Prediction errors disrupt hippocampal representations and update episodic memories

Alyssa H. Sinclair[a,b,1], Grace M. Manalili[b], Iva K. Brunec[c,d], R. Alison Adcock[a,e], and Morgan D. Barense[b,f]

[a]Department of Psychology & Neuroscience, Duke University, Durham, NC 27710; [b]Department of Psychology, University of Toronto, Toronto, ON M5S 3G3, Canada; [c]Department of Psychology, Temple University, Philadelphia, PA 19122; [d]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; [e]Department of Psychiatry & Behavioral Sciences, Duke University, Durham, NC 27710; and [f]Rotman Research Institute, Baycrest Hospital, Toronto, ON M64 1W1, Canada

Edited by Daniel Schacter, Psychology, Harvard University, Cambridge, MA; received October 1, 2021; accepted November 5, 2021

The brain supports adaptive behavior by generating predictions, learning from errors, and updating memories to incorporate new information. Prediction error, or surprise, triggers learning when reality contradicts expectations. Prior studies have shown that the hippocampus signals prediction errors, but the hypothesized link to memory updating has not been demonstrated. In a human functional MRI study, we elicited mnemonic prediction errors by interrupting familiar narrative videos immediately before the expected endings. We found that prediction errors reversed the relationship between univariate hippocampal activation and memory: greater hippocampal activation predicted memory preservation after expected endings, but memory updating after surprising endings. In contrast to previous studies, we show that univariate activation was insufficient for understanding hippocampal prediction error signals. We explain this surprising finding by tracking both the evolution of hippocampal activation patterns and the connectivity between the hippocampus and neuromodulatory regions. We found that hippocampal activation patterns stabilized as each narrative episode unfolded, suggesting sustained 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, supporting the idea that cholinergic modulation regulates attention and memory. We conclude that prediction errors create conditions that favor memory updating, prompting the hippocampus to abandon ongoing predictions and make memories malleable.

memory | cognitive neuroscience | hippocampus | prediction error | reconsolidation

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 discover misinformation in the news, receive feedback on an examination, 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 (1–6). The brain dynamically reconstructs memories during recall, recreating and revising past experiences based on current 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 each of 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, 21–23) or "mnemonic prediction error" (20). These past studies have shown that the hippocampus detects mnemonic prediction errors. Several theoretical frameworks have hypothesized that this hippocampal prediction error signal could update memories (17, 20, 24–27), but this crucial link for understanding how we learn from error has not yet been demonstrated.

What mechanisms could link hippocampal prediction errors to memory updating? A leading hypothesis is that prediction errors shift the focus of attention and adjust cognitive processing (20, 28–32). After episodes that align with expectations, we should continue generating predictions and shift attention internally, sustaining and reinforcing existing memories. However, after mnemonic prediction errors, we should reset our expectations and shift attention externally, preparing to encode new information and update memories. Consistent with this idea, mnemonic prediction errors have been shown to enhance the hippocampal input pathway that supports encoding, but suppress the output pathway that supports retrieval (20). We propose that surprising events may also change intrinsic hippocampal processing, changing the effect of hippocampal activation on memory outcomes.

Neuromodulation may be a critical factor that regulates hippocampal processing and enables memory updating. Currently, there is mixed evidence supporting two hypotheses: acetylcholine or dopamine could act upon the hippocampus to regulate processing after surprising events (24–27, 29, 31, 33, 34).

Significance

Our brains draw on memories to predict the future; when our predictions are incorrect, we must update our memories to improve future predictions. Past studies have demonstrated that the hippocampus signals prediction error (i.e., surprise) but have not linked this neural signal to memory updating. Here, we uncover this missing connection. We show that mnemonic 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 temporal continuity of hippocampal patterns. We propose that prediction errors disrupt sustained representations and enable memory updating. Our findings bear implications for improving education, understanding eyewitness memory distortion, and treating pathological memories.

Author contributions: A.H.S. and M.D.B. designed research; A.H.S. and G.M.M. performed research; A.H.S. and I.K.B. analyzed data; R.A.A. and M.D.B. provided conceptual input and supervision; and A.H.S., I.K.B., R.A.A., and M.D.B. wrote the paper.

The authors declare no competing interest

This article is a PNAS Direct Submission.

Published under the PNAS license.

[1]To whom correspondence may be addressed. Email: allie.sinclair@duke.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117625118/-/DCSupplemental.

Published December 15, 2021.

Downloaded at DUKE UNIVERSITY on December 15, 2021

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 (27–29, 35–38), thus allowing stored memories to be compared with perceptual input (31, 38, 39). After prediction errors, acetylcholine release could change hippocampal processing and enhance encoding or memory updating (26, 29, 33, 37, 39). 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 the hippocampus and VTA are coactivated after surprising events (40, 41). Other work has shown that coactivation of the hippocampus and VTA predicts memory encoding and integration (42–45). Overall, basal forebrain and VTA neuromodulation are both candidate mechanisms for regulating hippocampal processing and memory updating.

In the present study, we used an fMRI task with human participants to examine trial-wise hippocampal responses to prediction errors during narrative videos. During the "encoding phase," participants viewed 70 full-length videos that featured narrative episodes with salient endings (e.g., a baseball batter hitting a home run) (Fig. 1*A*). During the "reactivation phase" the following day, participants watched the videos again (Fig. 1*B*). 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 midswing). These surprising interruptions were comparable to the prediction errors employed in prior studies of memory updating (1). Half of the videos were presented in full-length form (*Full*, as previously seen during the encoding phase) and half were presented in interrupted form (*Interrupted*, eliciting prediction error).

During the "test phase," participants completed a memory test in the form of a structured interview (Fig. 1*C*). On each trial, participants were cued with the name of the video and recalled the narrative. The experimenter then probed for further details with predetermined questions (e.g., "Can you describe the baseball batter's ethnicity, age range, or clothing?"). Our critical measure of memory updating was "false memories," because the presence of a false memory indicates that the original memory was changed in some way. 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.

Prior research in human and animals has shown that some memory-updating effects only emerge after delays that allow protein synthesis to occur during consolidation and reconsolidation (1, 46–48). Therefore, to test our primary question about the neural correlates of memory updating, fMRI participants completed the encoding, reactivation, and test phases over 3 d, with 24-h between each session (Delayed group, $n = 24$). In addition, we tested the behavioral prediction that memory updating would require a delay (i.e., because transforming a memory trace requires protein synthesis) by recruiting a separate group of participants who completed the test phase immediately after the reactivation phase on day 2 (Immediate group, $n = 24$) (Fig. 1*D*). Delayed group participants completed the reactivation phase while undergoing an fMRI scan, whereas Immediate group participants ($n = 24$) were not scanned. Our primary fMRI analyses examined the fixation period immediately following the offset of Full and Interrupted videos (postvideo period) (Fig. 1 *B*, *Right*) during the reactivation phase in the Delayed group. Importantly, this design compares neural responses to surprising and expected video endings while controlling for visual and auditory input.

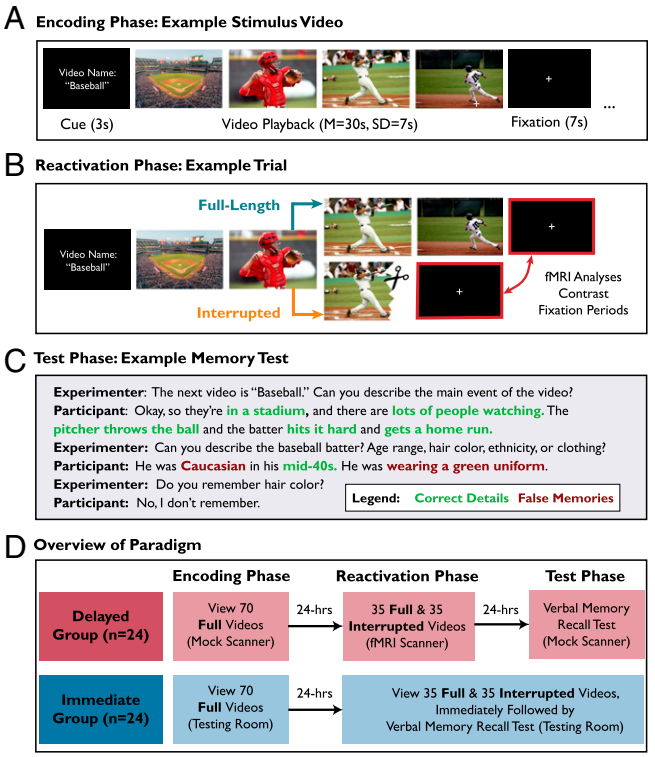

Fig. 1. Overview of experimental paradigm. (*A*) During the encoding phase, all videos were presented in full-length form. Here we show example frames depicting a stimulus video. (*B*) During the reactivation phase, participants viewed the 70 videos again, but half (35 videos) were interrupted to elicit mnemonic 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 midswing. fMRI analyses focused on the postvideo fixation periods (red highlighted boxes). Thus, visual and auditory stimulation were matched across Full and Interrupted conditions, allowing us to compare postvideo neural activation while controlling for perceptual input. (*C*) During the test phase, 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 the Encoding phase. Here we show example text illustrating the memory test format and scoring of correct details (our measure of memory preservation) and false memories (our measure of memory updating, because false memories indicate that the memory has been modified). 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 Delayed group (fMRI participants) completed the test phase 24 h after reactivation, because prior studies have shown that memory updating becomes evident only after a delay (e.g., to permit protein synthesis). The Immediate group completed the test phase immediately after reactivation and was not scanned. The purpose of the Immediate group was to test the behavioral prediction that memory updating required a delay.

Our approach allowed us to test several questions set up by the prior literature. First, we used naturalistic video stimuli to examine the effect of mnemonic prediction error on hippocampal activation and episodic memories. Second, to investigate hippocampal processing, we used multivariate analyses to track how episodic representations were sustained or disrupted over time. Third, to test hypotheses about neuromodulatory mechanisms, we related hippocampal activation and memory updating to activation in the basal forebrain and VTA.

Results

Behavioral Results. We transcribed and scored memory tests for two key measures: number of unique correct details (Fig. 2*A*)

Downloaded at DUKE UNIVERSITY on December 15, 2021

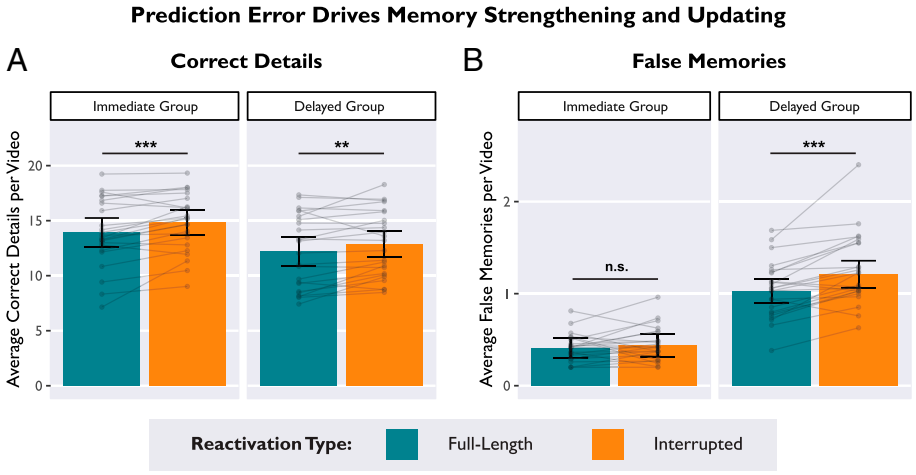

Fig. 2. Prediction errors strengthened and updated memories over distinct time-courses. (*A*) In both the Delayed and Immediate groups, average correct details were higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error can strengthen memory recall both immediately and after a delay. (*B*) In the Delayed group (but not the Immediate group), 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. Bars depict estimated marginal means from a linear mixed effects model. Subject averages are overlaid on top to display the distribution: Dots indicate average scores by-participant, and lines connect within-subjects measures. Error bars depict 95% CI. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

and false memories (Fig. 2*B*), reflecting memory preservation and updating, respectively. We also collected confidence ratings and scored the number of forgotten videos (*SI Appendix, Confidence and Forgetting* and Fig. S1). We defined false memories as distorted details that the participant recalled from the episode (e.g., "The pitcher wore a green uniform"). Void responses (e.g., "I don't remember") were not counted as false memories, but were missed opportunities to earn points for correct details. Importantly, our measures for correct details and false memories were independent; there was no limit to the number of details a participant could recall about a video, and each detail was scored as correct or false. We conducted linear mixed-effects regression to predict memory outcomes (either correct details or false memories) from the factors group (Delayed vs. Immediate) and reactivation type (Full vs. Interrupted). In all models, we included random effects to account for by-subject and by-video variability (*SI Appendix, Supplementary Methods*).

Prediction error increased correct details. We found that prediction errors during memory reactivation enhanced recall of correct details (main effect of reactivation type: $\beta = -0.07$, 95% confidence intervals [CI] [−0.12, −0.02], $t = -2.75$, $P = 0.008$) (Fig. 2*A*), such that participants in both groups reported more correct details for Interrupted videos than Full videos (Delayed group: $\beta = -0.05$, $t = -2.07$, $P = 0.042$; Immediate group: $\beta = -0.08$, $t = -2.50$, $P = 0.015$) (*SI Appendix*, Table S1*A*). Even though the video endings were omitted, prediction errors strengthened and preserved existing memories. Participants in the Delayed group recalled fewer correct details than participants in the Immediate group (main effect of group: $\beta = 0.16$, 95% CI [0.01, 0.31], $t = 2.16$, $P = 0.036$), likely because the Delayed group completed the memory test after 24 h. There was no interaction between group and reactivation type ($\beta = -0.01$, 95% CI [−0.04, 0.02], $t = -0.68$, $P = 0.495$), indicating that the effect of prediction error enhancing correct details did not require a delay.

Prediction error increased false memories. We found that prediction errors selectively increased false memories in the Delayed group, replicating our past behavioral results (48) (significant interaction between reactivation type and group, $\beta = 0.04$, 95% CI [0.01, 0.07], $t = 2.61$, $P = 0.010$) (Fig. 2*B* and *SI Appendix*, Table S1*B*). In other words, Interrupted videos increased false memories in the Delayed group ($\beta = -0.09$, $t = -3.50$, $P < 0.001$), but not the Immediate group ($\beta = -0.01$, $t = -0.78$, $P = 0.437$). We also found a main effect of group indicating more false memories in the Delayed group ($\beta = -0.36$, 95% CI [−0.43, −0.29], $t = -9.68$, $P < 0.001$). Lastly, there was a main effect of reactivation type ($\beta = -0.05$, 95% CI [−0.08, −0.02], $t = -3.31$, $P = 0.001$), driven by the effect of prediction error increasing false memories in the Delayed group. To ensure that the effect of prediction error on memory updating was not driven by the first few trials (which are presumably most surprising), we also conducted a control analysis that found no effect of trial number on false memories (*SI Appendix, Trial Number Control* and Table S2).

In sum, our behavioral results showed a dissociation between reinforcing and updating memories: Prediction errors during memory reactivation strengthened memories, evident both immediately and after a delay (Fig. 2*A*). However, memory updating (as revealed by false memories) was not evident until after a delay (Fig. 2*B*), consistent with prior studies that have been interpreted in terms of reconsolidation theory (1, 46–48).

Surprise ratings and semantic similarity predicted false memories. Expanding on the results reported above, we recruited an independent sample to watch the videos and rate (on a 5-point Likert-style scale) the degree of surprise elicited by the narrative interruptions (*SI Appendix, Supplementary Methods*). We then calculated the average surprise rating for each video and then related these surprise ratings to memory outcomes in our laboratory sample. Using linear mixed-effects regression, we predicted subsequent false memories from the variables reactivation type (Full vs. Interrupted), group (Delayed vs. Immediate), surprise ratings (continuous), and all relevant interactions. All model parameters are reported in *SI Appendix*, Table S3*A*. There was a significant interaction between surprise ratings and group ($\beta = -0.03$, 95% CI [−0.06, −0.01], $t = -2.07$, $P = 0.039$), such that more surprising videos were associated with more false memories selectively in the Delayed group (Delayed: $\beta = 0.10$, $z = 2.49$, $P = 0.013$; Immediate: $\beta = 0.03$, $z = 0.95$, $P = 0.344$). However, the three-way interaction among surprise ratings, group, and reactivation type was not significant ($\beta = 0.01$, 95% CI [−0.02, 0.04], $t = 0.75$, $P = 0.451$). Because our surprise ratings were collected from a separate online sample, this measure may not be sensitive enough to detect the expected three-way

Downloaded at DUKE UNIVERSITY on December 15, 2021

interaction when applied to the laboratory sample. In a separate model, we found that surprise ratings were not associated with correct details (*SI Appendix*, Table S3*B*).

In the present study, we indexed memory updating in terms of false memories; however, incorporating relevant 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 (*SI Appendix*, *Supplementary Methods*). Using linear mixed-effects regression, we predicted subsequent false memories from the variables reactivation type (Full vs. Interrupted), group (Delayed vs. Immediate), semantic similarity (continuous), and all relevant interactions. All model parameters are reported in *SI Appendix*, Table S4. We found that videos that were more semantically-similar to other videos in the stimulus set produced more false memories ($\beta = 0.11$, 95% CI [0.04, 0.19], $t = 3.09$, $P = 0.003$) (*SI Appendix*, Table S4). An interaction between semantic similarity and group predicted false memories ($\beta = -0.04$, 95% CI [−0.07, −0.01], $t = -2.77$, $P = 0.006$), such that the effect of semantic similarity was stronger in the Delayed group ($\beta = 0.16$, $z = 3.90$, $P < 0.001$) than in the Immediate group ($\beta = 0.07$, $z = 1.80$, $P = 0.073$). There was also an interaction between semantic similarity and reactivation type that predicted false memories ($\beta = -0.03$, 95% CI [−0.06, −0.01], $t = -2.08$, $P = 0.038$), such that the effect of semantic similarity was stronger for Interrupted videos ($\beta = 0.15$, $z = 3.44$, $P < 0.001$) than for Full videos ($\beta = 0.08$, $z = 2.18$, $P = 0.030$). However, the three-way interaction among reactivation type, group, and semantic similarity was not significant ($\beta = 0.02$, 95% CI [−0.02, 0.05], $t = 0.99$, $P = 0.323$). Overall, these results suggest that interfering details from semantically related videos distorted memories; this interference effect was strongest for Interrupted videos and for participants in the Delayed group. Consistent with an adaptive updating process, memories may have been updated with relevant information from other videos.

Univariate fMRI Results. The primary aim of our univariate fMRI analyses was to test the following questions: Is hippocampal activation related to reactivation type (Full vs. Interrupted) and memory updating as indexed by subsequent false memories? 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 Delayed group (the Immediate group was not scanned). Our analyses focused on the fixation screen presented during the postvideo period immediately after each video offset. This fixation screen was preceded by the ending of each video, which was either as expected (Full) or a surprising prediction error (Interrupted). Importantly, visual and auditory input was identical across conditions during this postvideo fixation screen, and thus, by analyzing neural activation during the postvideo period we controlled for differences in visual and auditory input (Fig. 2*B*). Another possibility is that condition differences in the duration of visual stimulation (during video playback) could affect the magnitude and duration of the ongoing BOLD response during the postvideo period. We conducted control analyses to confirm that our results were not confounded by video duration (*SI Appendix*, *Video Duration Control* and Table S18).

The effect of hippocampal activation on memory depended on prediction error. Whole-brain mass univariate results are provided in *SI Appendix* (*SI Appendix*, *Whole-Brain Analysis*, Fig. S2, and Table S5). To investigate our primary research questions, we used single-trial modeling to relate postvideo hippocampal activation to subsequent false memories. For our univariate analyses, we modeled a 2-s impulse during the postvideo period (fixation screen), convolved with the canonical double-γ hemodynamic response function (HRF) and phase-shifted 2 s after video offset. This 2-s shift targeted the peak postvideo hippocampal response identified in previous studies (49, 50). We isolated BOLD activation during the postvideo period on each trial and averaged parameter estimates across all voxels within each hippocampal region of interest (ROI) (*Methods*).

Some past studies have shown that prediction error signals are stronger in left hippocampus and anterior hippocampus (18, 20, 21, 51), whereas posterior hippocampus is more sensitive to video offsets (52). Other studies have shown that anterior and posterior hippocampus parse continuous experience at different timescales (53, 54). On the basis of these findings, we tested separate (nonoverlapping) ROIs for left anterior, right anterior, left posterior, and right posterior hippocampus (*SI Appendix*, *Supplementary Methods*). Activation estimates from these four ROIs are included within each model to test for left/right and anterior/posterior differences.

Using linear mixed-effects regression, we predicted trial-wise hippocampal activation from the following variables: reactivation type (Full vs. Interrupted), false memories (continuous measure), hemisphere (left vs. right), axis (anterior vs. posterior), and all relevant interactions.

We found a significant interaction between reactivation type and subsequent false memories predicting hippocampal activation ($\beta = -0.06$, 95% CI [−0.09, −0.03], $t = -4.33$, $P < 0.001$) (Fig. 3*A* and *SI Appendix*, Table S6). This interaction demonstrated that the relationship between hippocampal activation and subsequent memory differed between conditions. After Full videos, greater hippocampal activation was associated with fewer subsequent false memories (Fig. 3*A*, blue) ($\beta = -0.07$, $z = -2.81$, $P = 0.005$), consistent with the idea that the hippocampus reinforces memory for episodes that just concluded (49, 50, 55). However, we observed the opposite effect when events were surprising. After Interrupted videos, greater hippocampal activation was associated more subsequent false memories (Fig. 3*A*, orange) ($\beta = 0.05$, $z = 2.23$, $P = 0.026$), consistent with the idea that surprise drives memory updating. Overall, this interaction demonstrated that the same amount of hippocampal activation predicted different memory outcomes depending on whether the video was Full (fewer false memories) or Interrupted (more false memories).

Neither the main effect of reactivation type on hippocampal activation ($\beta = 0.04$, 95% CI [−0.01, 0.09], $t = 1.85$, $P = 0.069$), nor the main effect of false memories on hippocampal activation ($\beta = -0.01$, 95% CI [−0.05, 0.03], $t = -0.47$, $P = 0.640$) were statistically significant. These null results demonstrate the value of examining the effect of prediction error on both hippocampal activation and memory outcomes. There was a main effect of axis indicating that average activation was greater in anterior hippocampus than posterior hippocampus ($\beta = 0.04$, 95% CI [0.01, 0.06, $t = 3.12$, $P = 0.002$). There was no main effect of hemisphere, and the axis and hemisphere variables did not interact with reactivation type or false memories. Parameter estimates for all variables are provided in *SI Appendix*, Table S6.

Although our primary research questions pertained to false memories, we also tested a model that included both false memories and correct details (*SI Appendix*, Table S7). Consistent with the model reported above, there was a significant interaction between reactivation type and false memories predicting hippocampal activation ($\beta = -0.06$, 95% CI [−0.08, −0.03], $t = -4.01$, $P < 0.001$). Importantly, the associations between hippocampal activation and false memories remained significant after controlling for correct details (Full: $\beta = -0.06$, $z = -2.44$, $P = 0.015$; Interrupted: $\beta = 0.05$, $z = 2.07$, $P = 0.039$), demonstrating that the effect of memory updating was distinct from recall success for correct details. Mirroring these

Downloaded at DUKE UNIVERSITY on December 15, 2021

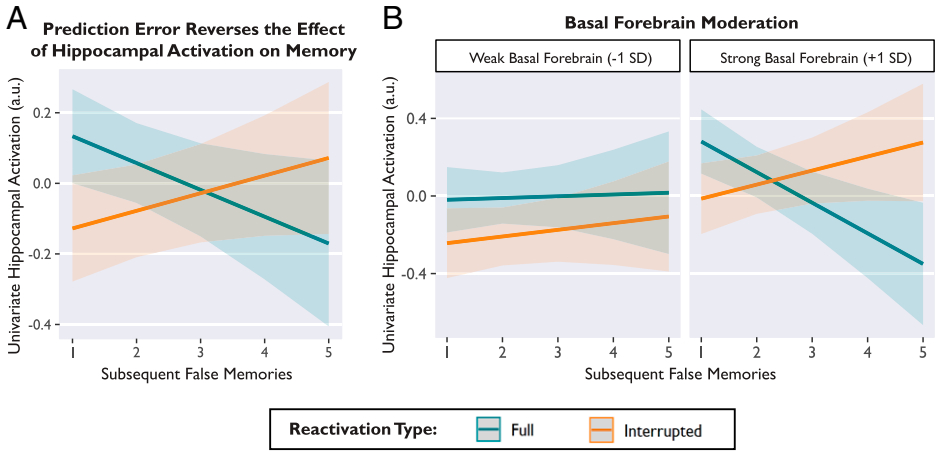

Fig. 3. Prediction error reversed the relationship between average hippocampal activation (arbitrary units, a.u.) and subsequent memory, and this effect depended on concurrent basal forebrain activation. (*A*) After Full videos, hippocampal activation was associated with memory *preservation*, predicting fewer false memories (blue). After Interrupted videos, hippocampal activation was associated with memory *updating*, predicting more false memories (orange). (*B*) The effect of prediction error on hippocampal activation and memory was observed only 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% CI. Model-derived estimates are shown instead of individual data points in order to show within-subjects effects, while controlling for subject and stimulus variance.

results, there was also a significant interaction between reactivation type and correct details predicting hippocampal activation ($\beta = 0.03$, 95% CI [0.01, 0.06], $t = 2.10$, $P = 0.036$). This interaction was driven by a negative association between correct details and hippocampal activation for Interrupted videos, although the slopes for each condition did not significantly differ from zero (Full: $\beta = 0.01$, $z = 0.33$, $P = 0.742$; Interrupted: $\beta = -0.05$, $z = -1.55$, $P = 0.120$). Numerically, the results for correct details contrasted with our results for false memories: after Full videos, hippocampal activation was negatively associated with false memories, but was not significantly associated with correct details; after Interrupted videos, hippocampal activation was positively associated with false memories, but negatively associated with correct details.

Hippocampal-basal forebrain connectivity predicted memory outcomes. Next, we tested hypotheses about neuromodulatory mechanisms by examining activation in the basal forebrain (which contains the medial septal nucleus, the primary source of acetylcholine in the hippocampus) (29, 31, 37) and the VTA (which contains dopaminergic neurons that project to the hippocampus) (16, 24, 56). First, we extracted average basal forebrain and VTA activation during the postvideo period (i.e., the 2-s segment of the postvideo fixation period, consistent with modeling of hippocampal activation).

Using linear mixed-effects regression, we predicted hippocampal activation from the variables reactivation type (Full vs. Interrupted), false memories (continuous), basal forebrain activation (continuous), hemisphere (left vs. right), axis (anterior vs. posterior), and all relevant interactions. There was a significant three-way interaction among basal forebrain activation, reactivation type, and false memories predicting hippocampal activation ($\beta = -0.04$, 95% CI [−0.06, −0.01], $t = -2.66$, $P = 0.008$) (Fig. 3*B*). This interaction demonstrated that the relationship between hippocampal activation and subsequent memory (Fig. 3*A*) was evident only when the basal forebrain was also strongly activated (Fig. 3 *B*, *Right*) (Full: $\beta = -0.12$, $z = -4.01$, $P < 0.001$; Interrupted: $\beta = 0.07$, $z = 2.37$, $P = 0.018$). When basal forebrain activation was weak, hippocampal activation was unrelated to memory (Fig. 3 *B*, *Left*) (Full: $\beta = 0.01$, $z = -0.33$, $P = 0.739$; Interrupted: $\beta = 0.04$, $z = 1.39$, $P = 0.166$). There was also a main effect of basal forebrain activation predicting hippocampal activation during the postvideo period ($\beta = 0.10$, 95% CI [0.02, 0.19], $t = 2.34$, $P = 0.029$). Other results from this extended model were consistent with the base model (without basal forebrain variables) described above. There were no significant interactions with hemisphere or axis. All parameter estimates are provided in *SI Appendix*, Table S8.

Next, we examined the role of VTA activation. We modified the model described above by replacing the basal forebrain activation variable (and interactions) with VTA activation parameters. VTA activation was positively related to hippocampal activation during the postvideo period ($\beta = 0.15$, 95% CI [0.05, 0.25], $t = 3.03$, $P = 0.006$). However, there was no interaction among VTA activation, reactivation type, and false memories ($\beta = -0.004$, 95% CI [−0.03, 0.02], $t = -0.32$, $P = 0.747$). Thus, there was no evidence that VTA moderated the effect of hippocampal activation on memory. All parameter estimates are provided in *SI Appendix*, Table S9. In separate models, we also tested for a main effect of prediction error on basal forebrain activation (*SI Appendix*, Table S10*A*) or VTA activation (*SI Appendix*, Table S10*B*) and found no significant effects.

Multivariate fMRI Results. Overall, our univariate results suggest that prediction error changed the role of the hippocampus: the same magnitude of hippocampal activation predicted opposing effects on memory depending on whether events were expected or surprising. Moreover, this effect of hippocampal activation on memory depended on concurrent basal forebrain activation, consistent with the idea that acetylcholine regulates hippocampal processing (27, 28). On the basis of our univariate findings, we proposed that during video playback, the hippocampus continually generates predictions and sustains episodic representations (57–59). If no prediction error is detected, these representations should be sustained, and the hippocampus should preserve and reinforce the memory (i.e., decreasing false memories). If a prediction error is detected, then the hippocampus should abandon ongoing predictions and prepare to update a memory (i.e., increasing false memories) (28, 29). Therefore, we hypothesized that prediction errors would disrupt sustained representations in the hippocampus, and that disrupting hippocampal representations would lead to memory updating. Furthermore, we predicted that activation in the basal forebrain and VTA would link hippocampal representations to memory

Downloaded at DUKE UNIVERSITY on December 15, 2021

outcomes, via neuromodulation of hippocampal processing (24–29, 34, 60, 61).

Prediction errors disrupted sustained representations in the hippocampus. Past studies in rodents and humans have used autocorrelation measures, which quantify similarity across neural patterns, to investigate hippocampal representations during naturalistic tasks (53, 54). Temporal autocorrelation is an index of multivariate information that is preserved over time; this measures moment-to-moment overlap of activation patterns (53, 58, 62). Intracranial recordings in humans have shown that temporal autocorrelation in the hippocampus ramps up over the course of familiar episodes (58). Ramping autocorrelation reflects sustained neural representations, consistent with the hippocampus generating predictions and anticipating upcoming stimuli (57, 58). 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 s (*Methods*). Importantly, we also investigated autocorrelation in two control regions (inferior lateral occipital cortex and white matter) to demonstrate that these representational changes were not a spurious whole-brain phenomenon (*SI Appendix*, *Autocorrelation Control*,).

First, we tested whether autocorrelation increased during video playback (binned into 5-s video segments over time). We used linear mixed-effects regression to predict hippocampal autocorrelation (averaged over 5-s bins) from the variables video segment (continuous), hemisphere, axis, and all interaction terms. We found that hippocampal autocorrelation increased linearly as videos progressed (β = 0.025, 95% CI [0.01, 0.04], t = 3.41, P = 0.002), suggesting that episodic representations were sustained and stabilized during video playback (Fig. 4*A*) (58). There were no significant effects of hemisphere or axis. All parameter estimates are provided in *SI Appendix*, Table S11*A*.

Next, we tested whether prediction error disrupted this ramping autocorrelation. To analyze postvideo change in autocorrelation, we calculated a difference score for each trial by subtracting the average autocorrelation value from the 5-s bin immediately before video offset from the average autocorrelation value from the 5-s video immediately after video offset. We then used linear mixed-effects regression to predict average postvideo change in autocorrelation from the variables reactivation type, hemisphere, axis, and interaction terms. There was a significant main effect of reactivation type (β = 0.04, 95% CI [0.01, 0.07], t = 2.21, P = 0.038), such that autocorrelation increased after the offset of Full videos but not Interrupted videos (Fig. 4*C*). In other words, prediction errors disrupted the continuity of hippocampal representations. This postvideo divergence is visualized in Fig. 4*B*. There were no significant interactions with hemisphere or axis. All parameter estimates are provided in *SI Appendix*, Table S11*B*.

Disruption of hippocampal autocorrelation predicted false memories. Next, we tested whether disruption of hippocampal representations predicted memory updating. Using linear mixed-effects regression, we predicted subsequent false memories from the variables reactivation type, postvideo change in autocorrelation (continuous), hemisphere, axis, and all relevant interactions. We also included a continuous covariate for univariate hippocampal activation (thus controlling for any autocorrelation effects that may be a consequence of univariate activation). There was a significant interaction between reactivation type and change in autocorrelation predicting false memories (β = 0.04, 95% CI [0.02, 0.07], t = 3.34, P < 0.001) (Fig. 5*A*). After Interrupted videos, disrupting hippocampal representations led to memory updating (β = −0.06, z = −2.89, P = 0.004). Conversely, after Full videos, hippocampal autocorrelation was not related to false memories (β = 0.02, z = 1.03, P = 0.303). There were no significant interactions with hemisphere or axis. All parameters are reported in *SI Appendix*, Table S12*A*.

Basal forebrain activation links hippocampal autocorrelation to memory. What determines whether hippocampal representations are sustained or disrupted? To investigate candidate neuromodulatory mechanisms, we extended the model described above by adding parameters for average postvideo basal forebrain activation and interaction terms. The model included all relevant interaction terms, reported in full in *SI Appendix*, Table S12*B*. Paralleling our univariate findings, we found a significant three-way interaction among basal forebrain activation, reactivation type, and hippocampal autocorrelation that predicted subsequent

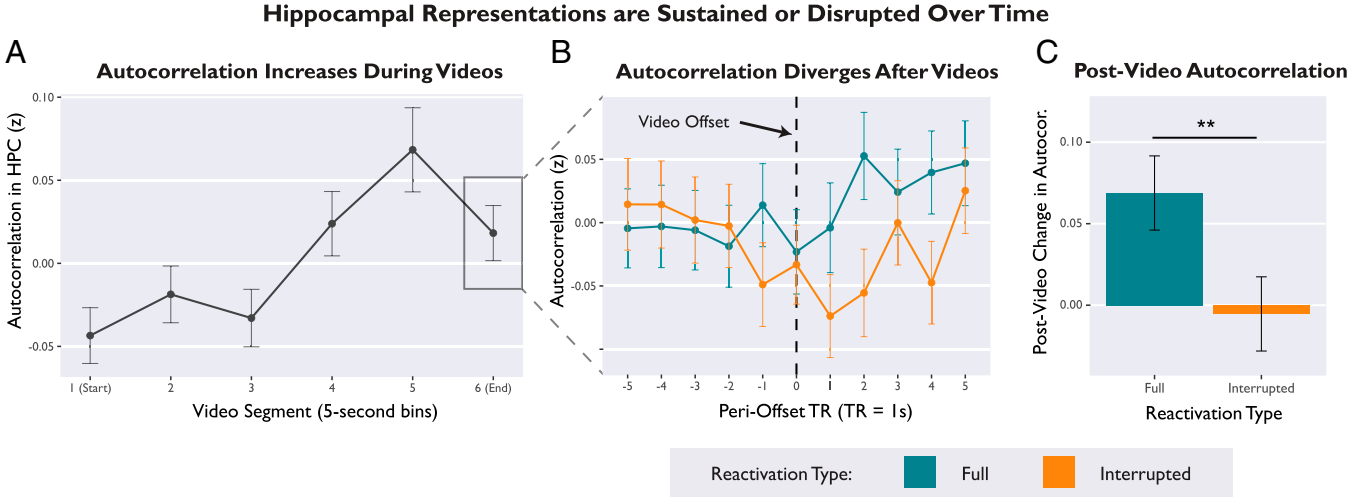

Fig. 4. Hippocampal representations are sustained or disrupted over time, depending on whether or not episodes align with expectations. (*A*) Temporal autocorrelation in the hippocampus gradually increased over the course of a video, suggesting that episodic representations were sustained over time. Autocorrelation values were averaged over 5-s bins of video playback. (*B*) Autocorrelation trajectories for Full and Interrupted videos diverged during the postvideo period. Plot visualizes second-by-second autocorrelation values in the hippocampus, time-locked to the moment of video offset (black dotted line). (*C*) Average postvideo change in autocorrelation (average autocorrelation scores for the 5-s 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. $^{**}P < 0.01$.

Downloaded at DUKE UNIVERSITY on December 15, 2021

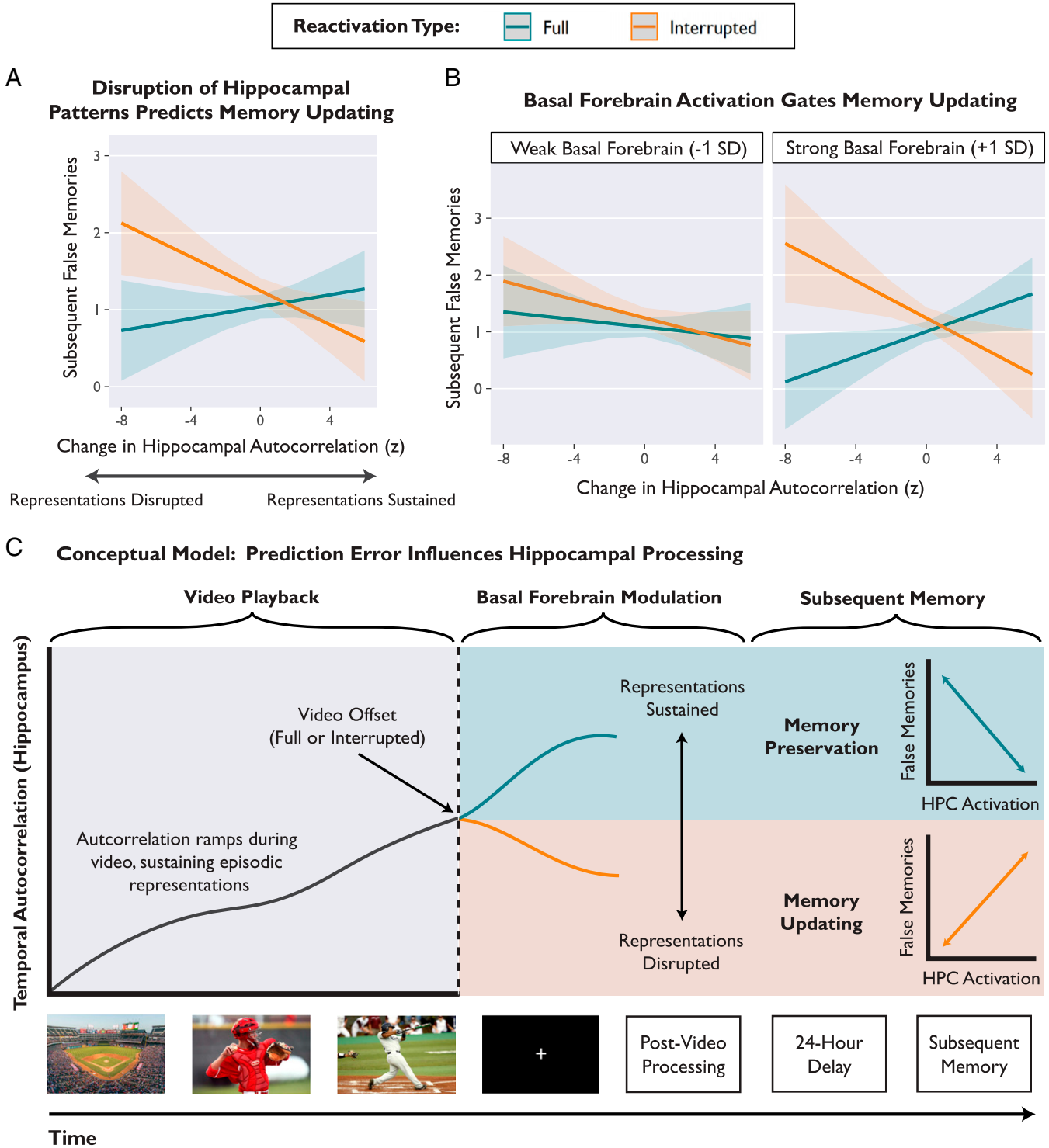

Fig. 5. Prediction errors elicited by Interrupted videos disrupted sustained hippocampal representations, and this disruption predicted memory updating. (*A*) Estimated values from a linear mixed-effects regression model predicting subsequent false memories from the interaction of reactivation type and 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% CIs around the regression line. In *A* and *B*, model-predicted estimates are depicted instead of individual data points in order to show within-subject 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, the hippocampus acts to preserve the memory (representations sustained, univariate activation predicts fewer false memories) or prepare for memory updating (representations disrupted, univariate activation predicts more false memories). The link between prediction error and memory outcomes depends on coactivation of the hippocampus and basal forebrain during the postvideo period.

Downloaded at DUKE UNIVERSITY on December 15, 2021

false memories (β = 0.03, 95% CI [0.01, 0.06], $t = 2.54$, $P = 0.011$) (Fig. 5*B*). In other words, prediction errors disrupted hippocampal representations and led to memory updating, but only when the basal forebrain was also strongly activated (at +1 SD basal forebrain activation, Interrupted: β = −0.07, $z = -2.46$, $P = 0.014$; Full: β = 0.08, $z = 2.92$, $P = 0.004$) (Fig. 5 *B*, *Right*).

Next, we tested whether VTA activation was related to hippocampal autocorrelation and memory. We modified the model described above (predicting subsequent false memories) by replacing the basal forebrain activation variable (and interaction terms) with the VTA activation variable. In contrast to our basal forebrain findings, there was no three-way interaction among VTA activation, reactivation type, and hippocampal autocorrelation (β = −0.01, $t = -0.54$, $P = 0.587$, 95% CI [-0.03, 0.02]). All parameter estimates are reported in *SI Appendix*, Table S12*C*. Overall, we found that our autocorrelation results paralleled our univariate findings: basal forebrain activation, but not VTA activation, was crucial for connecting hippocampal representations to memory outcomes.

Discussion

Here, we show that prediction errors modulate the function of the hippocampus and allow memories to be modified, consistent with an adaptive updating mechanism. 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 postvideo period. We conclude that prediction error, coupled with basal forebrain modulation, prompts the hippocampus to abandon ongoing predictions and prepare to update a memory by incorporating other information present in the environment (Fig. 5).

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, 63). 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, suggesting a shift in hippocampal processing (Fig. 3). After expected endings (Full videos), hippocampal activation protected against false memories, consistent with the idea that the hippocampus reinforces memory after the conclusion of an episode (49, 50). In contrast, after surprising endings (Interrupted videos), hippocampal activation predicted more false memories, consistent with the idea that prediction errors can destabilize memories and enable updating (1–3). Overall, this reversal supports the idea that the hippocampus acts to preserve memories after expected events, but update memories after surprising events. These divergent modes of processing parallel prior studies on encoding vs. retrieval modes (20, 27, 64) and internal vs. external attention (28).

To test the idea that prediction errors influence hippocampal processing, 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 (53, 58, 62). As narratives progressed, autocorrelation increased, reflecting stability and continuity; this increase in autocorrelation suggested that the hippocampus generated predictions (57, 58) and sustained episodic representations over time (49, 65) (Fig. 4*A*). Crucially, prediction errors disrupted the stability of hippocampal representations (Fig. 4 *B* and *C*), and this disruption predicted the degree of memory updating (Fig. 5*A*). Overall, we propose that disruption of hippocampal representations indicates a shift in processing: Prediction error prompts the hippocampus to abandon ongoing predictions and prepare to update a memory (Fig. 5*C*).

Our univariate findings also diverge from past studies of hippocampal prediction error responses, which have shown that hippocampal activation increases after prediction errors (17–19, 63). In contrast, our single-trial analyses showed no significant main effect of reactivation type on hippocampal activation; the effect of prediction error was revealed only when examining the link between hippocampal activation and subsequent memory. Moreover, our whole-brain mass univariate analyses revealed a significant cluster in the left hippocampus that was less activated after Interrupted videos than Full videos (*SI Appendix*, *Whole-Brain Analysis*). Our task elicits surprise by omitting expected endings, comparable to a negative prediction error. Previous studies have elicited surprise by replacing expected stimuli with novel stimuli, thus adding new information, comparable to a positive prediction error. Therefore, this reversal of activation (Full > Interrupted trials) is consistent with prior studies of reward (13) and information prediction errors (4, 66), which have shown increases in neural activation after positive prediction errors and decreases after negative prediction errors. Overall, our results suggest that hippocampal responses may depend on the stimuli and type of surprise, and demonstrate that prediction errors change the effect of hippocampal activation on memory.

Basal Forebrain Activation Relates to Hippocampal Processing. Past studies have suggested that either cholinergic (27–29, 35–37, 60) or dopaminergic (16, 24, 61) modulation could regulate hippocampal function after prediction errors, such as by enhancing plasticity and switching between processing modes. However, mixed evidence supporting both hypotheses has left the question unresolved (25, 26, 31, 33, 34). Here, we investigated whether activation of the basal forebrain or the VTA could explain the relationship between hippocampal activation after prediction error and subsequent memory. We found that the effect of prediction error on memory depended on coactivation of the hippocampus and basal forebrain, suggesting that connectivity between these regions is important for shifting hippocampal processing modes to either preserve or update a memory. Hippocampal activation was associated with memory updating after prediction errors, but only when the basal forebrain was also activated. Likewise, disrupting hippocampal representations led to memory updating after prediction errors, but only when the basal forebrain was also activated. Although fMRI cannot provide direct evidence of neuromodulation, our results are consistent with the idea that cholinergic modulation from the basal forebrain (27, 35–37, 39) influences hippocampal processing and memory outcomes.

Our findings are also relevant to the functional relationship between the VTA and hippocampus. Although we found a robust positive correlation between VTA and hippocampal activation during the postvideo period, VTA activation was unrelated to prediction error and did not link hippocampal activation with memory outcomes. These findings are consistent with our prior

Downloaded at DUKE UNIVERSITY on December 15, 2021

proposal that connectivity between the VTA and hippocampus reflects modulation of hippocampal learning states by sustained VTA activity (16, 67–69) rather than phasic VTA responses (25, 70–72). However, our paradigm was optimized for detecting memory updating instead of midbrain prediction error responses. It is also possible that prediction error signals could be transmitted to the hippocampus from the locus coeruleus (71, 72). Analyses of locus coeruleus in the current data did not reveal any relationships. Future research could disambiguate the roles of the basal forebrain, VTA, and locus coeruleus by examining both event-related and sustained connectivity with the hippocampus and their consequences for memory.

Prediction Error Both Strengthens and Updates Memories. Comparing behavioral results across the Delayed and Immediate groups revealed a dissociation: prediction error both strengthened and updated memories, but over distinct time courses and likely via different mechanisms (Fig. 2). Prediction error increased the number of correct details recalled, both immediately and after a 1-d delay. This finding is consistent with recent evidence that mnemonic prediction errors can promote detailed memories immediately (73), possibly by way of enhanced attention and pattern separation. In contrast, we found that prediction error increased false memories only after a delay, consistent with prior studies that have shown that memory updating requires a delay for protein synthesis to occur (1, 47, 74). Although it is possible that the relatively few false memories in the Immediate group made it more difficult to detect an effect of prediction error, a control analysis showed that excluding low-variance subjects from the Immediate group did not change our results (*SI Appendix*, *Behavioral Variance Control*). Overall, our finding that prediction error increased both correct details and false memories supports the idea that surprise drives adaptive updating.

In the present study, we used false memories as an index of memory updating. In the real world, however, memory updating can be adaptive: new information is not "false" per se, but a relevant addendum to prior knowledge. 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 new videos that were specifically chosen to interfere with reactivated memories (48). Here, we showed that videos that shared greater semantic similarity with the rest of the stimulus set produced more false memories (*SI Appendix*, Table S4). This finding suggests that prediction error increases source confusion or integration of information across related memories. Prior studies have shown that memory updating occurs when interference is introduced after a memory trace is reactivated (1, 47, 75). In our paradigm, interference could arise from semantically related details that result from the subject recalling related memories, or visual input from subsequent videos during the task. This finding accords with prior behavioral studies (1, 47, 75) and computational models of event segmentation (76, 77), which have both shown that interference among related episodes can produce false memories and source confusion after prediction error. However, memory updating is beneficial in other situations that require integrating old and new knowledge, or correcting erroneous information.

One possible mechanism for the memory updating we observed is reconsolidation, the process by which reactivating a memory trace can temporarily render it malleable. However, evidence for cellular reconsolidation processes in humans is lacking. Although previous human studies have used reconsolidation-like paradigms to demonstrate memory malleability (4, 47, 48, 75), it remains unknown whether the synaptic mechanisms of reconsolidation are consistent across animals and humans (1). A key prediction of reconsolidation theory is that memory updating effects should only emerge after a delay, because the process of modifying and restabilizing a memory trace requires protein synthesis that occurs over several hours. We found that the behavioral effect of prediction error on memory updating required a delay, which aligns with this theoretical prediction. Overall, our findings are broadly relevant to research on prediction error and memory, and reconsolidation theory offers one possible framework.

Limitations and Future Directions. We were unable to directly test whether the relationship between neural activation and memory updating required a delay, because the fMRI participants always completed the delayed memory test and may show effects of memory updating as well as forgetting over time. Although our prior behavioral work demonstrated an effect of prediction error that was not confounded with the delay (48), scanning a group of participants who experience a delay-to-test without memory reactivation would enable an investigation of the neural correlates of immediate vs. delayed memory effects.

In the present study, we elicited surprise by interrupting videos before their expected narrative endings. However, we were not able to directly measure subjective surprise, because asking participants to rate surprise after each video would have disrupted other cognitive processes and revealed the goal of the manipulation. To ensure that Interrupted videos would continue to elicit surprise after the first few trials, we included the following features in our experimental design: 1) Full-length videos were presented twice during encoding to set strong expectations; 2) Interrupted videos violated strongly expected action–outcome contingencies (e.g., a baseball batter halted midswing); 3) trials were pseudorandomized so that participants could not anticipate whether each video would be Full or Interrupted; and 4) participants did not know when each video might be interrupted. Consistent with these design considerations, we found that the effect of prediction error on false memories did not interact with trial number, suggesting that surprise did not diminish over the course of the experiment (*SI Appendix*, Table S2).

Prior studies with humans and animals have used incomplete reminders (e.g., a conditioned stimulus without the expected outcome) to elicit prediction error (1, 3, 75). Here, we mimicked this approach by interrupting narrative videos and omitting the expected endings. Incomplete reminders may be particularly effective because they elicit memory reactivation, and memory reactivation supports plasticity (78–80). After Interrupted videos, participants may actively retrieve the missing endings; this memory reactivation could contribute to memory malleability. As discussed in our prior review (1), both prediction error and memory reactivation may interact to support memory updating. Memory reactivation may be a prerequisite for generating a prediction, and experiencing a prediction error may prompt further memory reactivation. Future studies could directly investigate the role of memory reactivation by asking participants to report their ability to recall the missing endings, or by testing encoding-retrieval pattern similarity.

Lastly, a limitation of the present study is that we were unable to determine the temporal distribution or the source of each individual false memory. Most details that participants recalled (both correct and incorrect) pertained to the entire video, such as perceptual details about the characters and setting. Many of these perceptual details were also shared across multiple videos (e.g., several characters with green shirts), making it impossible to determine the source of each specific detail. Interestingly, we found that the omitted endings on Interrupted trials were rarely associated with false memories. Recall of the narrative content of the video endings was very accurate: in the Delayed group, only 2% of all false memories specifically

Downloaded at DUKE UNIVERSITY on December 15, 2021

pertained to the missing ending. Because the salient action–outcome contingencies define the videos, even when interrupted, these central details were not likely to be affected. Instead, we found that prediction error was more likely to induce a holistic distortion of perceptual content from throughout each video (e.g., details about the setting or a character).

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 mnemonic prediction errors prompt the hippocampus to abandon ongoing predictions and update memories by incorporating relevant details from subsequent experiences. In this way, surprise modulates hippocampal processing and determines the fate of episodic memories. This theoretical framework of memory updating bears implications for eyewitness testimony, education, and treating pathological memories (e.g., in posttraumatic stress disorder). Beyond memory research, our results offer insights for theories on the whole-brain predictive processes that govern attention, perception, action, and decision-making.

Methods

Data, Code, and Materials. Brief descriptions of the stimulus videos are provided in *SI Appendix*, Table S13. 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/). Additionally, fMRI data are available in a repository hosted by OpenNeuro (DOI: 10.18112/openneuro.ds003835.v1.0.2).

Participants. We recruited 55 paid participants from the University of Toronto community (Delayed group: $70, Immediate group: $40). Seven participants were excluded (*SI Appendix, Exclusions*), yielding a final sample of 48 participants. The sample size was determined a priori to satisfy the following conditions: 1) achieve at least 90% power to detect the interaction effect found in a prior study ($\eta_p^2 = 0.17$) (48), and 2) evenly allocate participants to six pseudorandomized trial order lists. Participants were healthy young adults (age: mean = 22.42, SD = 2.41, range [18 to 30]; gender: 75% female, 25% male). Inclusion criteria were as follows: between the ages of 18 and 30, normal or corrected-to-normal vision and hearing, no history of neurological or psychiatric disorders, and fluency in English. fMRI participants were all right-handed. All participants provided informed consent, and the study was approved by the University of Toronto Institutional Review Board, Protocol #00035787.

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 h (SD = 1.31) between the day 1 and day 2 sessions, and Delayed group participants slept an average of 7.02 h (SD = 1) between the day 2 and day 3 sessions.

Stimuli. We used 70 videos that featured distinct narrative events (duration mean = 30 s, SD = 7 s). The Interrupted version of each video ended abruptly at the narrative climax, omitting the salient ending and violating expectations (duration mean = 25 s, SD = 4 s). For more information, refer to *SI Appendix* (*SI Appendix, Supplementary Methods* and Table S3) and the additional materials in an online repository (https://osf.io/xb7sq/).

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 (35 Full videos, 35 Interrupted videos). Videos were played in a pseudorandom order (six trial order lists, counterbalanced across participants) such that there were never more than two consecutive Interrupted videos. Participants could not reliably anticipate whether each video would be Interrupted, or where the interruption might occur. Additionally, Full and Interrupted versions of each video were counterbalanced across participants. We also performed eye-tracking during the encoding and reactivation sessions for participants in both the Delayed 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.

Finally, 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 asked a predetermined list of 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 updating, we instructed participants not to guess and permitted them to skip details they could not recall.

Overall, the experiment took place over 3 d for participants in the Delayed group (24-h delays between encoding, reactivation, and test), or over 2 d for participants in the Immediate group (24-h delay between encoding and reactivation, no delay between reactivation and test). Only the Delayed group underwent neuroimaging.

Consistent with past studies (81–83), we maintained consistent contextual factors between encoding, reactivation, and test sessions. Delayed group participants completed the encoding session in a mock scanner. The mock scanner room was adjacent to the real scanner room and similar in appearance. Delayed group participants completed the reactivation session in the real scanner and the test session at a desk in the mock scanner room. Participants in the Immediate group completed all three sessions in the same behavioral testing room.

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 2.7-mm isotropic voxels (repetition time [TR]: 1,000 ms, echo time [TE]: 29 ms, flip angle: 50°, 60 slices at transversal orientation, phase encoding: A > P, field-of-view [FoV]: 210 mm, partial Fourier: 0.875, multiband factor: 4). High-resolution T1-weighted anatomical images were acquired with a magnetization-prepared rapid-acquisition gradient-echo pulse sequence (voxel size: 1 mm isotropic, TE: 24 ms, TR: 2,000 ms, inversion time [TI]: 1,100 ms, flip angle: 9°) to allow three-dimensional reconstruction and volume-based statistical analysis.

Statistical Analysis. For both behavioral and neural data, we conducted trial-wise analyses with linear mixed-effects regression models. Details about model construction are provided in the *SI Appendix, Supplementary Methods*).

Univariate fMRI analyses. Preprocessing steps, whole-brain mass univariate results, and ROI masks are reported in *SI Appendix, fMRI Preprcessing*, *Whole-Brain Analysis*, Fig. S2, and Table S5). To model neural responses on each individual trial, we employed the least-squares separate approach and constructed a separate generalized linear model (GLM) for each trial (84, 85). We modeled each trial as a 2-s impulse in the postvideo period, convolved with the canonical double-γ HRF and phase-shifted 2 s after video offset. This 2-s shift targeted the peak hippocampal response previously identified in studies of postvideo processing (49, 50). Within each GLM, the target trial (2-s event) was isolated as one regressor, and all other events were modeled with a separate regressor for each type of event (e.g., video playback, video name cues, other fixation periods). For each trial, we masked the processed data and averaged across voxels within each ROI to yield an average activation value.

Multivariate fMRI analyses. Multivariate temporal autocorrelation analyses (53, 58) were conducted on preprocessed data (prior to single-trial GLM analysis). We extracted the whole-run time series from every voxel within each ROI using the *fslmeants* utility. For control analyses (white matter and iLOC ROIs), autocorrelation was calculated on 200 contiguous voxels, approximately matching the size of the hippocampal ROIs. For every TR, we calculated temporal autocorrelation as the Pearson product-moment correlation between all voxel activation values at timepoint T and timepoint T+1 s. Autocorrelation values were then standardized (Fisher's z).

Next, we aligned multivariate time series data with event onset and duration markers. Comparable to past research, we phase-shifted the time series by 4 s to account for HRF lag (86). This manual shifting is necessary because event onset regressors have not been convolved with the HRF (unlike in standard GLM analyses used for our univariate analyses). Note that our univariate analyses included a 2-s shift in addition to the standard HRF shift; this allowed us to target the peak postvideo hippocampal response on each trial, but was not necessary for the autocorrelation analyses that yield TR-by-TR values.

After alignment, we calculated average autocorrelation values that were time-locked to events. For statistical analyses, autocorrelation values were averaged across 5-s bins during and after each video. To analyze signal history

Downloaded at DUKE UNIVERSITY on December 15, 2021

over the course of video playback, we related the video segment number (5-s bins) to average autocorrelation values. For each video, we included the first 5-s (timepoints 0 to 4), the next four middle segments (timepoints 5 to 9, 10 to 14, 15 to 19, and 20 to 24), and the last 5 s (variable depending on the length of the video). This binning scheme spanned the average video length of 30 s; additional middle segments from videos >30 s were omitted from this analysis. To analyze postvideo changes in autocorrelation, we calculated trial-by-trial difference scores by subtracting the average autocorrelation values for the 5-s bin immediately before video offset and the 5-s bin immediately after video offset.

Data Availability. Raw neuroimaging data have been deposited in OpenNeuro (https://doi.org/10.18112/openneuro.ds003835.v1.0.2) (87). Anonymized data and code necessary to reproduce all results have been deposited in the Open Science Framework (https://osf.io/XB7SQ/) (88).

ACKNOWLEDGMENTS. We thank Carolyn Chung, Tolulemi Gbile, and Aria Fallah for their invaluable contributions to data collection, transcription, and scoring; and Jia-Hou Poh for helpful comments on the manuscript. This research was funded by grants awarded from the James S. McDonnell Foundation (Scholar Award in Understanding Human Cognition, to M.D.B.) and the Natural Sciences and Engineering Research Council of Canada Discovery Grant and Accelerator Supplement, RGPIN-2014-05959 and RGPIN-2020-05747 (to M.D.B.). A.H.S. has been supported by awards from the NSF (Graduate Research Fellowship) and the Natural Sciences and Engineering Research Council of Canada (Postgraduate Doctoral Scholarship, Undergraduate Student Research Award).

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**, 1673 (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. Lond. 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* (MIT Press, ed. 2, 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. C. Hindy, E. W. Avery, N. B. Turk-Browne, Hippocampal-neocortical interactions sharpen over time for predictive actions. *Nat. Commun.* **10**, 3989 (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. K. C. Dickerson, J. Li, M. R. Delgado, Parallel contributions of distinct human memory systems during probabilistic learning. *Neuroimage* **55**, 266–276 (2011).
24. J. E. Lisman, A. A. Grace, The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron* **46**, 703–713 (2005).
25. 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).
26. 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).
27. M. E. Hasselmo, B. P. Wyble, G. V. Wallenstein, Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* **6**, 693–708 (1996).
28. 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).
29. 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).
30. B. E. Sherman, N. B. Turk-Browne, Statistical prediction of the future impairs episodic encoding of the present. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22760–22770 (2020).
31. 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).
32. M. M. Chun, J. D. Golomb, N. B. Turk-Browne, A taxonomy of external and internal attention. *Annu. Rev. Psychol.* **62**, 73–101 (2011).
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. 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).
35. 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).
36. 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).
37. M. E. Hasselmo, The role of acetylcholine in learning and memory. *Curr. Opin. Neurobiol.* **16**, 710–715 (2006).
38. 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).
39. A. L. Decker, K. Duncan, Acetylcholine and the complex interdependence of memory and attention. *Curr. Opin. Behav. Sci.* **32**, 21–28 (2020).
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* **24**, 2160–2168 (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. K. Nader, E. O. Einarsson, Memory reconsolidation: An update. *Ann. N. Y. Acad. Sci.* **1191**, 27–41 (2010).
47. 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.
48. A. H. Sinclair, M. D. Barense, Surprise and destabilize: Prediction error influences episodic memory reconsolidation. *Learn. Mem.* **25**, 369–381 (2018).
49. 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).
50. 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).
51. R. A. Cooper, M. Ritchey, Progression from feature-specific brain activity to hippocampal binding during episodic encoding. *J. Neurosci.* **40**, 1701–1709 (2020).
52. 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**, 3980 (2020).
53. I. K. Brunec *et al.*, Multiple scales of representation along the hippocampal anteroposterior axis in humans. *Curr. Biol.* **28**, 2129–2135.e6 (2018).

Downloaded at DUKE UNIVERSITY on December 15, 2021

54. K. B. Kjelstrup *et al.*, Finite scale of spatial representation in the hippocampus. *Science* **321**, 140–143 (2008).
55. C. Baldassano *et al.*, Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).
56. M. Zubair *et al.*, Divergent whole brain projections from the ventral midbrain in macaques. *Cereb. Cortex* **31**, 2913–2931 (2021).
57. 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).
58. 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).
59. A. Tambini, L. Davachi, Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19591–19596 (2013).
60. K. H. Jardine *et al.*, Activation of cortical M_1 muscarinic receptors and related intracellular signaling is necessary for reactivation-induced object memory updating. *Sci. Rep.* **10**, 9209 (2020).
61. J. I. Rossato *et al.*, State-dependent effect of dopamine D_1/D_5 receptors inactivation on memory destabilization and reconsolidation. *Behav. Brain Res.* **285**, 194–199 (2015).
62. R. V. Raut, A. Z. Snyder, M. E. Raichle, Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 20890–20897 (2020).
63. D. Kumaran, E. A. Maguire, Match mismatch processes underlie human hippocampal responses to associative novelty. *J. Neurosci.* **27**, 8517–8524 (2007).
64. K. Duncan, A. Tompary, L. Davachi, Associative encoding and retrieval are predicted by functional connectivity in distinct hippocampal area CA1 pathways. *J. Neurosci.* **34**, 11188–11198 (2014).
65. S. DuBrow, L. Davachi, Temporal binding within and across events. *Neurobiol. Learn. Mem.* **134** (Pt A), 107–114 (2016).
66. E. S. Bromberg-Martin, O. Hikosaka, Lateral habenula neurons signal errors in the prediction of reward information. *Nat. Neurosci.* **14**, 1209–1216 (2011).
67. 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* **27**, 1660–1669 (2017).
68. 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).
69. 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).
70. 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).
71. C. G. McNamara, D. Dupret, Two sources of dopamine for the hippocampus. *Trends Neurosci.* **40**, 383–384 (2017).
72. 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. U.S.A.* **113**, 14835–14840 (2016).
73. O. Bein, N. A. Plotkin, L. Davachi, Mnemonic prediction errors promote detailed memories. Learn. Mem. 28, 422-434 (2021).
74. K. Nader, Memory traces unbound. *Trends Neurosci.* **26**, 65–72 (2003).
75. 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).
76. 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).
77. G. A. Radvansky, J. M. Zacks, Event boundaries in memory and cognition. *Curr. Opin. Behav. Sci.* **17**, 133–140 (2017).
78. 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).
79. D. Stawarczyk, C. Wahlheim, J. Etzel, A. Snyder, J. Zacks, Aging and the encoding of changes in events: The role of neural activity pattern reinstatement. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29346–29353 (2020).
80. D. Stawarczyk, M. A. Bezdek, J. M. Zacks, Event representations and predictive processing: The role of the midline default network core. *Top. Cogn. Sci.* **13**, 164–186 (2019).
81. 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).
82. A. Hupbach, R. Gomez, L. Nadel, Episodic memory updating: The role of context familiarity. *Psychon. Bull. Rev.* **18**, 787–797 (2011).
83. 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).
84. 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).
85. J. A. Mumford, T. Davis, R. A. Poldrack, The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage* **103**, 130–138 (2014).
86. T. Sadeh, J. Chen, Y. Goshen-Gottstein, M. Moscovitch, Overlap between hippocampal pre-encoding and encoding patterns supports episodic memory. *Hippocampus* **29**, 836–847 (2019).
87. A. H. Sinclair, M. Barense, Prediction Errors Disrupt Hippocampal Representations and Update Episodic Memories. OpenNeuro. https://doi.org/10.18112/openneuro.ds003835.v1.0.2. Accessed 30 November 2021.
88. A. H. Sinclair, G. Manalili, I. Brunec, R. A. Adcock, M. Barense, Dataset: Prediction Errors Disrupt Hippocampal Representations and Update Episodic Memories. Open Science Framework. https://doi.org/10.17605/OSF.IO/XB7SQ. Accessed 30 November 2021.

Downloaded at DUKE UNIVERSITY on December 15, 2021