of memory modification, we instructed participants not to guess and permitted them to skip details they could not recall.

Overall, the experiment took place over three days for participants in the Reconsolidation group (24-hour delays between Encoding, Reactivation, and Test), or over two days for participants in the Immediate control group (24-hour delay between Encoding and Reactivation, no delay between Reactivation and Test). Only the Reconsolidation group underwent neuroimaging.

Consistent with past reconsolidation studies (83–85), we maintained consistent contextual factors between Encoding, Reactivation, and Test sessions. Reconsolidation group participants completed the encoding session in a mock scanner (shell of a retired 1.5T Siemens Avanto scanner), while recorded MRI sounds were played in the background. Reconsolidation group participants completed the Reactivation session in the real fMRI scanner and the Test session at a desk in the mock scanner room. Participants in the Immediate control group completed all three sessions in the same behavioral testing room. In both groups, participants completed all three sessions with the same experimenter.

fMRI Scanning

Scanning was performed with a 3T Siemens Prisma MRI scanner located at the Toronto Neuroimaging Center, University of Toronto. High-resolution functional images were collected with a T2*-weighted multiband-accelerated echo-planar imaging (EPI) pulse sequence, and a 32-channel head coil. Foam padding was used to minimize head motion. We acquired whole-brain BOLD activation estimates with a spatial resolution of 2.7mm isotropic voxels (TR: 1000 ms, TE: 29 ms, flip angle: 50°, 60 slices at transversal orientation, phase encoding: A>P, FoV: 210mm, Partial Fourier: 0.875, multiband factor: 4). High resolution T1-weighted anatomical images were acquired with a magnetization-prepared rapid-acquisition gradient-echo (MP-

RAGE) pulse sequence (voxel size: 1mm isotropic, TE: 24 ms, TR: 2000 ms, TI: 1100 ms, flip angle: 9°) to allow 3D reconstruction and volume-based statistical analysis.

Scoring of Memory Tests

We transcribed memory tests with *Temi*, an automated voice-to-text tool, then manually edited transcripts to verify accuracy (https://www.temi.com/). We coded videos as "forgotten" if the participant entirely failed to retrieve a memory when cued with the name of the video and a hint from a pre-determined list (brief descriptions of each video, provided in Supplementary Table 8). Scoring of details was conducted with *NVivo 12*, a program for qualitative analysis of transcripts. Research assistants manually labelled each detail as correct or false. Scorers were blinded to subject identity and reactivation type (Full vs. Interrupted) while scoring the memory tests. The number of false memories per-trial ranged from 0-6, but there were very few trials with 5 or 6 false memories. To account for these high outliers, we winsorized the false memories variable to the 95th percentile (4 false memories). Winsorizing improved model fits but did not affect the statistical significance of our results.

Lastly, we quantified semantic similarity among the videos by using the Cluster Analysis function in *NVivo*. Across all transcripts, we pooled the phrases used to describe each video, excluding false memories and irrelevant words (e.g., *the, um, and, maybe, confidence, remember*). We then calculated pairwise Pearson correlations on the basis of the most frequent 100 words used to describe each video. For each video, we calculated an overall semantic similarity score by averaging the correlation values; this metric summarizes how much the content of a given video relates to the rest of the stimulus set. A heatmap displaying all pairwise correlation values is provided in Supplementary Figure 2.

Online Ratings of Stimulus Videos

We recruited 3,913 participants online using Amazon's Mechanical Turk. Participants were paid $0.50 to complete a Qualtrics survey that took approximately 3 minutes. Each participant was randomly assigned to view one stimulus video, first as the Full version and then as the Interrupted version. We included timing constraints to ensure that participants could not progress to the next page of the survey before the video had finished playing. Participants were excluded for the following reasons: (1) they failed the attention check question ("If you are paying attention, choose 4 below."), (2) they failed the comprehension check question ("In general, not just in the video, is the emotion 'happiness' positive or negative?"), (3) they reported that they had experienced playback issues, or (4) they reported that they had seen the video clip prior to the survey. After exclusions, our sample size was 1,907 (20-41 raters per video). On 5-point Likert scales, participants rated how surprising each video felt when the ending was interrupted, as well as video memorability and emotional valence/intensity (Supplementary Tables 9-11).

Exclusions

In the Immediate Control group, two participants were excluded due to technical issues. In the Reconsolidation group, three participants were excluded due to a counterbalancing error and audio playback problems, and two participants were excluded because they had previously completed a pilot version of the study. Additionally, one full run of fMRI data (14 trials) was excluded for one participant due to audio playback failure and excessive motion. On a trial-by-trial basis, videos were excluded if technical issues arose (e.g., audio issues) (10 trials), the participant was falling asleep (as determined by eyetracking) (20 trials), or the participant reported having seen the video prior to the experiment (103 trials). In total, there were 147 trials that were excluded for the above reasons (out of all 48 participants in both the Reconsolidation

and Immediate groups). The total number of excluded trials for Full and Interrupted videos was approximately equal (Full: 70; Interrupted: 77). Additionally, subsequently forgotten videos were excluded from single-trial brain-to-behavior analyses (63 trials across the 24 participants in the Reconsolidation group). Overall, only 4.4% of all trials were excluded.

Linear Mixed-Effects Regression

All linear mixed-effects regression models reported in the main text included random intercepts for *subject* (identity of each participant) and *video* (identity of each stimulus item), along with random slopes for *reactivation type*. In accordance with current best practices (86), the random effects structure was determined by the maximal amount of complexity that was supported by the data (allowing model convergence and optimizing model fits as determined by the Akaike Information Criterion). All models converged successfully; we used the BOBYQA controller with 10,000 maximum iterations. We used restricted maximum likelihood estimation and assessed significance of predictors with a Type III ANOVA and Satterthwaite approximations of degrees of freedom. In R (v3.6), we constructed models with the *lme4* package (87) and evaluated significance with the *lmerTest* package (88). Variables for *reactivation type* and *group* were treated as factors, and all continuous variables were standardized/mean-centered. These model parameters applied to analysis of behavioral data, single-trial univariate neural activation, and temporal autocorrelation. Parameter estimates from all models are reported in the Supplementary Tables. Plots were generated with the packages *ggplot2* and *sjPlot* (89, 90).

fMRI Preprocessing

All data were preprocessed and analyzed using FSL v6.0, in conjunction with in-house R code (v3.6). Initial volumes were discarded by the scanner to allow for signal saturation.

Preprocessing steps included fieldmap distortion correction, spatial realignment, removal of head-motion artifacts (six regressors), nuisance regression of average white matter and CSF timeseries, slice-timing correction for an interleaved multiband acquisition, and high-pass frequency filtering (120s). For native-space ROI analyses (single-trial univariate and autocorrelation analyses), data were minimally smoothed with a 2-mm kernel to preserve spatial specificity and multivariate information.

Region of Interest Masks

We used FreeSurfer (v6.0) to automatically create binarized hippocampal masks in each subject's native space. After FreeSurfer segmentation, hippocampal masks were manually inspected and segmented into ROIs for left anterior, left posterior, right anterior, and right posterior hippocampus. Anterior and posterior regions were split along the long-axis at the uncal apex. We found that our effects were very consistent among the four hippocampal ROIs (Supplemental Material, ROI Differences). Therefore, results reported in the main text (single-trial univariate and autocorrelation analyses) are averaged across the entire hippocampus bilaterally. White matter masks were obtained with FSL segmentation utilities. Inferior Lateral Occipital Cortex (LOC) masks were taken from the Harvard-Oxford Cortical Atlas. VTA masks were taken from a probabilistic midbrain atlas developed by the Adcock lab (91). Basal forebrain masks were taken from the probabilistic cytoarchitectonic Julich-Brain atlas. We used ROIs for bilateral cholinergic nuclei Ch123, including the medial septal nucleus. This region (in contrast to Ch4) exhibits resting-state functional connectivity with the hippocampus (92). We investigated temporal signal-to-noise in the basal forebrain to ensure that results were not driven by noise (Supplemental Material, Basal Forebrain SNR). All standard space masks were

transformed into native space for each functional run, using the inverse deformation field from preprocessing and registration.

Univariate fMRI Analyses

Whole-brain mass univariate results are reported in the Supplemental Material (Whole-Brain Analysis, Supplementary Figure 3, Supplementary Table 4). The primary findings reported in the main text reflect a single-trial modelling approach that estimated hippocampal responses to each video during the task. In order to isolate responses on each trial, we employed the Least Squares-Single approach and constructed a separate GLM for each trial (93, 94). We modelled each trial as a 2s impulse in the post-video period, convolved with the canonical double-gamma hemodynamic response function and phase-shifted 2s after video offset. This 2s shift targets the peak hippocampal response previously identified in studies of post-video processing (60, 61). Within each GLM, the target trial (2s event) was isolated as one regressor, and all other events were modelled with a separate regressor for each type of event (e.g., video playback, video name cues, other fixation periods). This approach yielded whole-brain parameter estimates for each trial, in native space. For each trial, we masked the processed data and averaged across voxels within each ROI. Average activation values within each ROI were then submitted to linear mixed-effects regression, thus linking trial-wise ROI activation to reactivation type and subsequent memory.

Multivariate fMRI Analyses

Multivariate temporal autocorrelation analyses (58, 66) were conducted on the same preprocessed data described above. We extracted the whole-run timeseries from every voxel within each ROI using the *fslmeants* utility. For control analyses (white matter and LOC ROIs), autocorrelation was calculated on 200 contiguous voxels, approximately matching the size of the

hippocampal ROIs. Comparable to past research, we phase-shifted the timeseries by 4 seconds in order to account for HRF lag (95). Temporal autocorrelation was defined as the Pearson product-moment correlation between all voxel activation values at timepoint T and timepoint T+1s. This method yielded an autocorrelation value for every second of each functional run, excluding the final TR. Autocorrelation values were standardized (Fisher's *z*) prior to statistical analysis.

Next, we aligned multivariate timeseries data with event onset and duration markers. After alignment, we calculated average autocorrelation values that were time-locked to events. For statistical analyses, autocorrelation values were averaged across 5-second bins during and after each video. To analyze signal history over the course of video playback, we related the video segment number (5s bins) to average autocorrelation values. For each video, we included the first five seconds (timepoints 0-4), the next four middle segments (timepoints 5-9, 10-14, 15-19, and 20-24), and the last five seconds (variable depending on the length of the video). This binning scheme spans the average video length of 30 seconds; additional middle segments from videos that were longer than 30 seconds were omitted. Lastly, to compare post-offset changes in autocorrelation, we calculated difference scores between the 5-second bins immediately before and after video offset. Autocorrelation values and difference scores for each trial were then submitted to linear mixed effects regression.

References

1. A. H. Sinclair, M. D. Barense, Prediction Error and Memory Reactivation: How Incomplete Reminders Drive Reconsolidation. *Trends Neurosci.* **42**, 727–739 (2019).

2. R. N. Henson, P. Gagnepain, Predictive, Interactive Multiple Memory Systems. *Hippocampus* **20**, 1315–1326 (2010).

3. M. T. J. Exton-Mcguinness, J. L. C. Lee, A. C. Reichelt, Updating memories—The role of prediction errors in memory reconsolidation. *Behav. Brain Res.* **278**, 375–384 (2015).

4. A. Pine, N. Sadeh, A. Ben-Yakov, Y. Dudai, A. Mendelsohn, Knowledge acquisition is governed by striatal prediction errors. *Nat. Commun.* **9**, 1–14 (2018).

5. G. Kim, K. A. Norman, N. B. Turk-Browne, Neural differentiation of incorrectly predicted memories. *J. Neurosci.* **37**, 2022–2031 (2017).

6. J. A. Quent, R. N. Henson, A. Greve, A predictive account of how novelty influences declarative memory. *Neurobiol. Learn. Mem.* **179**, 107382 (2021).

7. D. L. Schacter, D. R. Addis, The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philos. Trans. R. Soc. B Biol. Sci.* **362**, 773–786 (2007).

8. L. F. Barrett, W. K. Simmons, Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* **16**, 419–429 (2015).

9. K. Friston, The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).

10. G. B. Keller, T. D. Mrsic-Flogel, Predictive Processing: A Canonical Cortical Computation. *Neuron* **100**, 424–435 (2018).

11. M. W. Spratling, A review of predictive coding algorithms. *Brain Cogn.* **112**, 92–97 (2017).

12. R. S. Sutton, A. G. Barto, *Reinforcement learning: An introduction*, 2nd Ed. (MIT Press, 1998).

13. M. Watabe-Uchida, N. Eshel, N. Uchida, Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* **40**, 373–394 (2017).

14. N. E. Wheeler, *et al.*, Ideology and predictive processing: coordination, bias, and polarization in socially constrained error minimization. *Curr. Opin. Behav. Sci.* **34**, 192–198 (2020).

15. N. Hindy, E. Avery, N. Turk-Browne, Hippocampal-neocortical interactions sharpen over time for predictive actions. *Nat. Commun.* **10** (2019).

16. D. Shohamy, R. A. Adcock, Dopamine and adaptive memory. *Trends Cogn. Sci.* **14**, 464–472 (2010).

17. J. Chen, P. A. Cook, A. D. Wagner, Prediction strength modulates responses in human area CA1 to sequence violations. *J. Neurophysiol.* **114**, 1227–1238 (2015).

18. K. Duncan, N. Ketz, S. J. Inati, L. Davachi, Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus* **22**, 389–398 (2012).

19. D. Kumaran, E. A. Maguire, An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS Biol.* **4**, e424 (2006).

20. O. Bein, K. Duncan, L. Davachi, Mnemonic prediction errors bias hippocampal states. *Nat. Commun.* **11**, 3451 (2020).

21. J. Chen, R. K. Olsen, A. R. Preston, G. H. Glover, A. D. Wagner, Associative retrieval processes in the human medial temporal lobe: Hippocampal retrieval success and CA1 mismatch detection. *Learn. Mem.* **18**, 523–528 (2011).

22. K. Duncan, C. Curtis, L. Davachi, Distinct memory signatures in the hippocampus: intentional states distinguish match and mismatch enhancement signals. *J. Neurosci.* **29**, 131–139 (2009).

23. J. E. Lisman, A. A. Grace, The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron* **46**, 703–713 (2005).

24. J. E. Lisman, N. A. Otmakhova, Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus* **11**, 551–568 (2001).

25. J. Schomaker, M. Meeter, Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. *Neurosci. Biobehav. Rev.* **55**, 268–279 (2015).

26. M. Hasselmo, B. Wyble, G. Wallenstein, Encoding and Retrieval of Episodic Memories: Role of Cholinergic and GABAergic Modulation in the Hippocampus. *Hippocampus* **6** (1996).

27. C. J. Honey, E. L. Newman, A. C. Schapiro, Switching between internal and external modes: A multiscale learning principle. *Netw. Neurosci.* **1**, 339–356 (2017).

28. M. Meeter, J. M. J. Murre, L. M. Talamini, Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus* **14**, 722–741 (2004).

29. B. E. Sherman, N. B. Turk-Browne, Statistical prediction of the future impairs episodic encoding of the present. *Proc. Natl. Acad. Sci.* **8** (2020).

30. M. F. Carr, L. M. Frank, A single microcircuit with multiple functions: state dependent information processing in the hippocampus. *Curr. Opin. Neurobiol.* **22**, 704–708 (2012).

31. M. M. Chun, J. D. Golomb, N. B. Turk-Browne, A taxonomy of external and internal attention. *Annu. Rev. Psychol.* **62**, 73–101 (2011).

32. L. L. Colgin, Rhythms of the hippocampal network. *Nat. Rev. Neurosci.* **17**, 239–249 (2016).

33. H. Tarder-Stoll, M. Jayakumar, H. R. Dimsdale-Zucker, E. Günseli, M. Aly, Dynamic internal states shape memory retrieval. *Neuropsychologia* **138**, 107328 (2020).

34. E. L. Newman, S. N. Gillet, J. R. Climer, M. E. Hasselmo, Cholinergic blockade reduces theta-gamma phase amplitude coupling and speed modulation of theta frequency consistent with behavioral effects on encoding. *J. Neurosci.* **33**, 19635–19646 (2013).

35. L. M. Giocomo, M. E. Hasselmo, Neuromodulation by glutamate and acetylcholine can change circuit dynamics by regulating the relative influence of afferent input and excitatory feedback. *Mol. Neurobiol.* **36**, 184–200 (2007).

36. M. E. Hasselmo, The role of acetylcholine in learning and memory. *Curr. Opin. Neurobiol.* **16**, 710–715 (2006).

37. C. Kemere, M. F. Carr, M. P. Karlsson, L. M. Frank, Rapid and Continuous Modulation of Hippocampal Network State during Exploration of New Places. *PLOS ONE* **8**, e73114 (2013).

38. A. L. Decker, K. Duncan, Acetylcholine and the complex interdependence of memory and attention. *Curr. Opin. Behav. Sci.* **32**, 21–28 (2020).

39. D. R. Vago, A. Bevan, R. P. Kesner, The role of the direct perforant path input to the CA1 subregion of the dorsal hippocampus in memory retention and retrieval. *Hippocampus* **17**, 977–987 (2007).

40. N. Bunzeck, E. Düzel, Absolute Coding of Stimulus Novelty in the Human Substantia Nigra/VTA. *Neuron* **51**, 369–379 (2006).

41. B. C. Wittmann, N. Bunzeck, R. J. Dolan, E. Düzel, Anticipation of novelty recruits reward system and hippocampus while promoting recollection. *NeuroImage* **38**, 194–202 (2007).

42. V. P. Murty, R. A. Adcock, Enriched encoding: reward motivation organizes cortical networks for hippocampal detection of unexpected events. *Cereb. Cortex N. Y. N 1991* **24**, 2160–8 (2014).

43. D. Shohamy, A. D. Wagner, Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron* **60**, 378–389 (2008).

44. R. A. Adcock, A. Thangavel, S. Whitfield-Gabrieli, B. Knutson, J. D. E. Gabrieli, Reward-Motivated Learning: Mesolimbic Activation Precedes Memory Formation. *Neuron* **50**, 507–517 (2006).

45. A. Tompary, K. Duncan, L. Davachi, Consolidation of associative and item memory is related to post-encoding functional connectivity between the ventral tegmental area and different medial temporal lobe subregions during an unrelated task. *J. Neurosci.* **35**, 7326–7331 (2015).

46. J. Debiec, J. E. LeDoux, K. Nader, Cellular and systems reconsolidation in the hippocampus. *Neuron* **36**, 527–538 (2002).

47. J. L. C. Lee, Reconsolidation: maintaining memory relevance. *Trends Neurosci.* **32**, 413–420 (2009).

48. K. Nader, G. E. Schafe, J. E. Le Doux, Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406**, 722–726 (2000).

49. K. Nader, E. O. Einarsson, Memory reconsolidation: an update. *Ann N Acad Sci* **1191**, 27–41 (2010).

50. C. E. Wideman, K. H. Jardine, B. D. Winters, Involvement of classical neurotransmitter systems in memory reconsolidation: Focus on destabilization. *Neurobiol. Learn. Mem.* **156**, 68–79 (2018).

51. K. H. Jardine, *et al.*, Activation of cortical M1 muscarinic receptors and related intracellular signaling is necessary for reactivation-induced object memory updating. *Sci. Rep.* **10**, 9209 (2020).

52. C. Forcato, P. Argibay, M. Pedreira, H. Maldonado, Human reconsolidation does not always occur when a memory is retrieved: The relevance of the reminder structure. *Neurobiol. Learn. Mem.* **91**, 50–57 (2009).

53. A. H. Sinclair, M. D. Barense, Surprise and destabilize: Prediction error influences episodic memory reconsolidation. *Learn. Mem.* **25**, 369–381 (2018).

54. A. Hupbach, R. Gomez, L. Nadel, "Episodic memory reconsolidation: An update" in *Memory Reconsolidation*, C. M. Alberini, Ed. (Elsevier Academic Press, 2013), pp. 233–247.

55. M. T. R. van Kesteren, D. J. Ruiter, G. Fernández, R. N. Henson, How schema and novelty augment memory formation. *Trends Neurosci.* **35**, 211–219 (2012).

56. R. A. Cooper, M. Ritchey, Progression from feature-specific brain activity to hippocampal binding during episodic encoding. *J. Neurosci.* **40**, 1701–1709 (2020).

57. Z. M. Reagh, A. I. Delarazan, A. Garber, C. Ranganath, Aging alters neural activity at event boundaries in the hippocampus and Posterior Medial network. *Nat. Commun.* **11**, 1–12 (2020).

58. I. K. Brunec, *et al.*, Multiple Scales of Representation along the Hippocampal Anteroposterior Axis in Humans. *Curr. Biol.* **28**, 2129-2135.e6 (2018).

59. K. B. Kjelstrup, *et al.*, Finite scale of spatial representation in the hippocampus. *Science* **321**, 140–143 (2008).

60. A. Ben-Yakov, N. Eshel, Y. Dudai, Hippocampal Immediate Poststimulus Activity in the Encoding of Consecutive Naturalistic Episodes. *J. Exp. Psychol. Gen.* **142**, 1255–1263 (2013).

61. A. Ben-Yakov, Y. Dudai, Constructing Realistic Engrams: Poststimulus Activity of Hippocampus and Dorsal Striatum Predicts Subsequent Episodic Memory. *J. Neurosci.* **31**, 9032–9042 (2011).

62. C. Baldassano, *et al.*, Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* **95**, 709-721.e5 (2017).

63. K. Duncan, A. Sadanand, L. Davachi, Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. *Science* **337**, 485–7 (2012).

64. D. Kumaran, E. A. Maguire, Match-mismatch processes underlie human hippocampal responses to associative novelty. *J. Neurosci.* **27**, 8517–8524 (2007).

65. R. U. Haque, S. K. Inati, A. I. Levey, K. A. Zaghloul, Feedforward prediction error signals during episodic memory retrieval. *Nat. Commun.* **11**, 6075 (2020).

66. R. Paz, *et al.*, A neural substrate in the human hippocampus for linking successive events. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 6046–6051 (2010).

67. R. Raut, A. Snyder, M. Raichle, Hierarchical dynamics as a macroscopic organizing principle of the human brain. *PNAS* **117**, 20890–20897 (2020).

68. S. DuBrow, L. Davachi, Temporal binding within and across events. *Neurobiol. Learn. Mem.* **134**, 107–114 (2016).

69. V. P. Murty, I. C. Ballard, R. A. Adcock, Hippocampus and Prefrontal Cortex Predict Distinct Timescales of Activation in the Human Ventral Tegmental Area. *Cereb. Cortex N. Y. NY* **27**, 1660–1669 (2017).

70. J. Lisman, A. A. Grace, E. Duzel, A neoHebbian framework for episodic memory; role of dopamine-dependent late LTP. *Trends Neurosci.* **34**, 536–547 (2011).

71. A. A. Grace, S. B. Floresco, Y. Goto, D. J. Lodge, Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends Neurosci.* **30**, 220–227 (2007).

72. C. G. McNamara, Á. Tejero-Cantero, S. Trouche, N. Campo-Urriza, D. Dupret, Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nat. Neurosci.* **17**, 1658–1660 (2014).

73. C. G. McNamara, D. Dupret, Two sources of dopamine for the hippocampus. *Trends Neurosci.* **40**, 383–384 (2017).

74. K. A. Kempadoo, E. V. Mosharov, S. J. Choi, D. Sulzer, E. R. Kandel, Dopamine release from the locus coeruleus to the dorsal hippocampus promotes spatial learning and memory. *Proc. Natl. Acad. Sci.* **113**, 14835–14840 (2016).

75. K. Nader, Memory traces unbound. *Trends Neurosci.* **26**, 65–72 (2003).

76. C. Forcato, M. L. C. Rodríguez, M. E. Pedreira, H. Maldonado, Reconsolidation in humans opens up declarative memory to the entrance of new information. *Neurobiol. Learn. Mem.* **93**, 77–84 (2010).

77. N. T. Franklin, K. A. Norman, C. Ranganath, J. M. Zacks, S. J. Gershman, Structured Event Memory: A neuro-symbolic model of event cognition. *Psychol. Rev.* **127**, 327–361 (2020).

78. G. A. Radvansky, J. M. Zacks, Event boundaries in memory and cognition. *Curr. Opin. Behav. Sci.* **17**, 133–140 (2017).

79. V. J. H. Ritvo, N. B. Turk-Browne, K. A. Norman, Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends Cogn. Sci.* **23**, 726–742 (2019).

80. D. Stawarczyk, C. Wahlheim, J. Etzel, A. Snyder, J. Zacks, Aging and the encoding of event changes: The role of neural activity pattern reinstatement. *bioRxiv* (2020) https:/doi.org/https://doi.org/10.1101/809806.

81. D. Stawarczyk, M. A. Bezdek, J. M. Zacks, Event Representations and Predictive Processing: The Role of the Midline Default Network Core. *Top. Cogn. Sci.* (2019) https:/doi.org/10.1111/tops.12450.

82. F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).

83. A. M. Capelo, P. B. Albuquerque, S. Cadavid, Exploring the role of context on the existing evidence for reconsolidation of episodic memory. *Memory* **27**, 280–294 (2019).

84. A. Hupbach, R. Gomez, L. Nadel, Episodic memory updating: The role of context familiarity. *Psychon. Bull. Rev.* **18**, 787–797 (2011).

85. A. Hupbach, R. Gomez, O. Hardt, L. Nadel, Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learn. Mem.* **14**, 47–53 (2007).

86. H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, D. Bates, Balancing Type I error and power in linear mixed models. *J. Mem. Lang.* **94**, 305–315 (2017).

87. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4 (2014).

88. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* **82**, 1–26 (2017).

89. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

90. D. Lüdecke, *sjPlot: Data Visualization for Statistics in Social Science* (2021).

91. V. P. Murty, *et al.*, Resting state networks distinguish human ventral tegmental area from substantia nigra. *NeuroImage* **100**, 580–589 (2014).

92. R. D. Markello, R. N. Spreng, W.-M. Luh, A. K. Anderson, E. De Rosa, Segregation of the human basal forebrain using resting state functional MRI. *NeuroImage* **173**, 287–297 (2018).

93. J. A. Mumford, B. O. Turner, F. G. Ashby, R. A. Poldrack, Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).

94. J. A. Mumford, T. Davis, R. A. Poldrack, The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage* (2014) https:/doi.org/10.1016/j.neuroimage.2014.09.026.

95. T. Sadeh, J. Chen, Y. Goshen-Gottstein, M. Moscovitch, Overlap between hippocampal pre-encoding and encoding patterns supports episodic memory. *Hippocampus* (2019) https:/doi.org/10.1002/hipo.23079.