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

* Corresponding Author: allie.sinclair@duke.edu

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

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 temporal continuity of hippocampal patterns. We propose that prediction errors disrupt sustained representations and destabilize memories, enabling updating. 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 incorporate 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 reveal this missing link. In a human fMRI study, we elicited mnemonic prediction errors by interrupting familiar narrative videos immediately before the expected endings. We found that prediction error 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 showed that univariate activation was insufficient for understanding hippocampal prediction error signals. We explained this surprising finding by tracking both 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, consistent with 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 abandon ongoing predictions and make memories malleable.

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 discover 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 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* 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 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.

A separate body of research has demonstrated how memories can be updated. Animal and human research on *reconsolidation* has shown that recalling a memory can temporarily destabilize the memory trace in the brain, creating a window of opportunity for memory updating (33–36). Prediction error is crucial for destabilizing and updating memories (1, 3, 37). Reconsolidation paradigms have elicited prediction errors by reactivating memories with incomplete reminders, which are imperfect replications of the original encoding experience (e.g., presenting a conditioned stimulus without the expected outcome) (1, 3, 38, 39). Yet, past reconsolidation studies have not measured the implied neural prediction error signals (in the hippocampus or elsewhere) or linked them to memory updating.

Neuromodulation may be a critical factor that regulates hippocampal processing and enables memory updating. Currently, there is mixed evidence supporting two hypotheses: acetylcholine and/or dopamine could act upon the hippocampus to regulate processing after surprising events (24–27, 29, 31, 40). 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, 41–44), thus allowing stored memories to be compared with perceptual input (31, 44, 45). After prediction errors, acetylcholine release could change

hippocampal processing and enhance encoding or memory updating (26, 29, 40, 43, 45). 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 co-activated after surprising events (46, 47). Other work has shown that co-activation of hippocampus and VTA predicts memory encoding and integration (48–51). Similarly, reconsolidation studies have suggested that both dopamine and acetylcholine may contribute to destabilizing memories, especially hippocampally-dependent memories (37, 52, 53). Overall, extant evidence suggests that basal forebrain and VTA neuromodulation are both candidate mechanisms for regulating hippocampal processing and memory updating.

In the present study, we adapted methods from reconsolidation paradigms (which are optimized to update memories) to identify the missing link between hippocampal prediction error signals and memory updating. Using an fMRI task with human participants, 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*, 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. 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 (36, 54). Therefore, the Immediate control group should not exhibit any memory effects that require protein synthesis-dependent reconsolidation.

Our approach allowed us to test several questions set up by the prior literature. First, to link hippocampal prediction error signals to memory updating, we used a reconsolidation-

inspired paradigm to create and then distort naturalistic episodic memories. Second, to link memory updating to 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 processing and memory updating to activation in the basal forebrain and VTA.

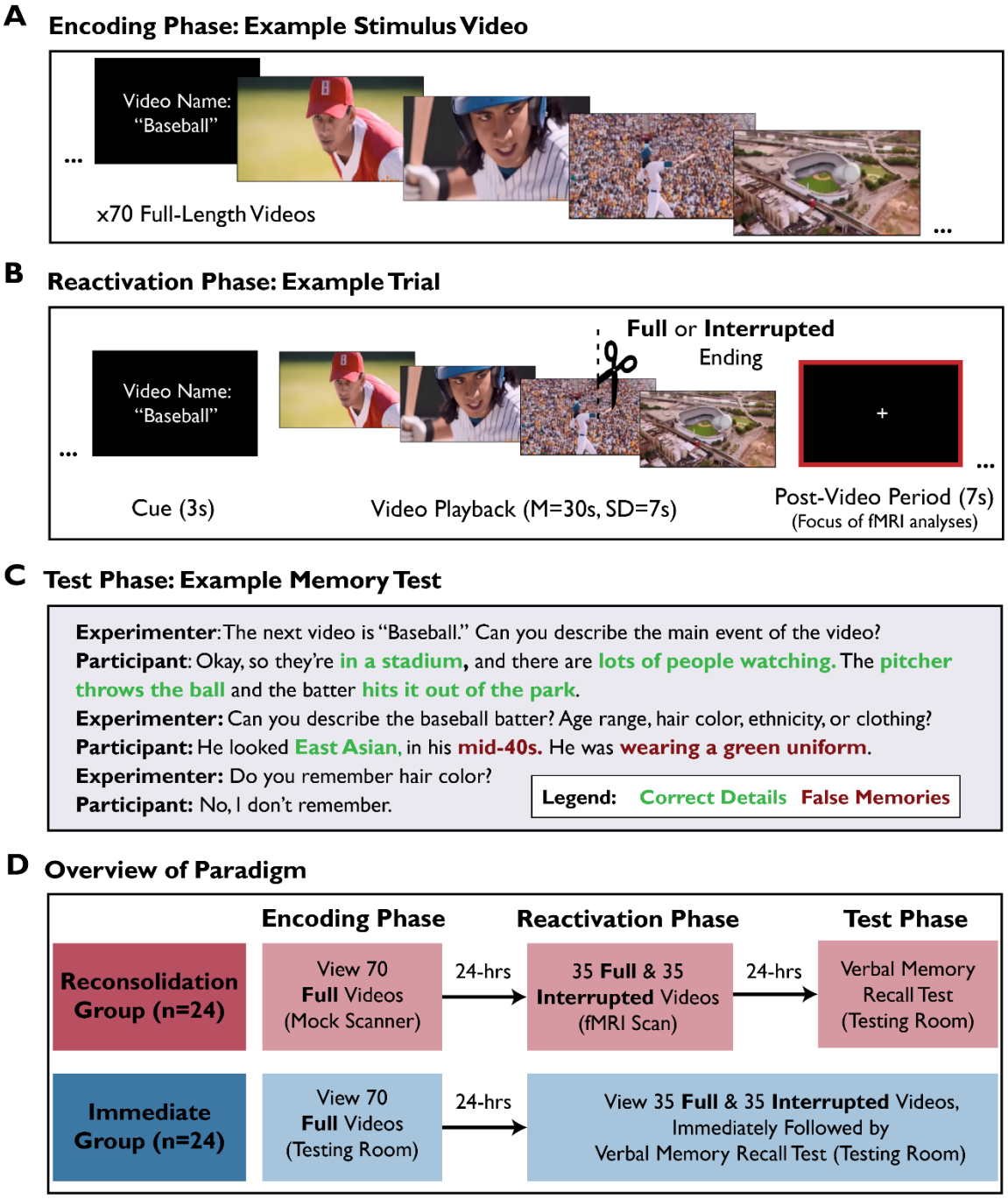

Figure 1. Overview of experimental paradigm. A) During the Day 1 Encoding session, all videos were presented in Full-length form. Here we show frames from a stimulus video named "Baseball", depicting a home run. B) During the Day 2 Reactivation session, participants viewed the videos again, but half 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 (red 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 (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 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) – reflecting memory preservation and updating, respectively. We also collected *confidence ratings* and scored the number of *forgotten videos* (Supplementary Information, *Confidence and Forgetting*, Figure S1). 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. 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 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 (Supplementary Information, *Supplementary Methods, Linear Mixed Effects Regression*).

Prediction Error Increased 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 ($\beta = -0.07$, 95% CI [-0.12, -0.02], $t = -2.75$, $p = .008$) (Table S1A). 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 ($\beta = 0.16$, 95% CI [0.01, 0.31], $t = 2.16$, $p < .036$), likely because the Reconsolidation group completed the memory test after a 24-hour delay. There was no interaction between group and reactivation type ($\beta = -0.01$, 95% CI [-0.04, 0.02], $t = -0.68$, $p = .495$), 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 after a 24-hour delay in the Reconsolidation group, replicating our past behavioral results (39) (significant interaction between reactivation type and group ($\beta = 0.04$, 95% CI [0.01, 0.07], $t = 2.61$, $p = .010$) (Figure 2B, Table S1B). In other words, Interrupted videos increased false memories in the Reconsolidation group ($\beta = -0.09$, $p < .001$), but not the Immediate group ($\beta = -0.01$, $p = .437$) . We also found main effects of group ($\beta = -0.36$, 95% CI [-0.43, -0.29], $t = -9.68$, $p < .001$), and

reactivation type (β = -0.05, 95% CI [-0.08, -0.02], t = -3.31, p = .001), both driven by the effect of prediction error increasing false memories in the Reconsolidation 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 supplementary analysis that included variables for trial number and interactions (Table S2). There was no main effect of trial number on false memories (β = 0.01, 95% CI [-0.03, 0.04], t = 0.48, p = .730), no interaction between trial number and reactivation type (β = 0.00, 95% CI [-0.03, 0.03], t = -0.09, p = .997), and no interaction among trial number, reactivation type, and group (β = 0.00, 95% CI [-0.03, 0.03], t = -0.40, p = .987). Overall, these null results suggest that the effect of Interrupted videos on false memories did not diminish over the course of the experiment.

In sum, our behavioral results showed a novel dissociation between reinforcing and updating memories: Prediction errors during memory reactivation strengthened memories, evident both immediately and after a delay (Figure 2A). However, memory updating (as revealed by false memories) was not evident until after a delay (Figure 2B), as predicted by reconsolidation theory.

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 scale) the degree of surprise elicited by the narrative interruptions (Supplementary Information, *Supplementary Methods, Online Video Ratings*). We found that whereas surprise ratings were unrelated to correct details recalled by our laboratory sample (Table S3), surprise was related to false memories: There was a significant interaction such that more surprising videos were associated with more false memories, selectively in the Reconsolidation group (β = -0.03, 95% CI [-0.06, -0.01], t = -2.07, p = .039). However, the

three-way interaction among surprise ratings, group, and reactivation type was not significant ($\beta = 0.01$, 95% CI [-0.02, 0.04], $t = 0.75$, $p = .451$). Because our surprise ratings were collected from a separate online sample, this measure may not be sensitive enough to detect the expected three-way interaction when applied to the laboratory sample.

In the current study, we index 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 (Supplementary Information, *Supplementary Methods*, *Scoring of Memory Tests*). All model parameters are reported in 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], $t = 3.09$, $p = .003$) (Table S4). There was an interaction between semantic similarity and group predicting false memories ($\beta = -0.04$, 95% CI [-0.07, -0.01], $t = -2.77$, $p = .006$), such that the effect of semantic similarity was stronger in the Reconsolidation group ($\beta = 0.16$, $z = 3.90$, $p < .001$) than in the Immediate group ($\beta = 0.07$, $z = 1.80$, $p = .073$). There was also an interaction between semantic similarity and reactivation type predicting false memories ($\beta = -0.03$, 95% CI [-0.06, -0.01], $t = -2.08$, $p = .038$), such that the effect of semantic similarity was stronger for Interrupted videos ($\beta = 0.15$, $z = 3.44$, $p < .001$) than for Full videos ($\beta = 0.08$, $z = 2.18$, $p = .030$). However, the three-way interaction among reactivation type, group, and semantic similarity was not significant ($\beta = 0.02$, 95% CI [-0.02, 0.05], $t = 0.99$, $p = .323$). 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 Reconsolidation group. Consistent with an adaptive updating process, memories may have been updated with relevant information from other videos.

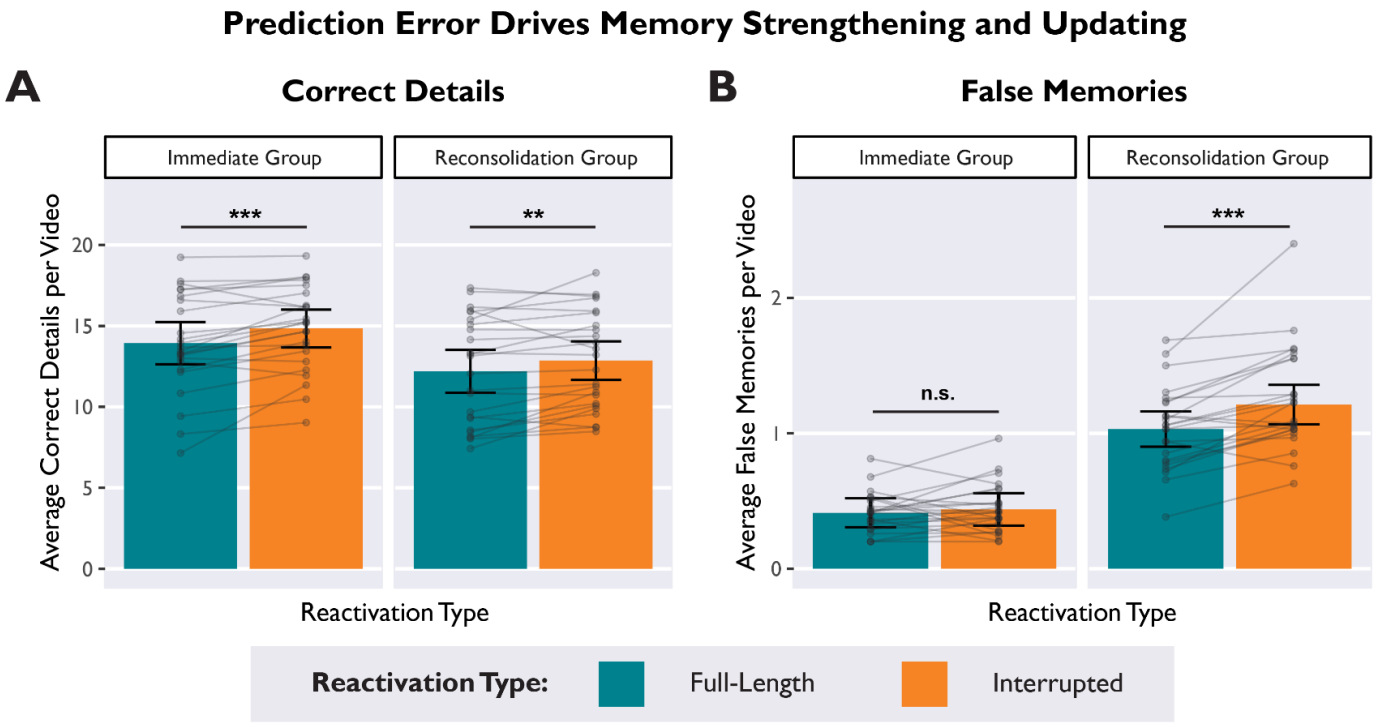

Figure 2. Prediction errors strengthened and updated memories over distinct time-courses. A) In both groups, average Correct Details were higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error can strengthen memory recall both immediately and after a delay. B) In the Reconsolidation 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 that permitted reconsolidation. 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% 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). Whole-brain mass univariate results are provided in the Supplementary Information (*Whole-Brain Analysis*, Table S5, Figure S2). Importantly, we controlled for differences in visual and auditory input by analyzing neural activation during the Post-Video fixation period when visual and auditory input was identical across conditions (Figure 2B).

Some past studies have shown that prediction error signals are stronger in left hippocampus and anterior hippocampus (18, 20, 21, 55), whereas posterior hippocampus is more sensitive to video offsets (56). Other studies have shown that anterior and posterior hippocampus parse continuous experience at different timescales (57, 58). On the basis of these findings, we tested separate (non-overlapping) ROIs for left anterior, right anterior, left posterior, and right posterior hippocampus (Supplementary Information, *Supplementary Methods, ROI Masks*). Activation estimates from these four ROIs are included within each model to test for left/right and anterior/posterior differences.

The Effect of Hippocampal Activation on Memory Depended on Prediction Error

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 targeted the peak Post-Video hippocampal response identified in previous studies (59, 60). We isolated BOLD activation during the Post-Video period on each trial and averaged parameter estimates across all voxels

within each hippocampal ROI (Methods, *Univariate fMRI Analyses*). 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 interactions.

We found a significant interaction between reactivation type and subsequent false memories associated with hippocampal activation (β = -0.06, 95% CI [-0.09, -0.04], t = -4.51, p < .001) (Figure 3A) (Table S6). As predicted, 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 (Figure 3A, blue; β = -0.07, z = -2.78, p = .006), consistent with the idea that the hippocampus reinforces memory for episodes that just concluded (59–61). However, we observed the opposite effect when events were surprising. After Interrupted videos, greater hippocampal activation was associated *more* subsequent false memories (Figure 3A, orange; β = 0.05, z = 2.26, p = .024), 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 (β = 0.04, 95% CI [-0.01, 0.09], t = 1.64, p = .106), nor the main effect of false memories on hippocampal activation (β = -0.01, 95% CI [-0.05, 0.03], t = -0.43, p = .673) was significant. These null results underscore the value of examining the effect of prediction error on both hippocampal activation and memory outcomes. Additionally, there were no significant main effects of hemisphere or

long axis position, nor any other significant interactions, indicating that effects did not differ across hippocampal ROIs. Parameter estimates for all variables are provided in Table S6.

Lastly, although our primary research questions pertained to false memories, we also tested a model that included both false memories and correct details (Table S7). We found that correct details also interacted with reactivation type to predict hippocampal activation, but contrasted with our results for false memories ($\beta = 0.03$, 95% CI [0.01, 0.06], $t = 2.31$, $p = .021$). After Full videos, hippocampal activation was positively associated with correct details but negatively associated with false memories. After Interrupted videos, these associations reversed direction: hippocampal activation was positively associated with false memories but negatively associated with correct details. However, the effect of hippocampal activation on correct details was weaker than for false memories, and the slopes for each reactivation type did not significantly differ from zero when tested independently (Full: $\beta = 0.01$, $z = 0.31$, $p = .759$; Interrupted: $\beta = -0.06$, $z = -1.90$, $p = .058$). Importantly, the effect of hippocampal activation on false memories remained unchanged after controlling for correct details (Full: $\beta = -0.07$, $z = -2.52$, $p = .012$; Interrupted: $\beta = 0.05$, $z = 2.03$, $p = .043$), demonstrating that the effect of memory updating was distinct from recall success for correct details (Table S7).

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, 43) and the VTA (which contains dopaminergic neurons that project to the hippocampus) (16, 24, 62). First, we added a parameter to the model described above to examine average basal forebrain activation during the Post-Video period (i.e., the 2-second segment of the Post-Video fixation period, consistent with modelling of

hippocampal activation). We also tested whether basal forebrain activation interacted with *reactivation type*, *false memories, hemisphere,* or *axis* to predict hippocampal activation.

There was a significant three-way interaction among basal forebrain activation, reactivation type, and false memories predicting hippocampal activation (β = -0.05, 95% CI [-0.07, -0.02], t = -3.34, p < .001) (Figure 3B). This interaction demonstrated that the relationship between hippocampal activation and subsequent memory (Figure 3A) was evident only when the basal forebrain was also strongly activated (Figure 3B, right) (Full: β = -0.14, z = -4.3, p < .001; Interrupted: β = 0.07, z = 2.4, p = .015). When basal forebrain activation was weak, hippocampal activation was unrelated to memory (Figure 3B, left) (Full: β = 0.01, z = 0.25, p = .800; Interrupted: β = 0.04, z = 1.24, p = .213). There was also a four-way interaction among hemisphere, basal forebrain activation, reactivation type, and false memories predicting hippocampal activation; this indicated that the three-way interaction described above was stronger in right hippocampus than in left hippocampus (β = -0.04, 95% CI [-0.08, -0.01], t = -2.32, p = .021). There was no effect of long axis position. All parameter estimates are provided in 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. We found that VTA activation was positively related to hippocampal activation during the Post-Video period (β = 0.16, 95% CI [0.05, 0.28], t = 2.90, p = .008). However, there was no interaction among VTA activation, reactivation type, and false memories (β = 0.002, 95% CI [-0.05, -0.05], t = 0.06, p = .953). Thus, there was no evidence that VTA moderated the effect of hippocampal activation on memory. All parameter estimates are provided in Table S9.

In a separate model, we also tested for a main effect of prediction error on VTA or basal forebrain activation, and found no significant effects (Table S10).

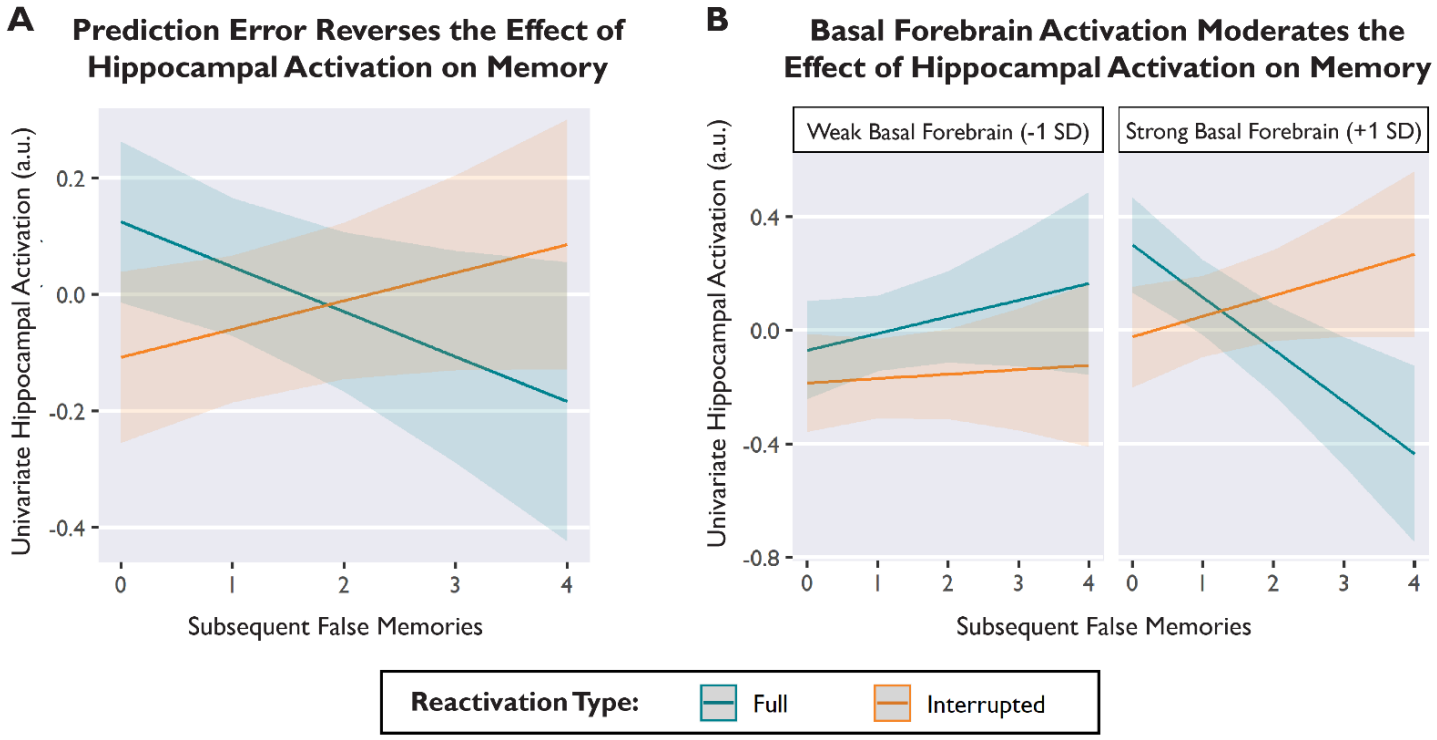

Figure 3. Prediction error reversed the relationship between average hippocampal activation (arbitrary units, a.u.) and subsequent memory, and this effect depended on concurrent basal forebrain activation. A) After Full-length videos, hippocampal activation was associated with memory *preservation*, 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% confidence interval. Model-derived estimates are shown instead of individual data points in order to show within-subjects effects, while controlling for subject and stimulus variance.

Multivariate fMRI Results

Overall, our univariate results suggest that prediction error changed the role of the hippocampus: the same magnitude of univariate 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. On the basis of our univariate findings, we propose that during video playback, the hippocampus continually generates predictions and sustains representations during an ongoing episode (63–65). 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/or VTA would link hippocampal representations to memory outcomes, via neuromodulation of hippocampal processing (24–29, 37, 52, 53).

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 (57, 58). *Temporal autocorrelation* is an index of multivariate information that is preserved over time; this measures moment-to-moment overlap of activation patterns (57, 64, 66). Intracranial recordings in humans have shown that temporal autocorrelation in the hippocampus ramps up over the course of familiar episodes (64). Ramping autocorrelation reflects sustained neural representations, consistent with the hippocampus generating predictions and anticipating upcoming stimuli (63, 64). 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*).

First, we tested whether autocorrelation increased during video playback. We used linear mixed effects regression to predict *hippocampal autocorrelation* (averaged over 5s bins) from the variables *video segment, hemisphere, axis,* and all interaction terms. We found that hippocampal autocorrelation increased linearly as videos progressed ($\beta = 0.024$, 95% CI [0.01, 0.04], $t = 3.37$, $p = .003$), suggesting that episodic representations were sustained and stabilized during video playback (Figure 4A) (64). There were no significant effects of hemisphere or axis. All parameter estimates are provided in Table S10A.

Next, we tested whether prediction error disrupted this ramping autocorrelation. To analyze Post-Video *change in autocorrelation*, we calculated a difference score for each trial by subtracting the average autocorrelation value from the 5s bin immediately *before* video offset from the average autocorrelation value from the 5s video immediately *after* video offset. We then used linear mixed effects regression to predict average *change in autocorrelation* from the variables *reactivation type*, *hemisphere*, *axis*, and interaction terms. There was a significant main effect of reactivation type ($\beta = 0.05$, 95% CI [0.02, 0.08], $t = 2.97$, $p = .006$), such that autocorrelation increased after the offset of Full videos but not Interrupted videos (Figure 4C). In other words, prediction errors disrupted the continuity of hippocampal representations. This Post-Video divergence is visualized in Figure 4B. There were no significant interactions with hemisphere or axis; all parameter estimates are provided in Table S10B.

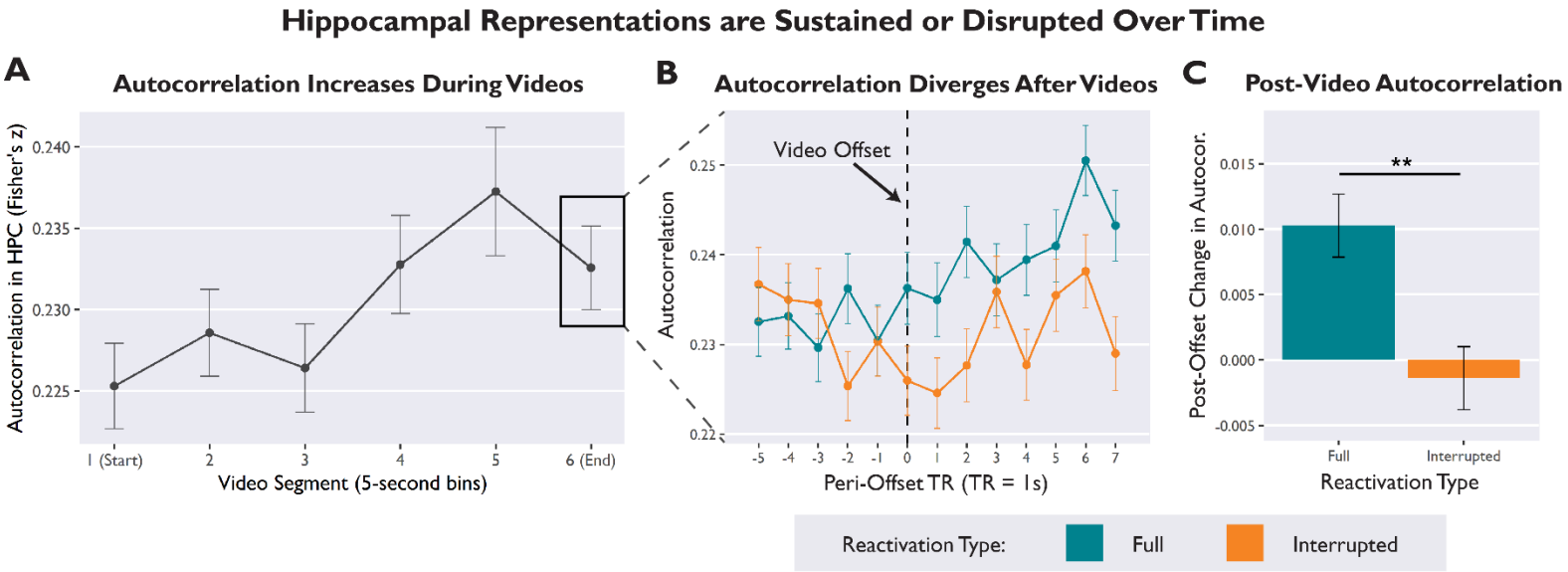

Figure 4. Hippocampal representations are sustained or disrupted over time, depending on whether or not episodes align with expectations. A) Temporal autocorrelation in the hippocampus gradually increased over the course of a video, suggesting that episodic representations were sustained over time. Autocorrelation values were averaged over 5-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.

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*, Post-Video *change in autocorrelation*, *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 [0.01, 0.06], t = 3.09, p = .002) (Figure 5A). After Interrupted videos, disrupting hippocampal representations led to memory updating (β = -0.06, z = -2.65, p = .008). Conversely, after Full videos, hippocampal autocorrelation was not related to memory (β = 0.02, z = 0.88, p = .377). There were no significant interactions with hemisphere or axis. All other parameters are reported in Table S11A.

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 Post-Video univariate basal forebrain activation and interaction terms. The model included all relevant interaction terms, reported in full in Table S11B. 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 (β = 0.03, 95% CI [0.01, 0.05], t = 2.74, p = .006) (Figure 5B). In other words, prediction errors disrupted hippocampal representations and led to memory updating, but only when the basal forebrain was also strongly activated (at +1SD basal forebrain activation, Full: β = 0.08, z = 3.24, p = .001; Interrupted: β = -0.06, z = -2.20, p = .028) (Figure 5B, 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, the three-way interaction among VTA activation, reactivation type, and hippocampal autocorrelation was not significant (β = -0.003, t = -0.26, p = .794, 95% CI [-0.03, 0.02]). All parameter estimates are reported in Table S11C.

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.

Autocorrelation Control Analyses

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 ($t(23) = -0.05$, $p < .96$, $d = 0.01$). Autocorrelation in white matter did not change post-offset ($t(23) = 0.99$, $p = .329$, $d = 0.14$) and did not differ by reactivation type ($t(23) = 0.86$, $p = .40$, $d = 0.18$). In summary, these control analyses indicated that our autocorrelation findings were not a brain-wide phenomenon.

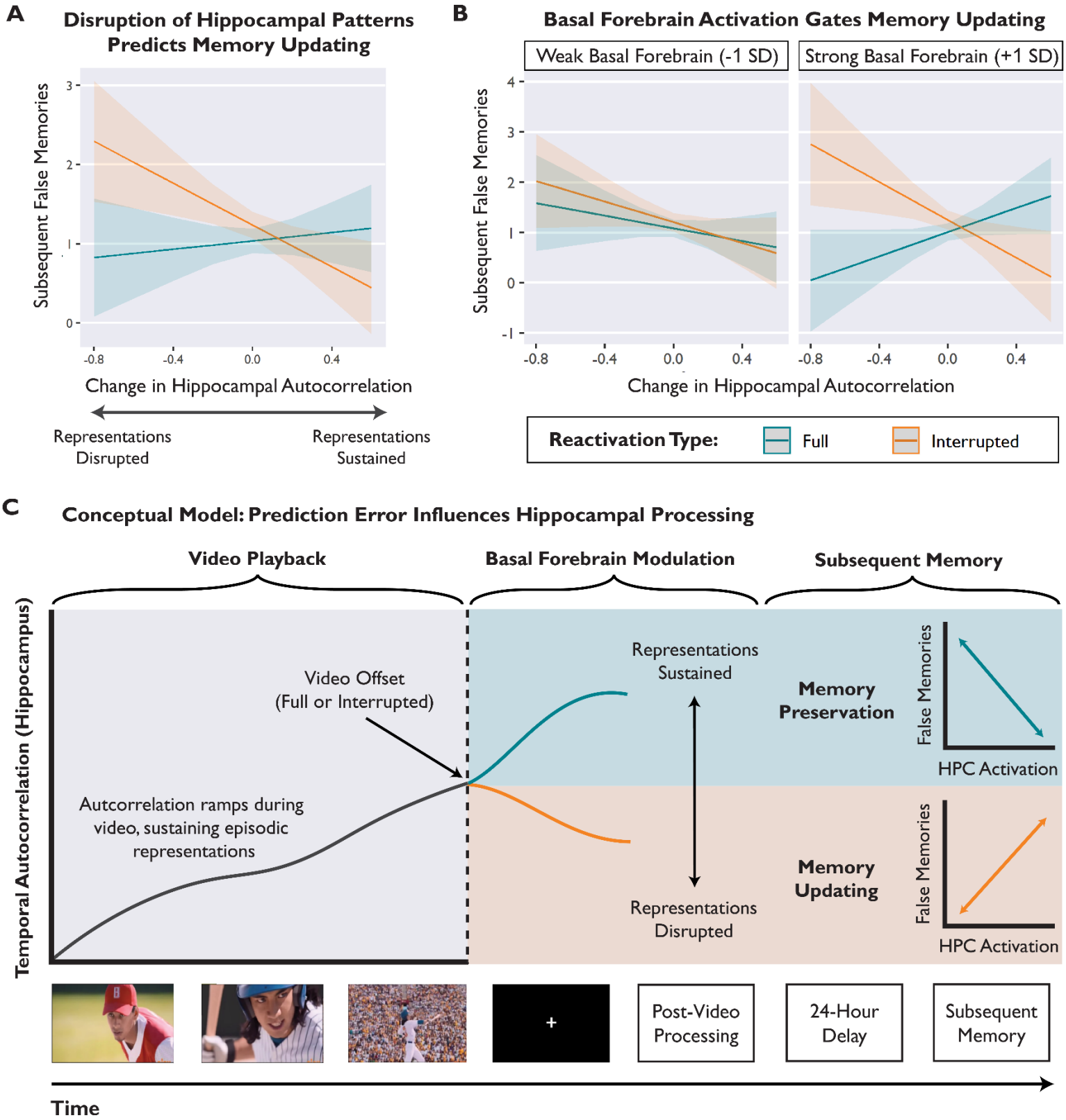
A
Disruption of Hippocampal Patterns Predicts Memory Updating
Subsequent False Memories
Change in Hippocampal Autocorrelation
Representations Disrupted
Representations Sustained
B
Basal Forebrain Activation Gates Memory Updating
Weak Basal Forebrain (-1 SD)
Strong Basal Forebrain (+1 SD)
Subsequent False Memories
Change in Hippocampal Autocorrelation
Reactivation Type:
Full
Interrupted
C
Conceptual Model: Prediction Error Influences Hippocampal Processing
Video Playback
Basal Forebrain Modulation
Subsequent Memory
Temporal Autocorrelation (Hippocampus)
Video Offset (Full or Interrupted)
Autocorrelation ramps during video, sustaining episodic representations
Representations Sustained
Representations Disrupted
Memory Preservation
Memory Updating
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-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, 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. The link between prediction error and memory outcomes depended 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 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 Post-Video 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 (Figure 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, 67). 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 shift in hippocampal processing (Figure 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 (59, 60). 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).

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 (57, 64, 66). As narratives progressed, autocorrelation increased, reflecting stability and continuity; this increase in autocorrelation suggested that the hippocampus generated predictions (63, 64) and sustained episodic representations over time (59, 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 representations indicates a shift in processing: Prediction error prompts the hippocampus to abandon ongoing predictions and prepare to update a memory (Figure 5C).

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, 67). 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 left hippocampus that was *less* activated after Interrupted videos than Full videos (Supplementary Information, *Whole-Brain Analysis*). Crucially, 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 is consistent with prior studies of reward (13) and information prediction errors (4, 69), 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, 41–43, 52) or dopaminergic (16, 24, 53) modulation could regulate hippocampal processing after prediction errors. However, mixed evidence supporting both hypotheses has left the question unresolved (25, 26, 31, 37, 40).

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 co-activation of the hippocampus and basal forebrain, suggesting that connectivity between these regions is important for shifting hippocampal processing. 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, 41–43, 45) gates memory updating and influences hippocampal processing.

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 Post-Video period, VTA activation was unrelated to prediction error and did not link hippocampal activation with memory updating. These findings are consistent with our prior proposal that connectivity between the VTA and hippocampus reflects modulation of hippocampal learning states by sustained VTA activity (16, 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 (74, 75). 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

Our behavioral results demonstrated a novel dissociation: prediction error both strengthened and updated memories, but over distinct timecourses and likely via different mechanisms (Figure 2). Prediction error increased the number of correct details recalled, both immediately and after a one-day delay. This finding is consistent with recent evidence that mnemonic prediction errors can promote detailed memories immediately (76), possibly by way of enhanced attention and pattern separation. In contrast, we found that prediction error increased false memories only after a delay, supporting that idea that reconsolidation specifically enables memory updating or distortion (1, 54, 77). 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 (39). Here, we showed that videos that shared greater semantic similarity with the rest of the stimulus set produced more false memories (Table S4). This finding suggests that prediction error increases source confusion or integration of information across related memories. Reconsolidation studies have shown that memory updating occurs when interference is introduced after a memory trace is reactivated (1, 54, 78). 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 reconsolidation

research (1, 54, 78) and computational models of event segmentation (79, 80), which have both shown that interference among related episodes can produce false memories and/or source confusion 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, 39, 54, 78), but it remains unknown whether the synaptic mechanisms of reconsolidation are consistent across animals and humans (1). We found that the behavioral effect of prediction error on memory updating required a delay, consistent with reconsolidation processes that rely on protein synthesis. We are 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. 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 track rapid changes in hippocampal representations during and after naturalistic episodes.

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. One possibility is that the strength of

surprise diminished over the course of the experiment; the first interrupted video was likely the most surprising. To address this limitation, we designed our experiment to maintain surprise as much as possible: Full-length videos were presented twice during encoding to set strong expectations; Interrupted videos violated strongly-expected action-outcome contingencies (e.g., a baseball batter halted mid-swing); we pseudorandomized our trials to ensure that participants could not anticipate whether the next video would be Full or Interrupted; and participants did not know *when* each video would 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 meaningfully diminish over the course of the experiment (Table S2).

Animal and human reconsolidation studies have previously used *incomplete reminders* (e.g., a conditioned stimulus without the expected outcome) to elicit prediction error (1, 3, 78). 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 (81–83). 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 be important for memory updating. These mechanisms could interact. Memory reactivation may be a prerequisite for generating a prediction, and experiencing a prediction error may prompt further memory reactivation. Future studies could investigate the role of memory reactivation, such as 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 red 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 Reconsolidation group, only 2% of all false memories specifically 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 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 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. We thank Jia-Hou Poh for helpful comments on the manuscript.

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 Table S12. 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 paid participants from the University of Toronto community (Reconsolidation group: $70, Immediate control group: $40). Seven participants were excluded (see *Exclusions*), yielding a final sample of 48 participants. The sample size was determined *a priori* to satisfy the following conditions: (1) reproduce the sample size and achieve at least 90% power to detect the interaction effect found in prior work ($\eta_p^2 = 0.17$) (39), and (2) evenly allocate participants to 6 pseudorandomized trial order lists. Participants were healthy young adults (age: $M = 22.42$, $SD = 2.41$, range [18, 30]; gender: 75% female, 25% male). 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. 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 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 scenes or characters. 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 Table S12 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). Importantly, the duration of the videos varied considerably, yielding overlapping distributions of Full and Interrupted videos (Figure S3). This variability ensured that our neural results could not be attributed to systematic timing differences between the Full and Interrupted videos (e.g., neural adaptation after given durations of visual stimulation).

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. thus preventing participants from anticipating

which videos would be interrupted. 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 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 (84–86), 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. The mock scanner room was adjacent to the real scanner room, and was very similar in size and appearance.

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.

Statistical Analysis

For both behavioral and neural data, we conducted trial-wise analyses with linear mixed-effects regression models. Detailed information about the random effects specification is provided in the Supplementary Information (*Supplementary Methods, Linear Mixed-Effects Regression*).

Univariate fMRI Analyses

Whole-brain mass univariate results are reported in the Supplementary Information (*Whole-Brain Analysis*, Figure S2, Table S5). Details about preprocessing steps and region-of-interest masks are provided in the Supplementary Information (*Supplementary Methods, fMRI Preprocessing, Region of Interest Masks*). 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 (87, 88). 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 (59, 60). 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 to yield an average activation value.

Multivariate fMRI Analyses

Multivariate temporal autocorrelation analyses (57, 64) were conducted on preprocessed data (prior to single-trial GLM analysis). 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. 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. Comparable to past research, we phase-shifted the timeseries by 4 seconds in order to account for HRF lag (89). This manual shifting is necessary because event onset regressors have not yet been convolved with the HRF (unlike in standard GLM analyses used for our univariate analyses). Note that our univariate analyses involved an additional 2s shift in addition to the standard HRF shift; this allowed us to target the peak post-video 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-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 analyze Post-Video changes in autocorrelation, we calculated trial-by-trial difference scores by subtracting the average autocorrelation value for the 5-second bin immediately *before* video offset from the average autocorrelation value for the 5-second bin immediately *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. 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. Hasselmo, B. Wyble, G. Wallenstein, Encoding and Retrieval of Episodic Memories: Role of Cholinergic and GABAergic Modulation in the Hippocampus. *Hippocampus* **6** (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.* **8** (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. J. Debiec, J. E. LeDoux, K. Nader, Cellular and systems reconsolidation in the hippocampus. *Neuron* **36**, 527–538 (2002).

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

35. 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).

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

37. 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).

38. 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).

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

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

41. 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).

42. 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).

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

44. 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).

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

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

47. 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).

48. 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).

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

50. 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).

51. 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).

52. 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).

53. J. I. Rossato, *et al.*, State-dependent effect of dopamine D1/D5 receptors inactivation on memory destabilization and reconsolidation. *Behav. Brain Res.* **285**, 194–199 (2015).

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

56. 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).

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

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

59. 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).

60. 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).

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

62. M. Zubair, *et al.*, Divergent Whole Brain Projections from the Ventral Midbrain in Macaques. *Cereb. Cortex N. Y. N 1991* **31**, 2913–2931 (2021).

63. 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).

64. 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).

65. A. Tambini, L. Davachi, Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19591–6 (2013).

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

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

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

69. E. S. Bromberg-Martin, O. Hikosaka, Lateral habenula neurons signal errors in the prediction of reward information. *Nat. Neurosci.* **14**, 1209–1218 (2011).

70. 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).

71. 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).

72. 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).

73. 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).

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

75. 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).

76. O. Bein, N. A. Plotkin, L. Davachi, Mnemonic prediction errors promote detailed memories (2021) https:/doi.org/10.31234/osf.io/b5z8a (May 11, 2021).

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

78. 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).

79. 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).

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

81. 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).

82. 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.

83. 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.

84. 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).

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

86. 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).

87. 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).

88. 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.

89. 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.