Prediction errors during naturalistic events modulate hippocampal representations and drive episodic memory updating

*Alyssa H. Sinclair[1,2] Grace M. Manalili[2,3], Iva K. Brunec[4,5],
R. Alison Adcock[1], & Morgan D. Barense[2,6]

[1] Duke University, [2] University of Toronto, [3] Queen's University, [4] Temple University,
[5] University of Pennsylvania, [6] Rotman Research Institute (Baycrest Hospital)

*Corresponding Author: Alyssa H. Sinclair, allie.sinclair@duke.edu

Main Text Word Count: 4,495

Abstract Word Count: 150

Abstract

The brain supports adaptive behavior by generating predictions, learning from errors, and updating memories. *Prediction error,* or surprise, is a known trigger for memory updating; however, the mechanisms that link prediction error, neural representations, and naturalistic memory updating remain unknown. In an fMRI study, we elicited prediction errors by interrupting familiar narrative videos immediately before an expected conclusion. We found that prediction errors reversed the effect of post-video univariate hippocampal activation on subsequent memory: hippocampal activation predicted false memories after prediction errors, but protected memories from distortion after expected events. Tracking second-by-second neural patterns revealed that prediction errors disrupted the temporal continuity of hippocampal representations. This disruption of signal history led to memory updating after prediction error. We conclude that prediction errors during memory reactivation prompt the hippocampus to abandon ongoing predictions and neural representations. Following prediction error, the hippocampus switches to an externally-oriented processing mode that supports memory updating.

Introduction

In daily life, we continuously draw on past experiences to predict the future. Expectation and surprise shape learning across many situations, such as when we encounter misinformation in the news, receive feedback on an exam, or make decisions based on past outcomes. When our predictions are incorrect, we must update our internal models of the world to support adaptive behavior. Converging evidence supports the idea that the brain signals *prediction error*, a measure of the discrepancy between expectation and reality; this surprise signal is related to memory updating[1–5]. However, it remains unknown how prediction error transforms neural representations of ongoing experiences, and thus allows memories to be updated with new information.

Prior research has implicated the hippocampus in the mnemonic functions required for learning from prediction errors: retrieving memories to make predictions, identifying discrepancies between past and present, and encoding potentially relevant new information[2,6–11]. Functional MRI (fMRI) studies have revealed that hippocampal activation increases after predictions are violated; this surprise response has been previously called *mismatch detection*[9,10,12,13], or *mnemonic prediction error*[11]. Evidence from human and animal research has implicated subfield CA1 as the origin of this surprise signal[8,9,11,12]. Recent work has shown that mnemonic prediction errors enhance connectivity between CA1 and the entorhinal cortex (a pathway that supports processing perceptual inputs)[14], but suppress connectivity between CA1 and CA3 (a pathway that supports memory retrieval)[11]. These findings support the idea that CA1 signals prediction errors and biases the hippocampus towards encoding incoming perceptual inputs, thus enabling memory updating. In other words, surprise may trigger a switch from an internally-oriented processing mode (which supports memory retrieval and ongoing predictions)

to an externally-oriented processing mode (which supports memory updating and new encoding)[11,15–17]. Despite these implied mechanisms, it remains unknown whether such mode switching affects hippocampal representations, and what effect mode switching has on subsequent memory.

Importantly, the memory malleability induced by surprise can be beneficial or harmful: prediction error may allow memories to be adaptively updated, but the same process can also produce distortion and false memories[1]. Animal and human research on *reconsolidation*[18–20], the process by which memory traces can be reactivated and temporarily destabilized, has shown that surprising reminders allow memories to be updated with new information[1,3]. In order to elicit prediction error, reconsolidation paradigms have used incomplete reminders that generate surprise by replicating part, but not all, of an encoding experience (e.g., a conditioned stimulus without the expected outcome)[1,3,21]. In previous behavioral research, we found that prediction error drives updating of naturalistic episodic memories in humans[22]. Using narrative videos, we showed that surprising reminders allowed memories to be updated with new information, thus producing false memories. These findings from the reconsolidation literature parallel past research on mnemonic prediction error and hippocampal mode switching, but these theoretical frameworks have not been bridged.

Present Study

Here, we provide the first demonstration of the neural mechanisms that enable updating of naturalistic episodic memories in humans. We sought to explain how surprise signals modulate hippocampal representations and determine the fate of complex episodic memories. We hypothesized that eliciting prediction errors during memory reactivation would alter neural representations and prompt the hippocampus to switch to an externally-oriented processing mode

that enables memory updating. Using fMRI, we showed that prediction errors reversed the relationship between hippocampal activation and subsequent memory, thus driving memory updating. Furthermore, we uncovered a new marker of mode switching by tracking dynamic changes in hippocampal representations during and after narrative videos. We show that the hippocampus builds increasingly stable representations during and after naturalistic events, that the temporal continuity of hippocampal representations is disrupted by prediction errors, and finally that these representational disruptions predict memory updating.

Using fMRI, we examined trial-wise hippocampal responses to prediction errors during narrative videos. During the Day 1 encoding session, participants viewed 70 audio-visual videos that featured narrative action-outcome events (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 prediction error, here defined as narrative surprise, by interrupting these videos immediately before the expected outcome (e.g., the video ends while the baseball batter is mid-swing). These surprising interruptions are comparable to the incomplete reminders that have been used to elicit prediction error 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). Participants in the Reconsolidation group (n=24) completed the Day 2 session while undergoing an fMRI scan, whereas participants in the Immediate control group (n=24) completed the study in a behavioral testing room and were not scanned. Our fMRI analyses focused on comparing neural responses to Full and Interrupted videos during the Day 2 session in the Reconsolidation group.

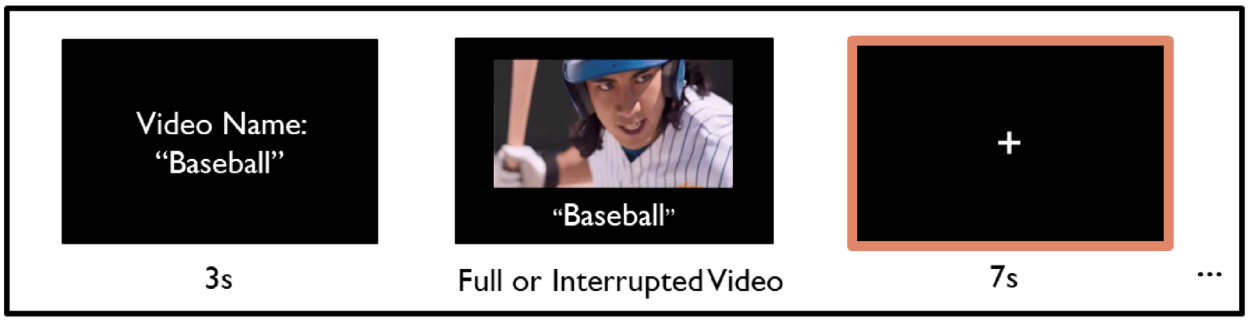
Lastly, participants completed a memory test in the form of a structured interview (Figure 1C). On each trial, participants were cued with the name of the video and freely recalled the

narrative. The experimenter then probed for further details with pre-determined questions (e.g., "Can you describe the baseball batter's ethnicity, age range, or clothing?"). Our critical measure of memory updating was *false memories*. Although real-world memories may be adaptively updated with relevant new information, we expected that our laboratory paradigm would induce false memories that integrated interfering details across many events. 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 procedure on Day 2, immediately after the reactivation session (Figure 1D). Reconsolidation theory states that memory updating requires a delay because re-stabilizing a memory involves protein synthesis that occurs over several hours[18,23]. Thus, in the Immediate control group, testing immediately after reactivation should limit any effects of prediction error on memory updating that require protein synthesis-dependent reconsolidation.

A

Example Stimulus Video: "Baseball"

B



C

Example Memory Test (Transcribed & Scored)

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

Legend: Correct Details False Memories

D

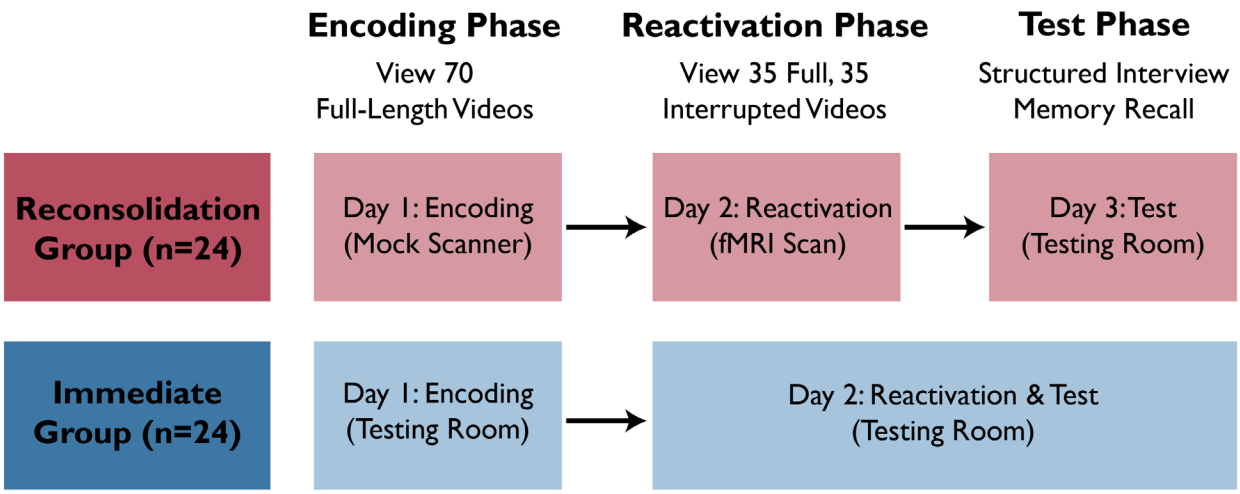


Figure 1. Overview of experimental paradigm. A) Frames from a stimulus video named "Baseball", depicting a home run. To elicit prediction error (surprise), this video was interrupted while the batter was mid-swing during the Reactivation Phase (B). At Reactivation, participants were cued with the video name, watched the video (Full or Interrupted), then viewed a fixation screen. fMRI analyses focused on the post-offset period after each video (highlighted box), treating the offset of the video as the stimulus. C) Example text illustrating the memory test format and scoring. Participants answered structured interview questions about all 70 videos, and were instructed to answer based on their memory of the full-length video originally shown during encoding. 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, while the Immediate control group did the Test Phase immediately after Reactivation, in order to investigate whether memory updating required a delay. Only the Reconsolidation group was scanned.

Results

Behavioral Results

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

Correct details. First, we tested whether prediction error during memory reactivation influenced the number of correct details recalled. Reactivation type was significantly related to correct details, $F_{(1,69)}$=7.59, p=.007, 95% CI [-0.67, -0.11], such that participants recalled *more*

correct details for Interrupted videos than Full videos (Figure 2A). Even though the video endings were omitted, prediction error strengthened and preserved existing memories. There was also a main effect of group, such that participants in the Reconsolidation group recalled fewer correct details than participants in the Immediate group, $F_{(1,46)}$=4.69, p=.036, 95% CI: [0.09, 1.77], likely because the Reconsolidation group completed the memory test after a 24-hour delay. Despite reporting fewer correct details, Reconsolidation group participants still demonstrated rich memory recall: The average number of correct details per-video (M=12.61, SD=5.84) was only 1.8 points below the average for the Immediate group (M=14.4, SD=5.54). Importantly, there was no interaction between group and reactivation type, $F_{(1,248)}$=0.48, p=.488, 95% CI [-0.24, 0.11], indicating that the effect of prediction error enhancing correct details did not require a delay.

False memories. We found that prediction error selectively increased false memories after a 24-hour delay in the Reconsolidation group, replicating our past behavioral results[22] (significant interaction between reactivation type and group, $F_{(1,270)}$=8.98, p=.003, 95% CI [0.02, 0.08], Figure 2B). In other words, Interrupted videos increased false memories in the Reconsolidation group (t(23)=-4.84, p < .001), but not the Immediate group (t(23)=-0.88, p=.387). We also found main effects of group, $F_{(1,46)}$=94.65, p < .0001, 95% CI [-0.43, -0.29], and reactivation type, $F_{(1,155)}$=12.30, p < .001, 95% CI [-0.09, -0.03], both driven by the effect of prediction error increasing false memories in the Reconsolidation group.

In sum, our results for correct details and false memories show a novel dissociation between memory strengthening and memory updating: Prediction error during memory reactivation can strengthen a memory immediately, but updating a memory trace with new information requires a delay, as predicted by reconsolidation theory.

Item analysis: Surprise ratings and semantic similarity. Expanding on the results reported above, we recruited an independent sample to watch the videos and rate (on a 5-point Likert scale) the degree of surprise elicited by the narrative interruptions (Methods, Online Video Ratings). We found that surprise ratings were unrelated to correct details (Supplementary Table 1), but there was a significant interaction between surprise ratings and group such that more surprising videos were associated with more false memories selectively in the Reconsolidation group, $F_{(1,2994)}$=5.72, p=.017, 95% CI [-0.07, -0.01].

Here, we use false memories as an index of memory updating; however, incorporating relevant new information into memory can be an adaptive function. In previous research, we presented semantically-related interference videos (e.g., a new video about baseball)[22] after memory reactivation, and found that prediction errors increased false memories that were intrusions from the semantically-related videos. Time constraints prevented the use of interference videos in the present study, so we quantified semantic similarity among the 70 videos with a text-based analysis (Methods, Memory Tests) (Supplementary Figure 2). We found that videos that were more semantically similar to other videos in the stimulus set yielded more false memories, $F_{(1,68)}$=6.40, p=.014, 95% CI [0.02, 0.17] (Supplementary Table 2). Interestingly, semantic similarity did not predict correct details overall, $F_{(1,67)}$=0.09, p=.769, 95% CI [-0.28, 0.21], but instead showed a significant interaction with reactivation type, $F_{(1,68)}$=8.22, p=.006, 95% CI [0.12, 0.64]. For Full videos, semantic similarity was positively associated with correct details (consistent with schema-based memory benefits). For Interrupted videos, semantic similarity was negatively related to correct details, suggesting a trade-off with false memories. This pattern of semantic similarity predicting fewer correct details but more false memories

indicates that memories may be updated with relevant new information, exactly as required for adaptive behavior.

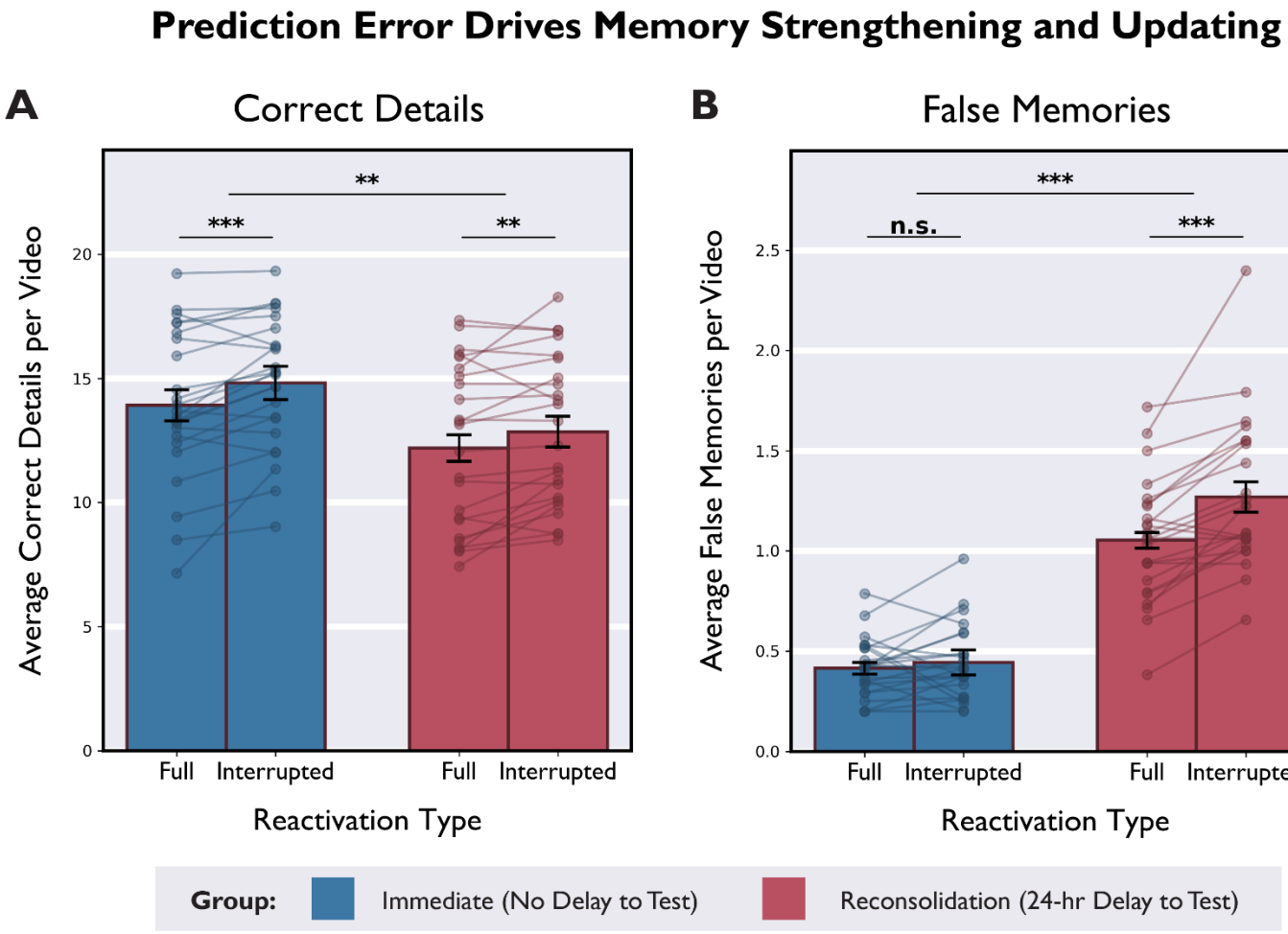


Figure 2. Prediction error modulates correct details versus false memories over distinct timecourses. A) In both groups, average Correct Details were higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error can strengthen memory recall both immediately and after a delay. B) Only in the Reconsolidation group were average False Memories higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error drives memory updating but requires a delay to permit the protein synthesis that underpins reconsolidation. Dots indicate average scores by-participant, and lines connect within-subjects measures. Error bars depict 95% confidence intervals. * $p < .05$, ** $p < .01$, *** $p < .001$.

Univariate fMRI Results

The primary aim of the fMRI analyses was to test whether blood oxygen-dependent (BOLD) activation in the hippocampus differed depending on *reactivation type* (Full vs. Interrupted) and subsequent *false memories* in the Reconsolidation group (the Immediate group was not scanned). Our analyses focused on the period immediately after the offset of each video. We treated the moment of video offset as the stimulus: the narrative conclusion was either as-expected (Full) or a surprising prediction error (Interrupted). After each video, participants viewed a fixation screen, thus controlling for visual and auditory input. Whole-brain mass univariate results are provided in the Supplemental Material (Whole-Brain Analysis, Supplementary Table 3, Supplementary Figure 3).

To address our primary research questions, we used single-trial modelling to relate hippocampal activation to subsequent false memories. For our univariate analyses, we modelled a 2s post-offset impulse after each video, convolved with the canonical double-gamma hemodynamic response function and phase-shifted 2s after video offset. This 2s shift targets the peak hippocampal offset response previously identified in studies of video processing[24,25]. We isolated fMRI activation during the post-offset period on each trial and averaged parameter estimates across all hippocampal voxels (Methods, fMRI Data Analysis).

Some past studies have shown that prediction error signals are stronger in left hippocampus and anterior hippocampus (which contains area CA1)[9,11,12,26], whereas posterior hippocampus is more sensitive to video offsets[27]. Relatedly, other studies have shown that anterior and posterior hippocampus parse continuous experience at different timescales[28,29]. On the basis of these findings, we tested separate ROIs for left, right, anterior, and posterior hippocampus (Methods, ROI Masks), but found that our effects were generally very consistent

across hippocampal ROIs (Supplemental Material, ROI Differences). Main text results are averaged across bilateral hippocampus, but results from individual ROIs are provided in Supplementary Tables 4-7 and Supplementary Figures 4-7. We used linear mixed-effects regression to predict trial-wise activation estimates from the variables *reactivation type* (Full vs. Interrupted), *false memories* (continuous measure), and their interaction.

We found a significant interaction between reactivation type and subsequent false memories predicting hippocampal activation, $F_{(1,1047)}$=7.04, p=.008, 95% CI [-0.11, -0.02] (Figure 3) (Table 2A). After Full videos, greater hippocampal activation was associated with fewer subsequent false memories (Figure 3, blue). This protective effect is consistent with the idea that post-offset hippocampal activation binds episodic details and reinforces memory for the event that just concluded[24,25,30]. After an event that aligns with expectations, the hippocampus should remain in an internal processing mode that favors ongoing predictions and memory retrieval[31]. However, when the ending of the video was surprising, we observed exactly the opposite effect. After Interrupted videos, greater hippocampal activation was associated with *more* false memories, consistent with the idea that surprise drives memory updating by triggering a switch to an external processing mode (Figure 3, orange).

Overall, we propose that prediction error changes the function of the hippocampus after an event, and thus determines the fate of episodic memories. During memory reactivation, the hippocampus retrieves a past episode, generates predictions, and checks for a mnemonic prediction error. If no prediction error is detected, then the hippocampus remains in an internal processing mode, reinforcing details from within the episode and protecting the memory from subsequent interference. In contrast, surprising events that violate expectations trigger the hippocampus to abandon ongoing predictions and switch to an external processing mode that

supports memory updating. Here, we show that univariate measures are insufficient for understanding the effect of prediction error on the hippocampus, because activation can exert opposing effects on memory. This novel finding now opens the question of whether and how prediction error impacts hippocampal *representations*, reflecting this change in processing mode.

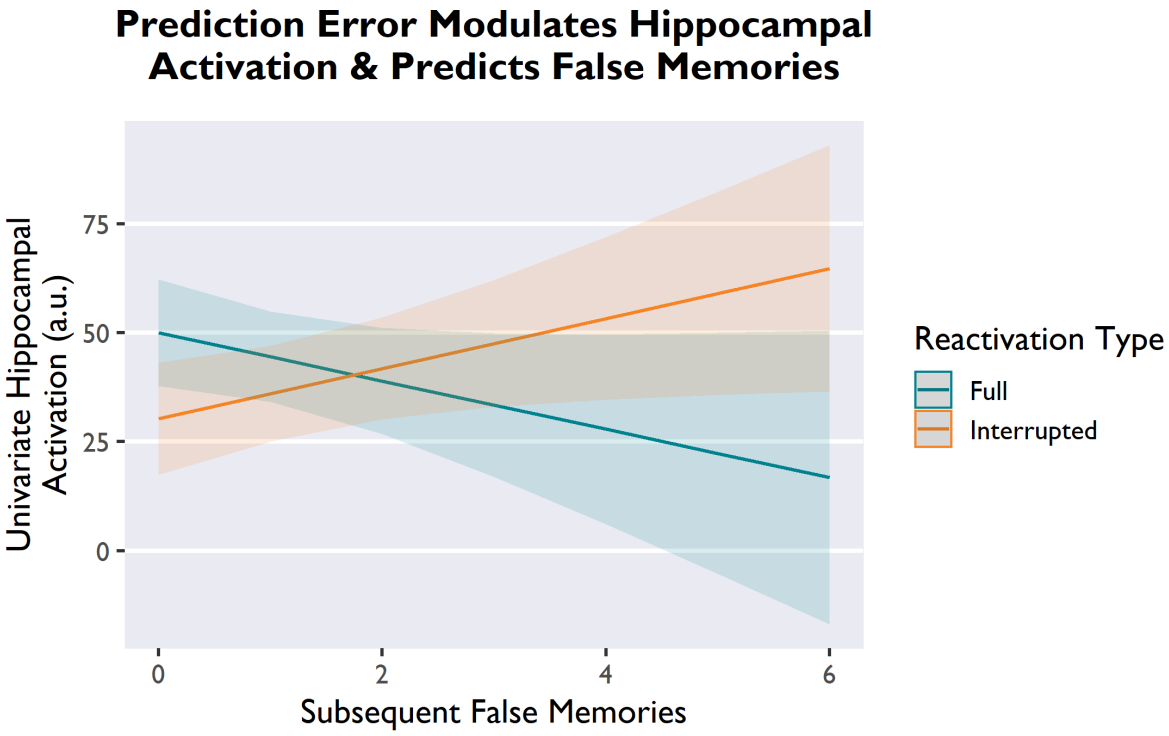


Figure 3. Prediction error reverses the relationship between hippocampal activation and subsequent memory. After Full-length videos, greater hippocampal activation had a protective effect, predicting fewer false memories (blue). After Interrupted videos, greater hippocampal activation predicted more false memories (orange). Lines depict model-predicted estimates, and shaded bands depict the 95% confidence interval.

Signal History in the Hippocampus

In our univariate analysis, we found that hippocampal activation produced different memory effects (preserving vs. updating) depending on whether an event was surprising. We hypothesized that these memory effects reflect switching processing modes[15]; if prediction error causes the hippocampus to switch modes, then the temporal continuity of hippocampal representations should be disrupted. Furthermore, this neural measure of mode switching should predict subsequent memory updating.

In order to test this hypothesis, we tracked *signal history* in the hippocampus by using an event-related *temporal autocorrelation* analysis that measures how multivariate fMRI patterns change dynamically over time[28,32,33]. Previously, an intracranial electrophysiological study found that firing rates of human hippocampal neurons became increasingly correlated as participants watched videos over multiple exposures[32]. Ramping autocorrelation during events may thus reflect ongoing predictions based on past experiences; the hippocampus anticipates upcoming stimuli and gradually stabilizes event representations[32]. Other studies in rodents and humans have used temporal and spatial autocorrelation to investigate representational scales along the hippocampal long axis, demonstrating a spectrum of coarse (anterior regions) to fine (posterior regions) representations during naturalistic tasks[28,29].

In the present study, our short TR (1s) enabled us to track second-by-second signal history in the hippocampus during and after events. Temporal autocorrelation measures moment-to-moment pattern overlap, or the amount of multivariate information that is preserved over time (i.e., signal history). High autocorrelation values indicate stable neural representations, whereas low autocorrelation values indicate that neural representations are changing dynamically. This multivariate analysis involves correlating the activation of all voxels within the hippocampus at

timepoint *T* with the activation pattern at timepoint *T+1 sec* (Methods, Autocorrelation Analyses).

First, we used linear mixed effects regression to test whether hippocampal autocorrelation increased over the course of an event. As predicted, we found that hippocampal autocorrelation linearly increased as videos progressed, $F_{(1,14893)}$=8.45, *p*=.004, 95% CI [0.001, 0.003] (Figure 4A, Table 2B). This increase in autocorrelation is consistent with the idea that signal history ramps as the hippocampus generates ongoing predictions and builds stable event representations[32]. Plotting second-by-second autocorrelation values before and after videos revealed that autocorrelation for Full and Interrupted videos diverged at the moment of offset (Table 4B). After Full videos, autocorrelation continued ramping into the post-offset period, whereas after Interrupted videos, the ramping of signal history ceased.

To test whether these post-offset signal history trajectories significantly differed for Full and Interrupted videos, we binned the 5-second periods immediately before and after video offset and calculated autocorrelation change scores driven by video offset. Paired t-tests revealed that the average post-offset increase in hippocampal autocorrelation was significantly greater after Full videos than Interrupted videos, *t*(23)=3.22, *p*=.004, 95% CI [0.004, 0.018], Cohen's *d*=0.66 (Figure 4C). In other words, ramping of signal history was disrupted after Interrupted videos. Thus, prediction errors disrupted the temporal continuity of representations in the hippocampus, consistent with a switch to an external processing mode.

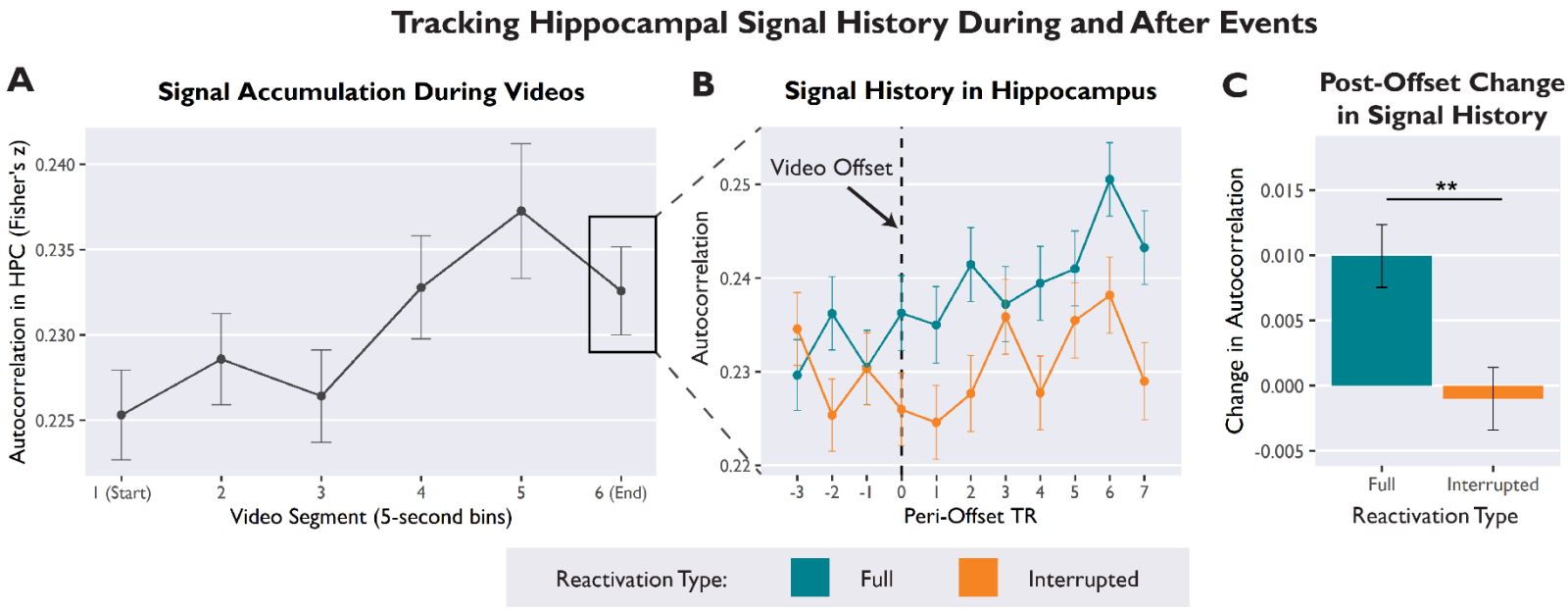


Figure 4. Tracking hippocampal signal history during and after events. A) Signal history (temporal autocorrelation) in the hippocampus gradually ramps over the course of a video. Autocorrelation values are averaged over 5-second bins of video playback. B) Second-by-second autocorrelation values in the hippocampus, time-locked to the moment of video offset (black dotted line). Signal history for Full and Interrupted videos diverges after video offset. C) Comparing post-offset change in signal history, using average autocorrelation scores for the 5-sec bins immediately before and after video offset. Prediction error halts signal accumulation after video offset. Error bars depict SEM.

Next, we tested whether disrupting the temporal continuity of hippocampal representations was related to subsequent false memories. Using linear mixed effects regression, we predicted false memories from the interaction between reactivation type and change in autocorrelation, while controlling for the level of univariate activation. We found a significant interaction between reactivation type and change in autocorrelation predicting subsequent false memories, $F_{(1,1347)}$=4.57, p=.033, 95% CI [0.06, 1.41] (Figure 5A, Table 2C). After Interrupted videos, hippocampal autocorrelation was inversely related to subsequent false memories. In other words, when signal history was reset after prediction error, more memory updating occurred (Figure 5B). Conversely, after Full videos, signal history was unrelated to memory updating.

Lastly, we tested differences in autocorrelation in two control regions, inferior lateral occipital cortex (LOC) and white matter. We predicted that autocorrelation in LOC would be sensitive to 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 offsets or prediction error. As predicted, autocorrelation in LOC significantly increased after videos ($t(23)=9.47$, $p < .001$, 95% CI [0.04, 0.07], Cohen's $d=1.37$), but did not differ by reactivation type ($t(23)=-0.05$, $p < .96$, 95% CI [-0.02, 0.02], Cohen's $d=0.01$). In contrast, autocorrelation in white matter did not change post-offset ($t(23)=0.99$, $p=.329$, 95% CI [-0.002, 0.01], Cohen's $d=0.14$) and did not differ by reactivation type ($t(23)= 0.86$, $p=.40$, 95% CI [-0.01, 0.01], Cohen's $d=0.18$). In summary, these control analyses indicate that prediction error was related to autocorrelation in the hippocampus, but not other brain regions.

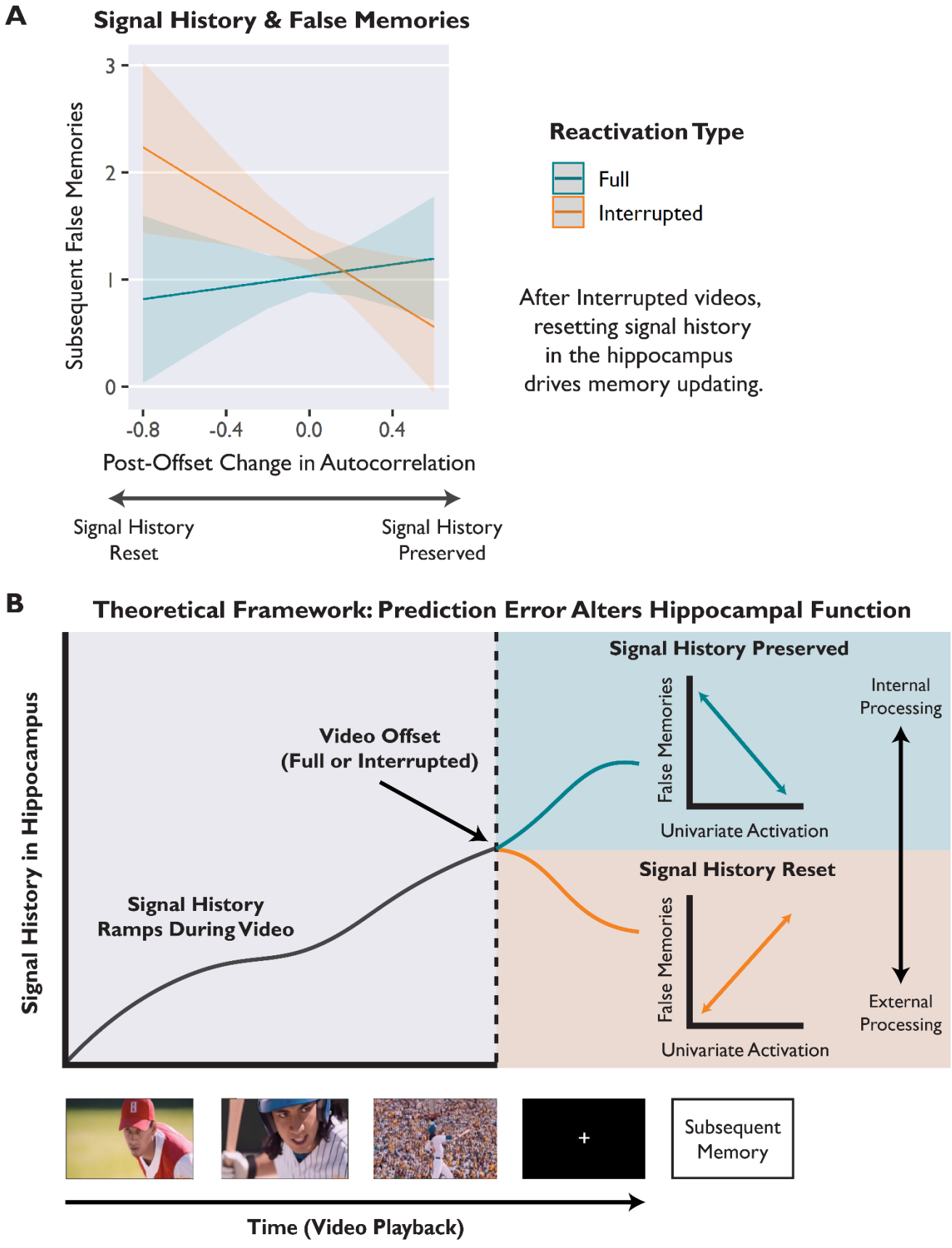


Figure 5. Prediction error resets signal history and drives memory updating. A) Estimated values from a linear regression model predicting subsequent false memories from the interaction of reactivation type and change in autocorrelation (i.e., signal history). After Interrupted videos, decreases in autocorrelation (negative values indicating resetting of signal history) were related to more memory updating. Shaded bands depict 95% confidence intervals around the regression line. B) Schematic depicting hippocampal mode switching and subsequent memory. During a video, signal history ramps in the hippocampus. After video offset, preservation or resetting of signal history indicates whether the hippocampus remains in an internal processing mode where univariate activation reflects ongoing prediction and retrieval (protecting against distortion), or an external processing mode where univariate activation reflects updating and encoding (increasing false memories).

Discussion

We found that narrative surprise, a unique type of mnemonic prediction error, drives memory updating. For the first time, we demonstrate that prediction errors modulate hippocampal representations and allow naturalistic episodic memories to be updated. Using a novel fMRI paradigm, we elicited prediction errors during narrative events by interrupting videos immediately before the expected conclusion. We showed that hippocampal activation exerted dissociable effects on memory depending on whether an event was surprising: After Interrupted videos, hippocampal activation led to memory updating, but after Full videos, hippocampal activation protected memories from distortion. Tracking dynamic changes in activation patterns revealed that signal history in the hippocampus ramped during and after events, but prediction error disrupted the temporal continuity of hippocampal representations. When signal history was reset after prediction errors, more memory updating occurred. To summarize, we show the following: (1) prediction error reverses the effect of hippocampal activation on memory, (2) prediction error disrupts the temporal continuity of neural representations, and (3) these representational disruptions predict subsequent memory updating. We conclude that prediction error biases the hippocampus towards an externally-oriented processing mode that promotes encoding and memory updating (Figure 5B).

Prediction Error Modulates Hippocampal States to Support Memory Updating

Critically, we showed that prediction error reversed the relationship between hippocampal activation and subsequent memory (Figure 3). We propose that these dissociable effects on memory arise from the hippocampus switching between processing modes. After full-length events (no prediction error), greater hippocampal activation predicted fewer false memories; this protective effect is consistent with an internal processing mode that supports ongoing predictions and binding episodic details. After surprising interruptions (prediction error), greater hippocampal activation predicted *more* false memories; this memory updating effect is consistent with an external processing mode that supports encoding new information[11,15,16]. To test the idea that prediction error triggers mode switching, we tracked signal history in the hippocampus by analyzing temporal autocorrelation[28,32,33], measured as the similarity between multivariate patterns of fMRI activation on a second-by-second basis. We found that hippocampal autocorrelation ramped up as events progressed, suggesting that the hippocampus generates predictions[32] and stabilizes event representations over time (Figure 4A).

For the first time, we show that prediction error disrupts hippocampal signal history and thus drives memory updating. After Full videos, autocorrelation continued to ramp during the post-offset period, suggesting that the hippocampus maintains event representations and binds episodic details[24,34]. However, ramping ceased after Interrupted videos, indicating that prediction error disrupts signal history and alters hippocampal states (Figure 4B, 4C). Critically, we also related this effect to subsequent memory (Figure 5A). After Interrupted videos, resetting of signal history (decrease in autocorrelation) predicted more false memories. Conversely, after Full videos, signal history was unrelated to memory. Overall, we show that prediction error disrupts the temporal continuity of hippocampal representations, and that this disruption leads to memory

updating. We propose that disruptions in signal history reflect switching from an internal processing mode that supports prediction and retrieval to an external processing mode that supports memory updating and encoding (Figure 5B). Our findings substantially advance past research on mnemonic prediction error[8,9,35] by disambiguating the effects of univariate hippocampal activation on memory, identifying a novel marker of mode switching, and linking mode switching to episodic memory updating.

Prediction Error Drives Memory Updating and Strengthening

Our behavioral results demonstrate a novel dissociation between memory strengthening and memory updating after prediction error (Figure 2). Prediction error increased the number of correct details recalled, both immediately and after a one-day delay. In contrast, the effect of prediction error on false memories required a delay, supporting that idea that reconsolidation enables memory updating[1,23,36]. The finding that prediction error also increased correct details demonstrates that the false memories arise from an adaptive updating mechanism, not forgetting. Overall, our results demonstrate that surprise can both enhance and update episodic memories, but these processes follow distinct timecourses.

Here, we use false memories as an index of memory updating; in the real world, it is adaptive to update memories with relevant new information. In our paradigm, interference from other stimulus videos likely produces false memories because information is integrated across videos. Previously, we found that prediction errors selectively updated memories with semantically-related information from interference videos[22]. Here, we showed that videos that had greater semantic overlap with the rest of the stimulus set produced more false memories (Supplementary Table 2). This finding accords with reconsolidation research[1,23,37] and computational models of event segmentation[38,39], which have both shown that interference

among related events can produce false memories after prediction error. However, memory updating is beneficial in other situations that require integrating old and new knowledge, or correcting erroneous information. Overall, we propose that prediction error supports adaptive memory updating by allowing relevant information to be incorporated into a memory trace.

Limitations

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[4,22,23,37], but it remains unknown whether the synaptic mechanisms of reconsolidation are consistent across animals and humans[1,40]. Here, we show that the effect of prediction error on memory updating requires a delay, consistent with reconsolidation processes that rely on protein synthesis. In past studies, we also showed that memory updating critically depends on a brief delay between memory reactivation and exposure to interference, consistent with the timecourse of protein degradation that underlies memory destabilization[22]. Overall, our findings are broadly relevant to research on prediction error and memory malleability even though the synaptic mechanisms remain unknown; reconsolidation theory offers one plausible framework for our results.

Additionally, the present data lack the spatial resolution required to segment hippocampal subfields. In the present study, we prioritized temporal resolution over spatial resolution, in order to more accurately track rapid changes in hippocampal representations during and after naturalistic events. As a result, we were unable to segment hippocampal subfields. Recent research has investigated mode switching in terms of connectivity between hippocampal subfields[11], but we are unable to analyze information flow between these small regions.

Future Directions

In our paradigm, we draw inspiration from reconsolidation research; we elicited prediction error by interrupting action-outcome events in narrative videos, comparable to the incomplete reminders (e.g., a conditioned stimulus without the expected outcome) that have been previously used in animal and human reconsolidation studies[1,3,37]. However, it remains unknown whether a reminder must be *incomplete* in order to initiate reconsolidation, or whether other surprising or novel stimuli (e.g., sounds, alternate endings) may also induce memory malleability. However, incomplete reminders may be particularly effective because participants actively recall missing associates. Past research has shown that *memory reactivation*, measured in terms of neural reinstatement, supports plasticity[41–43]. Future research could directly investigate memory reactivation by testing encoding-retrieval pattern similarity after an incomplete reminder.

Lastly, the underlying mechanisms of hippocampal mode switching remain largely unknown. Past research suggests that prediction error drives neuromodulatory systems that could induce mode switching: leading hypotheses focus on the role of acetylcholine (ACh)[15,16,44] and dopamine (DA)[7,45]. Several models have proposed that following prediction error or novelty, ACh upregulation supports encoding by adjusting the relative balance between input and output pathways among hippocampal subfields[16,46], and entraining theta and gamma frequencies[46,47]. An alternative hypothesis proposes that following prediction error or novelty, the ventral tegmental area (VTA) releases DA in the hippocampus and enhances encoding and plasticity[7,45]. Supporting this idea, multiple studies have shown that the hippocampus and VTA are co-activated after novel or surprising events[48,49], and DA strengthens synaptic transmission in the input pathway that connects hippocampal subfields[50]. In ongoing analyses, we are exploring how

prediction error influences connectivity among the hippocampus and neuromodulatory regions, and how functional connectivity relates to signal history and memory updating.

Conclusion

The brain continually generates predictions based on past experiences. When expectations do not align with reality, memories must be adaptively updated with new information. For the first time, we demonstrate that prediction error modulates hippocampal states and drives updating of naturalistic episodic memories. We show that prediction error reverses the relationship between hippocampal activation and subsequent memory: the hippocampus supports memory updating after prediction error, but protects memories from distortion after events that align with expectations. We track signal history in the hippocampus to show that prediction error disrupts the temporal continuity of neural representations and thus drives memory updating. Our findings suggest that prediction error prompts the hippocampus to abandon ongoing event representations and switch to an externally-oriented processing mode that supports memory updating and new encoding. In this way, surprising events modulate hippocampal states and determine the fate of episodic memories.

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.

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). 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.

Competing Interests: The authors have no competing interests to declare.

Methods

Participants

We recruited 55 participants from the University of Toronto community to participate in the study for monetary compensation (Reconsolidation group: $70, Immediate control group: $40). Of these participants, 7 were excluded (reasons stated below), leaving a final sample of 24 participants in each group. The sample size (N=48) was determined *a priori* to satisfy the following conditions: (a) achieve at least 90% power to detect the interaction effect previously found with a variant of this paradigm (η_p^2=0.17)[51], (b) reproduce the sample size previously used with a variant of this paradigm[22], and (c) 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) with fluency in English, normal or corrected-to-normal vision and hearing, and no history of neurological or psychiatric disorders. fMRI participants were all right-handed. In consideration of the effects of sleep on consolidation, we also asked participants to report approximate hours of sleep over the course of the study. Participants slept an average of 7.28 hours (*SD*=1.31) between the Day 1 and Day 2 sessions, and Reconsolidation group participants slept an average of 7.02 hours (*SD*=1) between the Day 2 and Day 3 sessions.

Exclusions. In the Immediate Control group, two participants were excluded due to technical failure with the video presentation software. In the Reconsolidation group, three participants were excluded due to a counterbalancing error and audio playback problems, and two participants were excluded because they had previously completed a pilot version of the study. Additionally, one full run of fMRI data (14 trials) was excluded for one participant due to audio playback failure and excessive motion. On a trial-by-trial basis, videos were excluded if

technical issues arose (e.g., audio failure) (10 trials), the participant was falling asleep (as determined by eyetracking) (20 trials), or the participant reported having seen the video prior to the experiment (103 trials). In total, there were 147 trials that were excluded for the above reasons (out of all 48 participants in both the Reconsolidation and Immediate groups). The total number of excluded trials for Full and Interrupted videos was approximately equal (Full: 70; Interrupted: 77). Additionally, subsequently forgotten videos were excluded from single-trial brain-to-behavior analyses (63 trials across the 24 participants in the Reconsolidation group). Overall, only 4.4% of all trials (70 trials per participant) were excluded.

Stimuli

Stimulus videos were sourced from movies, TV, and YouTube clips. We chose 70 videos that featured distinct action-outcome events. Semantic similarity varied across videos (e.g., several videos featured sporting events), but there were no overlapping sources, settings, or characters. The stimulus set included 18 videos that were previously used in a behavioral version of the paradigm. During pilot testing, we ensured that the videos would be infrequently recognized by our participants. The 70 videos used in the experiment are described in Supplementary Table 8 and publicly available on the Open Science Framework (https://osf.io/xb7sq/). The Interrupted version of each video ended abruptly at the narrative climax, violating the action-outcome contingency (duration M=24.79s, SD=4.08s).

For the fMRI version of the task (Reconsolidation group), stimuli were presented with EyeLink Experiment Builder (SR-Research) on a BOLDscreen display monitor (32”, 1920x1090, 100Hz refresh rate), viewed through a mirror attached to the head coil. Auditory stimulation was presented with in-ear MRI-compatible headphones (Sensimetrics, model S14). During the initial scout scan, we performed a sound test by playing the soundtrack of a movie

trailer. Participants reported whether the volume was appropriate for listening to the soundtrack over the scanner noise. The movie trailer used for the sound test was not included in the set of stimulus videos. For the behavioral version of the task (Immediate control group), videos were presented on a desktop computer and audio was presented with over-ear headphones.

Experimental Procedures

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

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

Test. Lastly, the Test session involved a structured interview-style recall test about details from each of the videos. Participants were cued with the name of each video and prompted to recall the narrative. The experimenter then probed the participant for more information with a pre-determined list of open-ended questions (e.g., “Can you describe the setting or context of the video?”, “Can you describe what the character looked like? Do you

remember gender, age range, hair color, or clothing?"). Participants were instructed to answer based on their memory of the full-length videos that had been originally presented during encoding. Because we were interested in false memories as a measure of memory updating, we instructed participants not to guess and permitted them to skip details they could not recall.

Group Differences. 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.

Previous reconsolidation studies have shown that context plays a critical role in memory updating[52–54], so we sought to maintain 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). Participants practiced lying still and viewed the videos while recorded MRI scanner sounds were played in the background. Reconsolidation group participants completed the Reactivation session in the real fMRI scanner and the Test session at a desk in the mock scanner room. Participants in the Immediate control group completed all three sessions in the same behavioral testing room. In both groups, participants completed all three sessions with the same experimenter.

Scoring of Memory Tests

We transcribed memory tests with *Temi*, an automated voice-to-text tool, then manually edited transcripts to verify accuracy. We coded videos as "forgotten" if the participant entirely failed to retrieve a memory when cued with the name of the video and a hint from a pre-determined list (brief descriptions of each video, provided in Supplementary Table 8). Scoring of

details was conducted with *NVivo 12*, a program for qualitative analysis of transcripts. Research assistants manually labelled each detail as correct or false. Lastly, we quantified semantic similarity among the videos by using the Cluster Analysis function in *NVivo* (Supplementary Figure 2). Across all transcripts, we pooled the phrases used to describe each video, excluding false memories and irrelevant words (e.g., *the, um, and, maybe, confidence, remember*). We then calculated pairwise Pearson correlations on the basis of the most frequent 100 words used to describe each video. For each video, we calculated an overall semantic similarity score by averaging the correlation values; this metric summarizes how much the content of a given video relates to the rest of the stimulus set. A heatmap displaying all pairwise correlation values is provided in Supplementary Figure 2.

Online Ratings of Stimulus Videos

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

interrupted, as well as video memorability and emotional valence/intensity (Supplementary Tables 9-11).

Linear Mixed-Effects Regression

All linear mixed-effects regression models reported in the main text included random intercepts for *subject* (identity of each participant) and *video* (identity of each stimulus item), along with random slopes for *reactivation type*. All models converged successfully; we used the BOBYQA controller with 5,000 maximum iterations. We used restricted maximum likelihood estimation and assessed significance of predictors with a type III ANOVA using Satterthwaite approximations of degrees of freedom. In R (v3.6), we constructed models with the *lme4* package[55] and evaluated significance with the *lmerTest* package[56]. Variables for *reactivation type* and *group* were treated as factors, and all predictive continuous variables were mean-centered. We used a profile procedure to calculate 95% confidence intervals on summary statistics[55]. These model parameters applied to analysis of behavioral data, single-trial univariate neural activation, and temporal autocorrelation.

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 blood-oxygen-level-dependent (BOLD) responses with a spatial resolution of 2.7mm isotropic voxels (TR: 1000 ms, TE: 29 ms, flip angle: 50°, 60 slices at transversal orientation, phase encoding: A>P, FoV: 210mm, Partial Fourier: 0.875, multiband factor: 4). High resolution T1-weighted anatomical images were acquired with a magnetization-prepared rapid-acquisition

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

fMRI Data Analysis

Preprocessing. All data were preprocessed and analyzed using FSL v6.0, in conjunction with in-house R code (v3.6). Initial volumes were discarded by the scanner to allow for signal saturation. Preprocessing steps included fieldmap distortion correction, spatial realignment, removal of head-motion artifacts (six regressors), nuisance regression of white matter and CSF timeseries, slice timing correction for an interleaved multiband acquisition, and high-pass frequency filtering (120s). For native-space ROI analyses (single-trial univariate and autocorrelation analyses), data were minimally smoothed with a 2-mm kernel in order to preserve spatial specificity and multivariate information.

Region of Interest (ROI) Masks. We used FreeSurfer (v6.0) to automatically create binarized hippocampal masks in each subject's native space. After FreeSurfer segmentation, hippocampal masks were manually inspected and segmented into ROIs for left anterior, left posterior, right anterior, and right posterior hippocampus. Anterior and posterior regions were split along the long-axis at the uncal apex. We found that our effects were very consistent among the four hippocampal ROIs (Supplemental Material, ROI Differences). Therefore, results reported in the main text (single-trial univariate and autocorrelation analyses) are averaged across the entire hippocampus bilaterally. White matter masks were obtained with FSL segmentation utilities. Inferior Lateral Occipital Cortex (LOC) masks were taken from the Harvard-Oxford Cortical Atlas and transformed into native space for each functional run, using the inverse deformation field from preprocessing and registration.

Univariate fMRI Analyses. Whole-brain mass univariate results are reported in the Supplemental Material (Whole-Brain Analysis, Supplementary Figure 3, Supplementary Table 3). 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 single trial, we employed the Least Squares-Single approach and constructed a separate GLM for each trial[57,58]. 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[24,25]. Within each GLM, the target trial (2s event) was isolated as one regressor, and all other events were modelled with a separate regressor for each type of event (e.g., video playback, video name cues, other fixation periods). This approach yielded whole-brain parameter estimates for each trial, in native space. For each trial, we masked the processed data and averaged across voxels within each ROI. Average activation values within each ROI were then submitted to linear mixed-effects regression, thus linking trial-wise ROI activation to reactivation type and subsequent memory.

Autocorrelation Analyses. Multivariate temporal autocorrelation analyses[28,32] were conducted on the same preprocessed data described above. We extracted the whole-run timeseries from every voxel within each ROI using the *fslmeants* utility. For control analyses (white matter and LOC ROIs), autocorrelation was calculated on 200 contiguous voxels, approximately matching the size of the hippocampal ROIs. Comparable to past research, we phase-shifted the timeseries by 4 seconds in order to account for HRF lag[59]. Temporal autocorrelation was defined as the Pearson product-moment correlation between all voxel activation values at timepoint T and timepoint T+1s. This method yielded an autocorrelation

value for every second of each functional run, excluding the final TR. Autocorrelation values were standardized (Fisher's *z*) prior to statistical analysis.

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

Data Availability. We have provided all derivative data and analysis code necessary to reproduce the results reported in the manuscript in a public repository hosted by the Open Science Framework (https://osf.io/xb7sq/). The full set of stimulus videos is also available for use in future studies.

References

1. Sinclair, A. H. & Barense, M. D. Prediction error and memory reactivation: How incomplete reminders drive reconsolidation. *Trends Neurosci.* **42**, 727–739 (2019).
2. Henson, R. N. & Gagnepain, P. Predictive, Interactive Multiple Memory Systems. *Hippocampus* **20**, 1315–1326 (2010).
3. Exton-Mcguinness, M. T. J., Lee, J. L. C. & Reichelt, A. C. Updating memories—The role of prediction errors in memory reconsolidation. *Behav. Brain Res.* **278**, 375–384 (2015).
4. Pine, A., Sadeh, N., Ben-Yakov, A., Dudai, Y. & Mendelsohn, A. Knowledge acquisition is governed by striatal prediction errors. *Nat. Commun.* **9**, 1–14 (2018).
5. Kim, G., Norman, K. A. & Turk-Browne, N. B. Neural differentiation of incorrectly predicted memories. *J. Neurosci.* **37**, 2022–2031 (2017).
6. Hindy, N., Avery, E. & Turk-Browne, N. Hippocampal-neocortical interactions sharpen over time for predictive actions. *Nat. Commun.* **10**, (2019).
7. Shohamy, D. & Adcock, R. A. Dopamine and adaptive memory. *Trends Cogn. Sci.* **14**, 464–472 (2010).
8. Chen, J., Cook, P. A. & Wagner, A. D. Prediction strength modulates responses in human area CA1 to sequence violations. *J. Neurophysiol.* **114**, 1227–1238 (2015).
9. Duncan, K., Ketz, N., Inati, S. J. & Davachi, L. Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus* **22**, 389–398 (2012).
10. Kumaran, D. & Maguire, E. A. An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS Biol.* **4**, e424 (2006).
11. Bein, O., Duncan, K. & Davachi, L. Mnemonic prediction errors bias hippocampal states.

Nat. Commun. **11**, 3451 (2020).

12. Chen, J., Olsen, R. K., Preston, A. R., Glover, G. H. & Wagner, A. D. Associative retrieval processes in the human medial temporal lobe: Hippocampal retrieval success and CA1 mismatch detection. *Learn. Mem.* **18**, 523–528 (2011).
13. Duncan, K., Curtis, C. & Davachi, L. Distinct memory signatures in the hippocampus: intentional states distinguish match and mismatch enhancement signals. *J. Neurosci.* **29**, 131–139 (2009).
14. Colgin, L. L. Rhythms of the hippocampal network. *Nature Reviews Neuroscience* vol. 17 239–249 (2016).
15. Honey, C. J., Newman, E. L. & Schapiro, A. C. Switching between internal and external modes: A multiscale learning principle. *Netw. Neurosci.* **1**, 339–356 (2017).
16. Meeter, M., Murre, J. M. J. & Talamini, L. M. Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus* **14**, 722–741 (2004).
17. Sherman, B. E. & Turk-Browne, N. B. Statistical prediction of the future impairs episodic encoding of the present. *Proc. Natl. Acad. Sci.* **8**, (2020).
18. Nader, K. & Einarsson, E. O. Memory reconsolidation: an update. *Ann. N.Y. Acad. Sci* **1191**, 27–41 (2010).
19. Nader, K., Schafe, G. E. & Le Doux, J. E. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406**, 722–726 (2000).
20. Lee, J. L. C. Reconsolidation: maintaining memory relevance. *Trends Neurosci.* **32**, 413–420 (2009).
21. Forcato, C., Argibay, P., Pedreira, M. & Maldonado, H. Human reconsolidation does not

always occur when a memory is retrieved: The relevance of the reminder structure. *Neurobiol. Learn. Mem.* **91**, 50–57 (2009).

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

23. Hupbach, A., Gomez, R. & Nadel, L. Episodic memory reconsolidation: An update. in *Memory Reconsolidation* (ed. Alberini, C. M.) 233–247 (Elsevier Academic Press, 2013). doi:10.1016/B978-0-12-386892-3.00011-1.

24. Ben-Yakov, A., Eshel, N. & Dudai, Y. Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes. *J. Exp. Psychol. Gen.* **142**, 1255–1263 (2013).

25. Ben-Yakov, A. & Dudai, Y. Constructing realistic engrams: Poststimulus activity of hippocampus and dorsal striatum predicts subsequent episodic memory. *J. Neurosci.* **31**, 9032–9042 (2011).

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

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

28. Brunec, I. K. *et al.* Multiple scales of representation along the hippocampal anteroposterior axis in humans. *Curr. Biol.* **28**, 2129-2135.e6 (2018).

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

30. Baldassano, C. *et al.* Discovering event structure in continuous narrative perception and

memory. *Neuron* **95**, 709-721.e5 (2017).

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

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

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

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

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

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

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

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

39. Radvansky, G. A. & Zacks, J. M. Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences* vol. 17 133–140 (2017).

40. Sederberg, P. B., Gershman, S. J., Polyn, S. M. & Norman, K. A. Human memory reconsolidation can be explained using the temporal context model. *Psychon. Bull. Rev.* **18**, 455–468 (2011).

41. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic Plasticity: How memory retrieval drives learning. *Trends Cogn. Sci.* **23**, 726–742 (2019).

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

43. Stawarczyk, D., Bezdek, M. A. & Zacks, J. M. Event representations and predictive processing: The role of the midline Default Network core. *Top. Cogn. Sci.* (2019) doi:10.1111/tops.12450.

44. Hasselmo, M. E., Wyble, B. P. & Wallenstein, G. V. Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* **6**, (1996).

45. Lisman, J. E. & Grace, A. A. The Hippocampal-VTA Loop: Controlling the entry of information into long-term memory. *Neuron* **46**, 703–713 (2005).

46. Newman, E. L., Gillet, S. N., Climer, J. R. & Hasselmo, M. E. 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).

47. Vandecasteele, M. *et al.* Optogenetic activation of septal cholinergic neurons suppresses sharp wave ripples and enhances theta oscillations in the hippocampus. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13535–13540 (2014).

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

49. Bunzeck, N. & Düzel, E. Absolute coding of stimulus novelty in the human substantia

nigra/VTA. *Neuron* **51**, 369–379 (2006).

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

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

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

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

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

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

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

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

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

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

Tables

		Correct Details		
Predictors	*Estimates*	*CI*	*p*	*df*
(Intercept)	13.46 ***	12.60 – 14.32	**<0.001**	49.75
Reactivation Type	-0.39 **	-0.67 – -0.11	**0.007**	69.45
Group	0.93 *	0.09 – 1.77	**0.036**	45.88
Reactivation Type * Group	-0.06	-0.24 – 0.11	0.488	248.01
		False Memories		
Predictors	*Estimates*	*CI*	*p*	*df*
(Intercept)	0.78 ***	0.68 – 0.88	**<0.001**	94.65
Reactivation Type	-0.06 ***	-0.09 – -0.03	**0.001**	154.76
Group	-0.36 ***	-0.43 – -0.29	**<0.001**	45.60
Reactivation Type * Group	0.05 **	0.02 – 0.08	**0.003**	270.31

Table 1. Parameter estimates from linear mixed effects regression models predicting correct details and false memories. * $p<0.05$ ** $p<0.01$ *** $p<0.001$

A) Hippocampal Activation

Predictors	*Estimates*	*CI*	*p*	*df*
(Intercept)	40.16 ***	29.67 – 50.65	**<0.001**	38.12
Reactivation Type	9.83 **	2.96 – 16.71	**0.007**	57.48
False Memories	0.11	-4.17 – 4.38	0.961	1218.47
Reactivation Type * False Memories	-5.63 **	-9.80 – -1.47	**0.008**	1047.43

B) Hippocampal Autocorrelation

Predictors	*Estimates*	*CI*	*p*	*df*
(Intercept)	0.23 ***	0.20 – 0.25	**<0.001**	23.57
Video Segment	0.002 ***	0.001 – 0.003	**0.004**	7434.00

C) False Memories

Predictors	*Estimates*	*CI*	*p*	*df*
(Intercept)	1.16 ***	0.99 – 1.32	**<0.001**	42.76
Autocorrelation Change	-0.47	-1.15 – 0.22	0.180	1352.23
Reactivation Type	-0.12 ***	-0.18 – -0.06	**<0.001**	93.73
Univariate Activation	0.01	-0.05 – 0.06	0.831	1374.30
Reactivation Type * Autocorrelation Change	0.74 *	0.06 – 1.41	**0.033**	1346.74
Autocorrelation Change * Univariate Activation	-0.11	-0.65 – 0.44	0.701	1368.18
Reactivation Type * Univariate Activation	-0.06 *	-0.12 – -0.00	**0.038**	1349.18
Autocorrelation Change * Univariate Activation * Reactivation Type	0.15	-0.39 – 0.70	0.582	1356.33

Table 2. Parameter estimates from linear mixed effects regression models of hippocampal activation. A) Predicting trial-wise univariate hippocampal activation from reactivation type (Full vs. Interrupted), subsequent false memories, and their interaction. B) Predicting hippocampal autocorrelation from segment (5s bins) during video playback. C) Predicting subsequent false memories from reactivation type, change in autocorrelation, and univariate activation in hippocampus. Boldface indicates statistically significant parameters. * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Figure Legends

Figure 1. Overview of experimental paradigm. A) Frames from a stimulus video named "Baseball", depicting a home run. To elicit prediction error (surprise), this video was interrupted while the batter was mid-swing during the Reactivation Phase (B). At Reactivation, participants were cued with the video name, watched the video (Full or Interrupted), then viewed a fixation screen. fMRI analyses focused on the post-offset period after each video (highlighted box), treating the offset of the video as the stimulus. C) Example text illustrating the memory test format and scoring. Participants answered structured interview questions about all 70 videos, and were instructed to answer based on their memory of the full-length video originally shown during encoding. 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, while the Immediate control group did the Test Phase immediately after Reactivation, in order to investigate whether memory updating required a delay. Only the Reconsolidation group was scanned.

Figure 2. Prediction error modulates correct details versus false memories over distinct timecourses. A) In both groups, average Correct Details were higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error can strengthen memory recall both immediately and after a delay. B) Only in the Reconsolidation group were average False Memories higher for videos that were Interrupted during memory reactivation, demonstrating that prediction error drives memory updating but requires a delay to permit the protein synthesis that underpins reconsolidation. Dots indicate average scores by-participant, and lines connect within-subjects measures. Error bars depict 95% confidence intervals. * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 3. Prediction error reverses the relationship between hippocampal activation and subsequent memory. After Full-length videos, greater hippocampal activation had a protective effect, predicting fewer false memories (blue). After Interrupted videos, greater hippocampal activation predicted more false memories (orange). Lines depict model-predicted estimates, and shaded bands depict the 95% confidence interval.

Figure 4. Tracking hippocampal signal history during and after events. A) Signal history (temporal autocorrelation) in the hippocampus gradually ramps over the course of a video. Autocorrelation values are averaged over 5-second bins of video playback. B) Second-by-second autocorrelation values in the hippocampus, time-locked to the moment of video offset (black dotted line). Signal history for Full and Interrupted videos diverges after video offset. C) Comparing post-offset change in signal history, using average autocorrelation scores for the 5-

sec bins immediately before and after video offset. Prediction error halts signal accumulation after video offset. Error bars depict SEM.

Figure 5. Prediction error resets signal history and drives memory updating. A) Estimated values from a linear regression model predicting subsequent false memories from the interaction of reactivation type and change in autocorrelation (i.e., signal history). After Interrupted videos, decreases in autocorrelation (negative values indicating resetting of signal history) were related to more memory updating. Shaded bands depict 95% confidence intervals around the regression line. B) Schematic depicting hippocampal mode switching and subsequent memory. During a video, signal history ramps in the hippocampus. After video offset, preservation or resetting of signal history indicates whether the hippocampus remains in an internal processing mode where univariate activation reflects ongoing prediction and retrieval (protecting against distortion), or an external processing mode where univariate activation reflects updating and encoding (increasing false memories).