

Reviewed Preprint

v1 • February 20, 2026

Not revised

Reviewed Preprint

v2 • August 14, 2026

Revised by authors

✉ For correspondence:

alicexue@stanford.edu

Competing interests:

No competing interests declared

Funding:

See [page 34](#)

Reviewing editor: Lila Davachi, Yale University, United States

© 2026, Xue et al. This article is distributed under the terms of the

[Creative Commons Attribution](#)

[License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Strong mnemonic prediction errors increase cognitive control, attention, and arousal

Alice M Xue^{1,2} ✉, Jenna Jokhani^{1,2}, Anthony M Norcia^{1,2}, Anthony D Wagner^{1,2}

¹Stanford Department of Psychology, Stanford, United States • ²Stanford Wu Tsai Neurosciences Institute, Stanford, United States

eLife Assessment

The work presented represents **important** new evidence linking the cascade of neural processes triggered by memory-based prediction errors. The study uses an impressive collection of approaches and methods to characterize and measure cognitive control, arousal, and memory changes as a function of memory-based violations. Analyses are technically sophisticated and rigorous and, taken together, provide **solid** evidence that there are multiple processes accompanying prediction errors, and that they differentially relate to successful encoding.

<https://doi.org/10.7554/eLife.109678.2.sa3>

Abstract

Ongoing experience is continuously processed in the context of past events. Divergence between current experience and memory-based predictions (i.e., a mnemonic prediction error; MPE) is theorized to be a signal for the hippocampus that encoding, as opposed to retrieval, should be prioritized. We asked how MPEs place demands on cognitive and neural resources beyond the hippocampus, and whether these demands differ as a function of prediction strength. We investigated these questions across two experiments, wherein we recorded scalp electroencephalography and/or pupillometry as 101 young human adults performed an associative memory task. Strong MPEs, more so than weak MPEs, increased physiological indices of cognitive control (frontal theta), attention (posterior alpha and pupil size), and arousal (pupil size). Trial-level pupil-linked MPE responses scaled with the amount of attention (posterior alpha) allocated during prediction generation. Finally, greater cognitive control (frontal theta) during strong MPEs promoted better learning of prediction violating (i.e., unexpected) stimuli. Collectively, these findings reveal multiple pathways through which the mind and brain respond adaptively to violations of strong mnemonic predictions.

Introduction

Internal models of the world are constructed using memories of past experiences. Knowledge retained from events in the past enables predictions about ongoing and future events, allowing for adaptive behavior. As events unfold, alignment between what is experienced and what is predicted provides validation for the current internal model and signals that continuing to retrieve information from memory may be useful for guiding behavior. On the other hand, divergence of ongoing experience from mnemonic predictions signals that the environment has potentially changed and that past experiences may not be informative or relevant for current and future behavior. Such mnemonic prediction errors (MPEs) may elicit increased attention and arousal¹ and signal that more cognitive control — the collection of cognitive and neural mechanisms that effortfully influence information processing to be aligned with one's goals² — is needed.

Insights into the effects of MPEs on attention and arousal have come from studies of event segmentation. According to Event Segmentation Theory^{3,4}, the mismatch between one's internal model and ongoing events can elicit a prediction error, leading to event segmentation (i.e., the psychological perception of an event boundary along with accompanying neural computations), which discretizes current experience into “episodes” in memory^{5,6}. Extant data indicate that increases in pupil size are observed at event boundaries^{7,8} and in response to prediction errors of varying content, signs, and modalities^{9–11}. Such changes in pupil size reflect increases in attention and arousal^{12–14} and may be linked to noradrenergic activity in the locus coeruleus (LC)^{14,15}. The LC has been associated with signaling strong prediction errors^{16,17}, also characterized as “global model failure signals”¹⁶ that initiate a network reset¹⁸. The LC may further support adaptive behavior in the face of MPEs by enhancing memory encoding to facilitate model updating¹⁹. New learning following MPEs may depend on increases in attention, by direct modulation of hippocampal learning mechanisms¹⁹, or both. While MPE responses in the hippocampus appear to vary with prediction strength²⁰, it is unclear whether effects on attention and arousal are similarly modulated. Here, we hypothesize that pupil size will show a greater increase in response to strong compared to weak MPEs, that MPE-induced increases in attention and arousal will support learning, and that learning from MPEs will differ as a function of prediction strength¹¹.

Complementing evidence of MPE-related increases in pupil size^{7,8,11}, time-resolved electrophysiological measures of visuospatial attention — specifically, electroencephalography (EEG) posterior alpha power suppression^{21,22} — may also show that attention increases following MPEs, evidenced by greater alpha suppression²³. The underlying sources of these measures of attention may be partially shared, given that pupil size and posterior alpha are coupled in certain contexts^{24–26}. Moreover, LC lesions in monkeys reduce the amplitude of the P3 event-related potential (ERP), another electrophysiological measure of attention/alertness²⁷. However, it remains unclear whether LC activity additionally modulates posterior alpha suppression. Posterior alpha and pupil size also could be differentially influenced by other mechanisms of MPE detection, such as those subserved by the hippocampus and/or prefrontal cortex (PFC)²⁸.

The hippocampus is thought to compute MPEs^{29–32}, with the CA1 subfield being anatomically well-suited for mismatch detection given that it is a convergence zone for mnemonic predictions from hippocampal subregion CA3 and sensory input from entorhinal cortex^{33–36}. Functional neuroimaging studies provide support for CA1 as a mismatch “detector” in humans^{20,37,38}. Moreover, following MPE signals from the hippocampus, LC, or other regions, learning rates in the hippocampus are posited to increase³³. Such increases may be underpinned, at least in part, by a biasing of hippocampal functional connectivity towards an encoding-favored configuration³⁹, which may enhance the integration of bottom-up sensory information during hippocampal encoding computations⁴⁰. Separately, increases in hippocampal theta power^{41,42} may increase the likelihood of encoding novel information into memory⁴³. Finally, dynamic causal modeling of magnetoencephalography (MEG) data suggests that mismatch detection in the hippocampus, as assayed by theta power, may be driven by theta-related signals from ventromedial prefrontal cortex (vmPFC)⁴². Hippocampal theta, which may be challenging to detect with scalp EEG⁴⁴, could modulate or be modulated by cognitive control and attention signals that are often assayed using scalp EEG.

PFC theta, observed using MEG^{42,45} and scalp EEG^{46–48}, is known canonically as frontal theta⁴⁹ or frontal midline theta⁵⁰ and is considered a marker of cognitive control^{49,51}. Frontal theta increases following reward prediction errors^{46,48,49}, scales with negative prediction errors⁴⁶, and fosters behavioral adaptation⁴⁶. Behavioral adaptation in response to increases in frontal theta may be mediated by changes in information processing⁴⁹ by way of increased attention. Control may directly increase attention to and prioritization of bottom-up sensory information from the external environment and the overcoming of memory-guided, but now irrelevant, representations and behavioral responses⁵². In rodents, PFC activity can modulate LC activity⁵³, raising the possibility that the behavioral adaptation effects associated with increased frontal theta could be mediated in part by changes in LC activity and thus pupil-linked attention and arousal. Whether MPEs elicit (a) an increase in cognitive control that (b) consequently results in increases in

attention and arousal to facilitate adaptive behavior remain open questions. A primary aim of the current study was to test whether there are interactions among cognitive control, attention, and arousal upon detection of MPEs.

Characterization of potential process interactions may additionally provide critical insights into potential mechanisms by which MPEs facilitate learning in service of future behavior. Higher learning rates following MPEs may be mediated, in part, by increases in attention and arousal¹, which facilitate memory encoding^{11,54,55}. In addition, increases in pupil-linked attention and arousal following MPEs could reflect direct involvement of LC activity in fostering learning, given that optogenetic activation of LC neurons in mice enhances memory encoding⁵⁶ and higher LC integrity among older adults is associated with better episodic memory performance⁵⁷. LC effects on learning may also be mediated by the hippocampus, which receives profuse projections from LC neurons⁵⁶. Increases in attention and arousal following MPEs may require the initial detection that cognitive control, as instantiated via frontal theta⁴⁹, is needed. Higher frontal theta prior to⁵⁸ and during^{51,59} encoding has independently been associated with better subsequent memory, leaving open the additional possibility that MPE-related learning effects stem from changes in frontal theta. Finally, MPE-driven increases in frontal theta may also enhance learning by driving hippocampal theta⁴², which also is associated with better subsequent memory⁴³.

There are thus multiple possible pathways by which MPEs impact cognition. Here we focus on two mechanisms: (1) increases in cognitive control (assayed via frontal theta) and (2) increases in attention and arousal (assayed via posterior alpha and pupil size). To elicit MPEs, we used an associative memory task wherein participants were tested on their memory for verb-picture pairings they had studied four times (strong memories) or once (weak memories). Associative memory of a paired picture was cued by the presentation of a studied verb. On a subset of trials, verb presentation was followed by the presentation of a mismatched picture. Onset of a mismatch would indicate a violation of a mnemonic prediction and the elicitation of an MPE. Using time-resolved scalp EEG (Experiment 1) and pupillometry (Experiments 1 and 2) measured as participants completed the associative memory task, we document the effects of strong vs. weak MPEs on cognitive control, attention, and arousal. Then, we characterize how control signals interact with attention and arousal signals, and how EEG and pupil assays of attention/arousal relate to each other. Throughout, we examine how these electrophysiological and pupillary measures differ as a function of memory strength to investigate how trial-level assays of prediction strength relate to MPE responses. Finally, we examine whether MPE-related changes in each of these physiological assays relate to the encoding of prediction-violating stimuli, as evidenced by subsequent memory for mismatch probes.

Collectively, our findings show that strong MPEs, relative to weak MPEs, evoke greater increases in cognitive control (frontal theta), attention (posterior alpha suppression and pupil), and arousal (pupil). The magnitude of immediate MPE responses not only varies as a function of strong vs. weak predictions, but also with trial-level assays of the amount of attention allocated to the mnemonic prediction (posterior alpha suppression). Exploratory mediation analyses did not provide strong support for the hypothesis that increases in attention and/or arousal induced by MPEs can be explained in part by increases in cognitive control. Finally, increases in frontal theta, but not posterior alpha or pupil size, are linked to enhanced learning of strong prediction-violating (i.e., mismatch) stimuli.

Methods

We conducted and, for most analyses, pooled data across two experiments that used a common associative memory task paradigm with pupillometry. Experiments 1 and 2 differed in: participant recruitment pool; the duration of the delay between the Associative Memory Task and a final subsequent Recognition Test on the mismatch probes; day of administration of the gradual continuous performance task (gradCPT); inclusion of a Localizer Task (only administered in Experiment 1); and concurrent recording of EEG (only collected in Experiment 1). Images of animate and inanimate objects were taken from the BOSS database⁶⁰, an Animacy x Size stimulus set⁶¹, and the internet. Objects were isolated from their white or complex backgrounds and were

then overlaid on a gray background for stimulus presentation. Scene stimuli in the Localizer Task in Experiment 1 were taken from the “Massive Memory” Scene Categories stimulus set⁶². Luminance of image stimuli in all tasks, with the exception of the gradCPT, were controlled using the SHINE Toolbox⁶³.

Experiment 1: Associative Memory Task with Same-Day Mismatch Probe Recognition Test (scalp EEG and pupillometry)

Fifty-one healthy young adults (aged 18-35 yrs) enrolled in Experiment 1. One participant withdrew early and all data from this participant were excluded. The final sample included 50 participants (25 female; mean age = 23.1 yrs, s.d. = 3.3). Pupil data were not properly tracked from three participants and were excluded from the corresponding pupil analyses. Participants were recruited from Stanford University and the surrounding community, were right-handed, and had normal or corrected-to-normal vision. The study was approved by the Stanford University Institutional Review Board and all participants provided written informed consent. In a single experimental session, participants completed 2.5hrs of behavioral tasks, with the subsequent recognition memory task occurring 6-min following the associative memory task. Throughout, EEG data were recorded using a 128-channel HydroCell Sensor Net (Electrical Geodesics) and pupillometry data were collected using an Eyelink 1000 eye-tracker; EEG and pupillometry data were collected at a sampling rate of 1000 Hz. Following EEG net application, electrode positions were recorded using a LiDAR sensor on an iPad using the Polycam application (<https://poly.cam/>). Before each task (ranging from 3–16min in duration), EEG impedances were checked and calibration and validation of the eye-tracker were performed.

Experiment 2: Associative Memory Task with Delayed Mismatch Probe Recognition Test (pupillometry)

Fifty-three healthy young adults (aged 18-23 yrs) enrolled in Experiment 2. Participants were recruited from the Stanford Psychology Credit Pool and had normal or corrected-to-normal vision. Two participants withdrew on Day 1 and all data from these participants were excluded. One participant withdrew before Day 2 of the experiment; Day 1 data from this participant were retained. The final sample included 51 participants (23 female; mean age = 19.8 yrs, s.d. = 1.4). Other data loss/inclusion: one participant’s pupil was not properly tracked throughout the experiment and these pupil data were excluded from corresponding analyses; due to experimenter error, the subsequent memory test was administered incorrectly for four participants on Day 2 and these data were excluded from the corresponding analyses; one participant’s subsequent recognition data were excluded because the participant responded “Old” on all trials. The study was approved by the Stanford University Institutional Review Board and all participants provided written informed consent. In Session 1, participants completed 2hrs of behavioral tasks. Session 2 took place 48hrs after Session 1, during which participants completed the subsequent recognition task and the gradCPT. In both sessions of the experiment, pupillometry data were collected during behavioral tasks using an Eyelink 1000 eye-tracker at a sampling rate of 1000Hz. Calibration and validation were performed before each task.

Experimental Design

1.1 Associative Memory Task

Participants performed an associative memory task, which consisted of an associative encoding phase and an associative retrieval test. The associative memory task was administered in four study-test cycles, with encoding and retrieval phases separated by a 3min distractor task wherein participants reported the parity of sequentially presented integers ranging from 1 to 100. Each integer was presented immediately following a response. The purpose of the distractor task was to prevent participants from maintaining the studied pairs in working memory prior to the associative retrieval test.

In the associative encoding phase, participants were instructed to learn verb-picture pairs by forming a mental image involving the verb and the image in each pair (Fig. 1a). On each trial, an action verb was presented at the center of the screen for 1.5s. The subsequent inter-stimulus interval (ISI) ranged from 0.5-1.5s after which a picture of an animate or inanimate object was then displayed for 2.5s. During the presentation of the object, participants generated a mental image and reported (using a button press) whether their mental image of the verb-picture pairing was “very vivid” or “not very vivid.” The inter-trial interval (ITI) ranged from 1.25–1.5s. Participants were shown 52 pairs in each encoding phase; 26 were presented four times over the course of an encoding phase cycle (“Strong” pairings) and 26 were presented once (“Weak” pairings). Repeated exposures to strong pairings were distributed across the encoding phase cycle, such that each strong pairing was viewed once during each quartile of trials. As such, each cycle of encoding included 130 trials (26 pairs presented four times and 26 pairs presented once). A black fixation dot was presented at the center of the screen during all trials and participants were instructed to fixate throughout. Verb-picture pairings were randomly assigned per participant, as was the assignment of pairs to the Strong and Weak conditions.

After the distractor phase, an associative retrieval test examined memory for the verb-picture pairings (Fig. 1b). On each trial, a verb cue was presented for 1s. Participants were instructed to try to bring to mind the associated picture as soon as the cue appeared on the screen. Following a 1s delay, during which only the central fixation dot was presented on the screen, a picture was then presented for 0.75s. Participants were instructed to respond as quickly as possible, using a button press to indicate whether or not they thought they had studied the picture with that verb. Memory for all 52 pairs in the encoding phase of a given cycle was tested. The brief (0.75s) presentation duration of the test probe and instruction to respond rapidly were designed to encourage participants to retrieve the associate upon presentation of the verb cue, setting up a situation in which memory of the associate and perception of the probe either match or mismatch. To focus on trials on which retrieval likely preceded probe onset, retrieval task trials were retained for analysis if responses were made within 1s after probe onset. In each study-test cycle, 36 trials were “Match” trials, wherein the picture after the verb matched that studied in the encoding phase, and 16 were “Mismatch” trials, wherein the picture did not match what was studied. On Mismatch trials, the probe picture was from a different category as the correct associate (i.e., if the picture associated with a verb cue was animate, the mismatch probe was inanimate, and vice versa). To control for stimulus novelty during retrieval, mismatch stimuli were initially introduced to participants at the outset of the experiment in a pre-exposure 1-back task (see below). ITIs ranged from 1.25–1.5s. A black fixation dot was presented at the center of the screen at all times and participants were instructed to fixate throughout. The test order of pairings was randomized.

Each study-test cycle included unique verbs and pictures. The assignment of each picture to each verb for Match and Mismatch pairings was randomized across participants.

1.2 Pre-Exposure Task

At the outset of the experiment, prior to the associative memory task, participants performed a 1-back task in which they were shown a sequence of images (each presented for 2s) separated by variable-length ITIs (1.0-1.2s). For each stimulus, participants were instructed to report, using a button press, whether the current stimulus was the same as or different from the previous stimulus. To maximize the duration with which participants attended to each stimulus, they were instructed to provide their response when each stimulus had left the screen. There were 96 unique stimuli in total (48 animate and 48 inanimate objects); 64 of the stimuli were later presented as mismatch probes during the retrieval phase of the associative memory task (see above), and all stimuli were later presented in the subsequent recognition test; the order of the stimuli was randomized. Of the 32 items not presented in the retrieval phase of the associative memory task, nine were repeated in the 1-back task. A black fixation dot was presented at the center of the screen at all times during all trials and participants were instructed to fixate on the central dot throughout the task. The purpose of this pre-exposure task was to familiarize participants to the

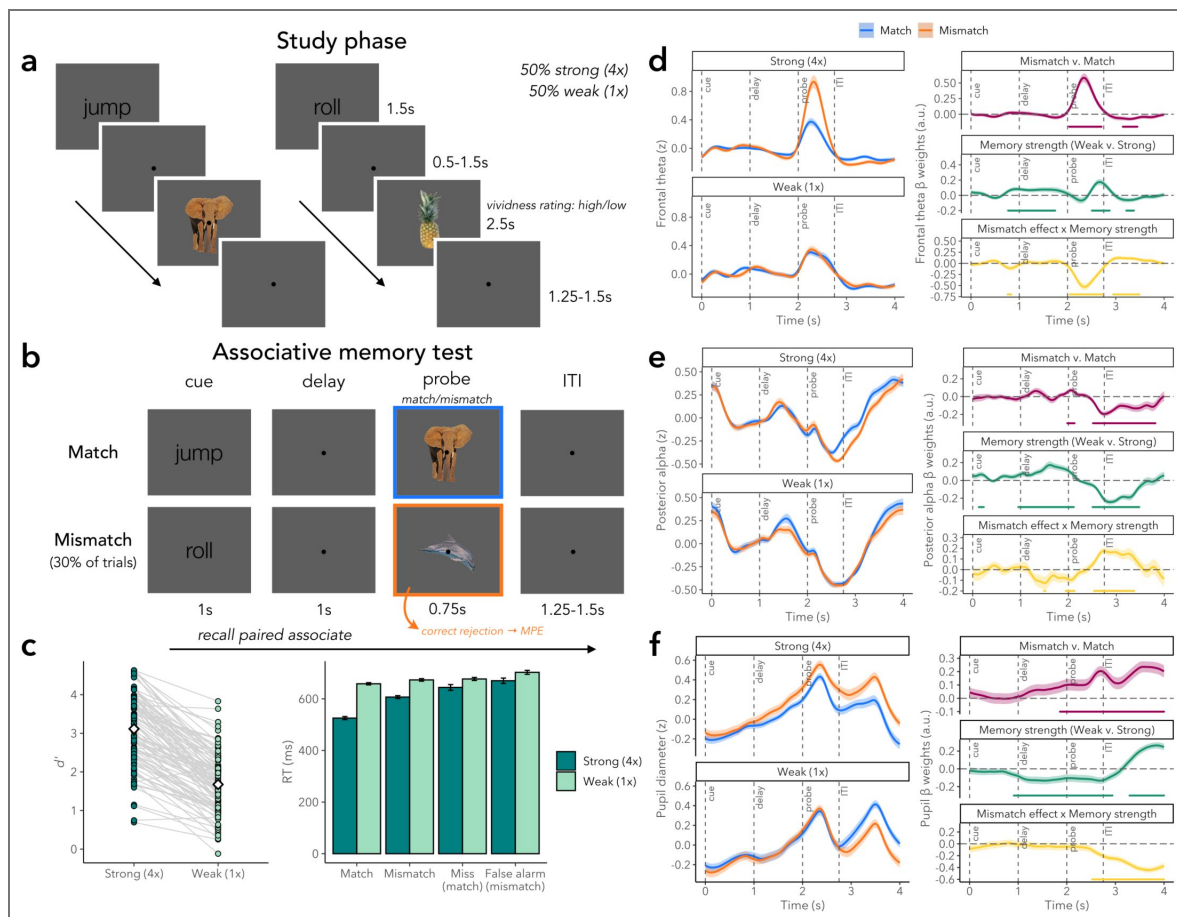


Figure 1. Associative memory task design and memory behavior.

(a) Study phase task design. (b) Match/mismatch test phase task design. (c) Behavior in the match/mismatch retrieval task. D' (left) and RTs (right) in dark teal for strong pairings and light teal for weak pairings. Frontal theta power (d), posterior alpha power (e), and pupil diameter (f) for strong (top) and weak (bottom) trials, color coded by match/mismatch condition (blue/orange). Subplots on the left depict mean and standard error of z-scored power/size across participants. Mean and standard error of participant-level betas for the effects of Mismatch, Strength, and Mismatch x Strength are depicted on the right. Horizontal lines below beta plots indicate significant clusters (permutation-corrected; all $p < 0.001$).

stimuli that would later serve as mismatch probes during the associative memory task. By using mismatch probes that were familiar rather than novel, item novelty could not be a primary source of memory evidence guiding decisions in the retrieval phase of the associative memory task.

1.3 GradCPT

To obtain a subject-level assay of attention, we collected 5-min of gradCPT⁶⁴ data in each experiment with modified stimulus presentation timings. In Experiment 1, participants performed the gradCPT after four cycles of the associative memory task to introduce a delay before the surprise recognition test (see below) and prevent maintenance of material studied in the associative memory task in working memory. In Experiment 2, where the surprise recognition test took place 48hrs after the associative memory test, the gradCPT was performed after the recognition test in Session 2. In the gradCPT, a centrally presented picture gradually transitioned between different mountain and city images (90% of images were of cities and 10% were of mountains). Each image was presented for 0.05s and transitioned to the next image over 0.8s. Participants were instructed to press the spacebar when they saw a city scene and to withhold responding for mountain scenes. Performance in the gradCPT was not central to the core hypotheses and is not reported in the current manuscript.

1.4 Recognition Test

To assess the effects of MPEs on learning, a recognition test was administered. In this surprise recognition memory test, participants were tested on their memory for the 64 mismatch probes shown during the associative retrieval test. Non-mismatch foils on the recognition test consisted of the 32 stimuli from the pre-exposure 1-back that were not repeated in that task nor used as mismatch probes during the associative retrieval test; as such, they were old but not associated with MPEs. An additional 64 novel foils were included in the recognition test. On each recognition trial, participants reported whether they had seen the image at any point during the experiment, by pressing one button for “old” and another for “new”. Each image was displayed for a minimum of 1.2s and a maximum of 2s, such that if participants failed to respond within 1.2s, the image remained on the screen for an additional 0.8s. ITIs varied from 0.5–1.5s and the order of the stimuli was randomized. A black fixation dot was presented at the center of the screen at all times and participants were instructed to fixate. In Experiment 1, the recognition test was administered after the gradCPT. In Experiment 2, the recognition test was the first task administered in Session 2. The recognition test examined subsequent memory for mismatch probes either ~5-10min (Expt 1) or ~48hrs (Expt 2) after the associative memory task.

1.5 Localizer Task

At the end of Experiment 1, participants performed a second 1-back task designed to enable training of an EEG classifier that discriminates between evoked electrophysiological activity patterns associated with animate and inanimate objects. The localizer task was included in Experiment 1 to address other research questions and results from this task are not reported in the current manuscript. For completeness, the task is described here. On each trial, an image of an animate object, an inanimate object, or a scene⁶² was displayed on the screen for 2s. Images were separated by a 1.0-1.2s ITI. The instructions were the same as for the pre-exposure 1-back task, such that participants were instructed to report whether a given image was the same as or different from the previous image, and to respond after the image had left the screen. The stimuli were novel (i.e., had not appeared in any other experimental phase), with 270 unique stimuli presented in total (90 per category); the order of the stimuli was randomized. Ten percent of the stimuli repeated once (as 1-backs) over the course of the task.

EEG and Pupillometry Analyses

2.1 EEG Preprocessing

EEG data from the associative retrieval test in Experiment 1 are a focus of the present manuscript, given the open questions/hypotheses regarding mnemonic prediction errors. Preprocessing of these EEG data was performed using the MNE package⁶⁵. EEG data were first downsampled from

1000 Hz to 250 Hz and then filtered with a high-pass filter of 1 Hz and a low-pass filter of 30 Hz. Bad channels were then identified via visual inspection of the power spectrum and were subsequently interpolated. An ICA was fitted to the data; ICs resembling blinks and muscle artifacts were identified automatically using functions in MNE and by visual inspection. Data were then re-referenced to the average.

2.2 EEG Spectral Analyses

For analyses of spectral power in the associative retrieval task, trial-level power was computed on epochs of data from -1s to 5s with respect to trial onset. Epoched data were decimated by a factor of two to reduce computation time, yielding a sampling frequency of 125Hz. Epochs wherein the voltage of any channel of interest exceeded 100 μ V were excluded from analyses. Frontal theta (4-7Hz) power was computed using electrodes 5, 6, 7, 11, 12, 13, 106, 112 and posterior alpha (8-12Hz) power was computed using electrodes 62, 67, 71, 72, 75, 76, 77 (Supplementary Fig. 1 [\[3\]](#)). These electrodes were selected based on prior research on attention lapsing and goal coding⁶⁶. Instantaneous power was computed at each time point using a Morlet wavelet with 7 cycles and epochs were subsequently trimmed to 0s–4s after trial onset for subsequent analyses.

2.3 Pupillometry Preprocessing

Pupillometry size data from both experiments were preprocessed⁶⁷, with data first being decimated from 1000 Hz to 100 Hz. A pupil dilation speed threshold was determined by first computing the median absolute deviation (MAD) in pupil size in each block of associative retrieval task data. Visual inspection was used to assign 10 as the constant value for multiplicatively scaling the MAD. A rejection threshold was then determined as the sum of this value and the median dilation speed. Samples of data for which the change in pupil diameter from the previous sample exceeded this threshold were removed. Blinks were padded by 50ms on each side and linear interpolation was used to fill in missing data due to blinks or dilation artifacts. The data were then smoothed using a third order Butterworth filter and additional outlier samples were identified using the dilation speed threshold procedure described above, but with a constant factor of 0.2. This procedure of interpolation, smoothing, and outlier detection was then performed a second time. Finally, trials were excluded if more than 10% of data points were missing in the original time series. Data were subsequently de-trended by taking the residuals of a linear regression model fit to the time series for each run of each task. For all analyses, the z-scored residuals were analyzed.

2.4 Statistical Models for Time Series Analysis

Trial-level regression analyses were conducted on the frontal theta, posterior alpha, and pupil time series from the associative retrieval test to identify temporal clusters that were sensitive to the factors of Mismatch (i.e., mismatch vs. match probes) and/or Strength (i.e., strong vs. weak associative pairs). For each participant and each time point, a linear regression model testing main effects of Mismatch and Strength, and a Mismatch $\times$ Strength interaction was run using R⁶⁸ to compute beta weights for each regressor. For these models, only correct trials (i.e., hits to match probes and correct rejections to mismatch probes) were included to identify temporal clusters associated with memory retrieval strength and/or MPEs. For frontal theta and posterior alpha analyses, two participants in Experiment 1 were excluded because they were missing trials in at least one of the four conditions and model fitting could not be performed; for pupil analyses, two participants from Experiment 2 were excluded for missing trials in at least one of the four conditions. For analyses of subsequent recognition memory for mismatch probes, 10 participants from Experiment 1 and seven from Experiment 2 were excluded for missing trials in at least one of the four conditions; of these, seven participants from Experiment 1 were excluded from the frontal theta and posterior alpha time series analyses of subsequent recognition memory. For primary pupil analyses, data were pooled across Experiments 1 and 2 (N=94); see Supplement for results by experiment.

Significance testing for each regressor was conducted at each time point and was performed using one-sample t-tests. The beta weights computed for each time point were averaged across participants for visualization of the effects of Strength, Mismatch, and Mismatch $\times$ Strength interactions for frontal theta, posterior alpha, and pupil size over time. Contiguous time points with significant t-values were identified; the significance of these temporal clusters was then tested using permutation tests in which the sign (positive or negative) of each participant's beta weights was randomly assigned; this procedure was repeated 1000 times and t-tests for each time point were computed on each iteration to generate a null distribution. Significance testing for each temporal cluster was performed by computing the probability of a cluster that was at least as long in the null distribution.

2.5 Pupillary Temporal PCA

In addition to canonical pupil size analyses, the pupil data were deconstructed using a temporal principal components analysis (PCA). Temporal PCA was used to obtain trial-level assays of attention and arousal (1) with greater sensitivity to cognitive processes engaged during the task and (2) that could be related to electrophysiological measures on a trial-by-trial basis. PCA was performed on the pupil time series aggregated across all correct trials and all participants ($N=96$); the size of the input matrix was the number of trials $\times$ the number of time points (10346×401 time points). We retained the first six components, using at least 95% variance explained as the cutoff; cumulative variance explained by all six components was 96.16%. To obtain interpretable scores for each trial and each PC, the loadings were then orthogonalized using a varimax rotation on the rotation matrix produced by the PCA. For each component, we fit linear mixed effects models using lme4⁶⁹ to examine the effects of Mismatch and Strength (score $\sim$ Mismatch $\times$ Strength + (Mismatch $\times$ Strength | participant)). We also examined how scores related to reaction time (RT) on hit and correct rejection trials (RT $\sim$ score $\sim$ Mismatch $\times$ Strength + (Mismatch $\times$ Strength | participant)). RTs were log transformed in all analyses to account for skewed distributions. A maximal random effects approach was adopted as recommended by Barr⁷⁰; for models that failed to converge, we successively simplified the random effects structure.

2.6 Multiple Mediation Analyses

To characterize the dynamics between cognitive control, attention, and arousal following a strong MPE, Bayesian multivariate mediation analyses⁷¹ were conducted using the brms⁷² package in R. All models were conducted on strong Mismatch trials. The first model tested whether the relationship between MPE magnitude, approximated using strong mismatch RTs, and immediate pupil MPE effects (PC3) was mediated by frontal theta effects (mean power in the second Mismatch $\times$ Strength cluster in Fig. 1d⁶⁸). The second model tested whether the relationship between strong mismatch RTs and posterior alpha MPE effects (mean power in the third Mismatch $\times$ Strength cluster in Fig. 1f⁶⁸) was mediated by frontal theta effects (mean power in the second Mismatch $\times$ Strength cluster in Fig. 1d⁶⁸). In all models, frontal theta power, PC3 scores, and posterior alpha power were z-scored within participants and intercept terms were excluded. RTs were log transformed to account for potential outliers and a skewed distribution, then z-scored. An additional exploratory model was tested, examining whether the relationship between mismatch RTs and frontal theta was mediated by posterior alpha.

2.7 Cross-Correlation Analyses

To further characterize the temporal dynamics between the different physiological measures, we conducted cross-correlation analyses on the trial-level 4s-long frontal theta, posterior alpha, and pupil retrieval time series for strong Mismatch trials. Because the pupil time series was downsampled in previous analyses to 100Hz whereas the EEG power time series were downsampled to 250Hz, the pupil time series was upsampled to match the frequency of the EEG power time series. Given the relatively slower dynamics of the pupil, this upsampling should not substantially impact the temporal profile of the pupil time series. A lag window of 1s was defined and the correlation coefficients for different lags were obtained for each pair of trials. Correlation coefficients were averaged across trials for each participant. Significance testing for each lag was performed using one-sampled t-tests and p-values were Bonferroni corrected.

Results

1. Effects of mnemonic prediction errors

Integrating behavior, electrophysiology, and pupillometry, we examined the core hypothesis that mnemonic prediction errors (MPEs) trigger an increase in cognitive control, attention, and arousal. We tested this hypothesis through condition comparisons of prediction errors following strong vs. weak associative encoding, and through trial-level assays of varying associative retrieval/prediction strength (via readout from EEG spectral power during retrieval; canonical and principal component pupil assays). We then explored whether such increases in control, attention, and/or arousal predict subsequent memory for the unexpected/mismatch stimuli.

Prediction strength impacts match/mismatch choice behavior

In a combined sample of 101 participants, associative memory d' ($Z_{\text{associative hit}} - Z_{\text{false alarm}}$) was higher for strong (3.11 ± 0.93) compared to weak pairings (1.67 ± 0.76) ($t_{100} = 20.988$, $p < 0.001$; Fig. 1c [↗](#); see Supplementary Fig. 2 [↗](#) for data separated by experiment; see Table 1 [↗](#) for hit and false alarm rates). Consistent with participants retrieving the associate prior to probe onset, probe-locked response times (RT) were relatively rapid, with across-participant mean RT occurring prior to offset of the probe. RTs were faster for strong compared to weak trials (main effect of Strength: $\beta = 0.211$, $CI = [0.202, 0.219]$, $p < 0.001$), correct than incorrect trials (Accuracy: $\beta = 0.200$, $CI = [0.171, 0.229]$, $p < 0.001$), and match than mismatch trials (Mismatch: $\beta = 0.165$, $CI = [0.154, 0.175]$, $p < 0.001$). The RT difference between (a) correct and incorrect trials was larger for strong than weak pairings (Strength $\times$ Accuracy: $\beta = -0.167$, $CI = [-0.194, -0.140]$, $p < 0.001$); (b) match and mismatch trials was larger for strong compared to weak pairings (Strength $\times$ Mismatch: $\beta = -0.126$, $CI = [-0.141, -0.111]$, $p < 0.001$); and (c) correct and incorrect trials was larger for match than mismatch trials (Accuracy $\times$ Mismatch: $\beta = -0.152$, $CI = [-0.189, -0.115]$, $p < 0.001$).

The Strength $\times$ Accuracy $\times$ Mismatch interaction also was significant ($\beta = 0.140$, $CI = [0.096, 0.184]$, $p < 0.001$). Collectively, these outcomes indicate that: participants followed task instructions, with retrieval of the cued associate tending to begin prior to probe onset, thus establishing conditions for a MPE on mismatch trials; memory was superior following repeated learning (higher d' and faster RTs in the strong condition); responses were slower when the probe was a mismatch; and strong MPEs had a greater impact on RTs relative to weak MPEs.

Strong MPEs signal a need for cognitive control

Frontal theta is posited to mark the engagement of cognitive control and support goal-directed behavior⁴⁹. We examined how MPEs impact frontal theta power, leveraging trial-level regression models to examine the effects of match vs. mismatch probes and condition-level differences between strong vs. weak predictions. We focus here on temporal clusters where there were effects of Strength, Mismatch, and/or an interaction that emerged after probe onset. Because some participants had few to no error trials, models only included correct trials (for models including incorrect trials, see Supplementary Fig. 3 and Supplementary Tables 1-3 [↗](#)).

Post-probe onset, there was a main effect of Mismatch (0.032 to 0.704s; mean $\beta = 0.35$, 95% $CI = [0.26, 0.44]$; $p < 0.001$; Fig. 1d [↗](#)), a main effect of Strength (0.504 to 0.872s; $\beta = 0.14$, $CI = [0.07, 0.21]$; $p < 0.001$) and a Mismatch $\times$ Strength interaction (0.048 to 0.728s; $\beta = -0.35$, $CI = [-0.46, -0.24]$; $p < 0.001$; all p -values for temporal clusters are permutation-corrected, unless indicated otherwise). The interaction reflects a greater difference in frontal theta power between mismatch and match trials on strong compared to weak trials (Mismatch effect size, strong: Cohen's $d = 1.28$; weak: $d = 0.85$). These data suggest that strong MPEs elicit a greater increase in cognitive control relative to weak MPEs, with these frontal theta power changes primarily preceding responses (Supplementary Fig. 4a [↗](#)).

After probe offset (i.e., during the ITI), there was a main effect of Mismatch (0.394 to 0.698s; $\beta = -0.07$, $CI = [-0.12, -0.02]$; $p < 0.001$), a main effect of Strength (0.466 to 0.626s; $\beta = -0.06$, $CI = [-0.12, 0.00]$; $p < 0.001$), and a Mismatch $\times$ Strength interaction (0.194 to 0.738s; $\beta = 0.11$, $CI = [0.05, 0.17]$;

	% Hits (SD)	% False alarms (SD)
Associative retrieval performance		
Strong (4x)	94.1 (9.9)	11.1 (13.5)
Weak (1x)	71.0 (18.8)	17.0 (16.2)
Recognition task performance		
Experiment 1		
Strong (4x)	75.6 (16.4)	
Weak (1x)	76.7 (18.3)	
Familiar item	43.9 (16.7)	
Novel		15.7 (9.3)
Experiment 2		
Strong (4x)	46.4 (18.7)	
Weak (1x)	48.6 (20.6)	
Familiar item	27.6 (15.1)	
Novel		14.2 (8.7)

Table 1. Hit and false alarm rates

$p < 0.001$); the signs of all three effects were inverted relative to the post-probe onset effects. Indeed, the interaction indicates that the difference in the magnitude of the post-probe offset dip in frontal theta between match and mismatch trials was larger for strong compared to weak trials.

Strong MPEs increase attentional engagement

Attentional engagement can be indexed by multiple physiological measures, including a decrease in posterior alpha power²¹. After probe onset, there was a main effect of Mismatch (0.512 to 0.816s; $\beta = -0.13$, CI=[-0.18, -0.08]; $p < 0.001$; Fig. 1e), a main effect of Strength (0.504 to 1.480s; $\beta = -0.18$, CI=[-0.22, -0.14]; $p < 0.001$), and a Mismatch $\times$ Strength interaction (0.536 to 1.376s; $\beta = 0.15$, CI=[0.10, 0.20]; $p < 0.001$). The interaction indicates that, relative to weak MPEs, strong MPEs evoked greater posterior alpha suppression, putatively marking greater attentional engagement. Moreover, weak matches evoked greater alpha suppression relative to strong matches. Response-locked analyses showed that qualitatively, responses coincided with a drop in posterior alpha, that main effects of Mismatch on posterior alpha were evident prior to responses, and that strength-modulated MPE effects in posterior alpha emerged primarily after responses (Supplementary Fig. 4b).

Attentional engagement and arousal also can be indexed by analyses of pupil size. Across 94 participants with suitable pupil data and sufficient trial counts per condition, we found that relative to probe onset, there was a main effect of Mismatch (-0.14 to 2.00s; $\beta = 0.16$, CI=[0.09, 0.23]; $p < 0.001$; Fig. 1f), a main effect of Strength (1.29 to 2.00s; $\beta = 0.22$, CI=[0.15, 0.29]; $p < 0.001$), and a Mismatch $\times$ Strength interaction (0.52 to 2.00s; $\beta = -0.32$, CI=[-0.40, -0.24]; $p < 0.001$) (see Supplementary Fig. 5 for data separated by experiment). Complementing the posterior alpha findings, the interaction indicates that the increase in pupil diameter was larger for strong compared to weak MPEs. This strength-modulated MPE effect specifically emerged after responses (Supplementary Fig. 4c).

Distinct pupil features reflect mnemonic prediction errors and attentional orienting

We leveraged temporal PCA over pupil data ($N=96$) to gain further insights into the sequence of cognitive operations that were engaged over the course of the associative retrieval trials and in response to MPEs. As detailed in corresponding Results sections below, the temporal profile of each of six components (Fig. 2a) reveal relationships with trial structure, condition, and retrieval behavior, and suggest that the analysis isolated: (a) the effects of pupil-linked arousal at the time of cue presentation on subsequent associative retrieval success (PC2); (b) responses to the strength of the cue (PC2 and PC5); (c) engagement of cognitive effort in support of controlled retrieval (PC5); (d) memory retrieval strength (PC1); (e) responses to probe onset and detection of mismatch stimuli (PC6); (f) responses to MPEs (PC3 and PC4); and (g) post-error orienting responses (PC4). In the following sections, we concentrate on and further delineate specific pupil components that most closely relate to our core hypothesis (for discussion of PC2 and PC6, see Supplementary Fig. 6-7 and Supplementary Tables 4-8).

By relating PC scores to behavior and experimental conditions (memory strength and match/mismatch), we identified two PCs (PC3 and PC4) that were sensitive to MPEs (Fig. 2b and 2c). PC3, accounting for 19.29% of variance, peaked 0.99s after probe onset (Fig. 2b; 95% CI = [0.97, 1.01]; loading = 0.88, 95% CI=[0.88, 0.90]) and had a full width at half maximum (FWHM) of 1.18s (CI=[1.16, 1.21]). PC4, accounting for 14.68% of variance, peaked towards the end of the trial (Fig. 2c; 1.82s after probe onset, CI=[1.81, 1.83]; loading = 0.93, CI=[0.92, 0.93]). To interpret these components and test their sensitivity to MPEs, we examined how the scores for each PC related to prediction strength and match/mismatch conditions; we further assessed how RTs related to PC scores and experimental conditions.

For PC3, there was a Mismatch $\times$ Strength interaction (Table 2), indicating a strength-sensitive MPE pupil response. Pairwise comparisons showed higher PC3 scores for strong mismatches (correct rejections) than strong matches (hits) ($\Delta_{\text{Match-Mismatch}} = -0.079$, CI=[-0.134, -0.024], $t(148.072) = -2.822$, $p = 0.005$); for weak pairings, PC3 scores were higher for matches than mismatches ($\Delta_{\text{Match-Mismatch}} = 0.093$, CI=[0.034, 0.153], $t(197.646) = 3.091$, $p = 0.002$). Using data from

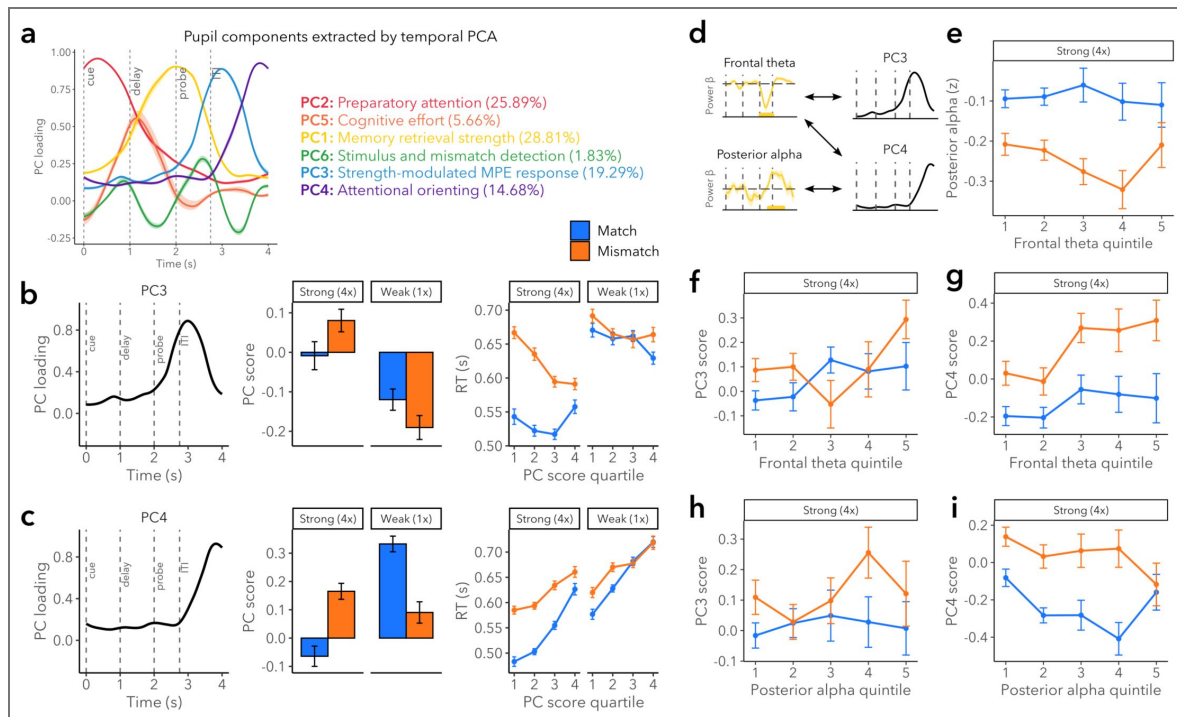


Fig. 2. Pupil components sensitive to MPEs and their relationships with frontal theta and posterior alpha.

(a) Pupil components extracted by temporal PCA with bootstrapped 95% confidence intervals. PC scores for PC3 (b) and PC4 (c) as a function of match/mismatch conditions and memory strength (middle) and as related to reaction times (right). (d) Schematic of observed relationships among frontal theta, posterior alpha, PC3 scores, and PC4 scores. Time windows over which frontal theta and posterior alpha were averaged in trial-level analyses are underlined in yellow and correspond to those displayed in Fig. 1. (e) Relationship between frontal theta cognitive control signals and attention signals in posterior alpha. (f and g) Relationships between frontal theta cognitive control signals and PC3 (f) and PC4 (g). (h and i) Relationships between posterior alpha MPE-related attention signals and PC3 (h) and PC4 (i). Quartiles and quintiles are shown for visualization purposes only; data were analyzed using continuous measures as described in the Methods.

all trials, including misses and false alarms, tPCA extracted highly similar components ($r(399) > 0.99$, $p < 0.001$ for all six PCs); using these component scores, we performed a model comparison to examine whether Accuracy explained a significant amount of variance in PC3 scores. Indeed, an augmented model including effects of and interactions with Accuracy (in addition to Mismatch and Strength) fit PC3 scores better than a model that only included effects of and an interaction between Mismatch and Strength ($\chi^2(8) = 63.90$, $p < 0.001$; Table 3). The enhanced fit of the augmented model further supports the interpretation that PC3 reflects a strength-sensitive MPE pupil response.

Altogether, the RT findings suggest that stronger MPEs (i.e., faster mismatch RTs) evoked larger pupillary responses. This was manifest in there being a Mismatch $\times$ Strength $\times$ PC3 score interaction (Table 4), with post-hoc analyses showing that higher PC3 scores for strong mismatches were coupled with faster responses (estimated slope = -0.046, CI = [-0.061, -0.031], $t(306.238) = -6.219$, $p < 0.001$), whereas RTs and PC3 scores were unrelated for strong matches (estimated slope = 0.009, CI = [-0.003, 0.021], $t(152.586) = 1.463$, $p = 0.145$); the difference between these slopes was significant (match – mismatch slope = 0.055, CI = [0.038, 0.072], $t(4019.950) = 8.196$, $p < 0.001$). Furthermore, RTs and PC3 scores were more strongly related for strong compared to weak mismatches (strong – weak slope = -0.032, CI = [-0.053, -0.012], $t(5920.054) = -4.134$, $p < 0.001$). Together, these results further support the interpretation that PC3 is a strength-sensitive MPE component that is sensitive to differences in prediction strength at both the condition-level (strong vs. weak) and trial-level (e.g., trial-level strong mismatch RTs).

A different pattern was observed for weak trials, with only a trend-level relationship between PC3 scores and RTs for mismatches (estimated slope = -0.014, CI = [-0.029, 0.001], $t(341.456) = -1.777$, $p = 0.077$), a negative relationship for matches (estimated slope = -0.022, CI = [-0.035, -0.010], $t(188.098) = -3.450$, $p = 0.001$), and no significant difference between these slopes (match – mismatch slope = -0.009, CI = [-0.027, 0.010], $t(4473.238) = -1.243$, $p = 0.599$). These outcomes are consistent with the behavioral findings that MPEs were experienced primarily on strong but not weak trials (Fig. 1). In addition, these findings indicate that within the strong condition, when stronger mnemonic predictions (as indexed by shorter strong mismatch RTs) were violated, larger pupillary responses were evoked.

Similar to PC3, PC4 responses displayed strength-sensitive MPE modulation (Fig 2c; Table 5). A Mismatch $\times$ Strength interaction revealed a larger difference between mismatch and match scores in the strong than the weak condition; strong mismatch scores were higher than strong match scores ($\Delta_{Match-Mismatch} = -0.275$, CI = [-0.327, -0.224], $t(10260.360) = -10.486$, $p < 0.001$) and weak mismatch scores were lower than weak match scores ($\Delta_{Match-Mismatch} = 0.202$, CI = [0.145, 0.258], $t(10269.580) = 6.947$, $p < 0.001$). PC4 scores from the dataset including error trials were better fit by a model that additionally included effects of and interactions with Accuracy ($\chi^2(8) = 213.83$, $p < 0.001$); moreover, component scores were highest for strong error trials (Table 6; Accuracy effect; Accuracy $\times$ Strength).

The RT model showed that weaker mnemonic predictions (or greater difficulty in correctly retrieving memories), as indexed by longer RTs, were associated with higher PC4 scores (Table 7; effect of PC4 score). The relationship between PC4 scores and RTs was stronger for weak compared to strong trials (PC4 score $\times$ Strength) and for match compared to mismatch trials (PC4 score $\times$ Mismatch). There was no three-way interaction (PC4 score $\times$ Mismatch $\times$ Strength). As such, more effortful memory retrieval, manifesting as slower RTs and/or erroneous responses, elicited larger delayed pupillary responses.

Together, these results suggest that PC4, which peaked after PC3 and after responses, reflects post-MPE cognitive processing and perhaps a change in attentional engagement. PC4 appears to primarily reflect an orienting response to errors – not only strong erroneous predictions, but also erroneous memory decisions. This component is sensitive to the amount of effort (assayed by RTs) exerted to generate those predictions and the strength of those errors (strong vs. weak). Elevated PC4 scores on weak hits, where no error was made, may instead reflect retrieval practice and thus internally oriented attention. On these trials, when cue-elicited retrieval may have been weaker, the probe may have provided additional support for pattern completion of the learning episode

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	0.002	[-0.119, 0.123]	0.979	0.061	90.970
Mismatch	0.079	[0.024, 0.134]	0.005	0.028	148.072
Strength	-0.113	[-0.177, -0.050]	0.001	0.032	114.654
Mismatch x Strength	-0.172	[-0.240, -0.104]	0.000	0.035	10236.929

Table 2. PC3 scores as a function of Mismatch and Strength

Table 3. PC3 scores as a function of Mismatch, Strength, and Memory Accuracy

Fixed effects	Estimate		CI	p-value	SE	DF
(Intercept)	0.034	[-0.087, 0.155]	0.577	0.061	92.514	
Mismatch	0.079	[0.027, 0.130]	0.003	0.026	215.623	
Accuracy	0.021	[-0.104, 0.146]	0.743	0.064	655.824	
Strength	-0.107	[-0.168, -0.046]	0.001	0.031	123.938	
Mismatch x Accuracy	-0.418	[-0.601, -0.235]	0.000	0.093	3369.869	
Mismatch x Strength	-0.176	[-0.244, -0.107]	0.000	0.035	11969.179	
Accuracy x Strength	-0.169	[-0.304, -0.035]	0.013	0.068	3106.734	
Mismatch x Accuracy x Strength	0.490	[0.275, 0.706]	0.000	0.110	7003.623	

Table 4. PC3 scores and RTs

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	-0.660	[-0.805, -0.516]	0.006	0.022	1.427
PC3 score	0.009	[-0.003, 0.021]	0.145	0.006	152.586
Strength	0.202	[0.187, 0.217]	0.000	0.008	105.059
Mismatch	0.161	[0.144, 0.178]	0.000	0.009	136.388
PC3 score x Strength	-0.031	[-0.043, -0.020]	0.000	0.006	3083.608
PC3 score x Mismatch	-0.055	[-0.068, -0.042]	0.000	0.007	4019.950
Strength x Mismatch	-0.128	[-0.146, -0.109]	0.000	0.009	10186.180
PC3 score x Strength x Mismatch	0.064	[0.046, 0.082]	0.000	0.009	10234.438

Table 5. PC4 scores as a function of Mismatch and Strength

Fixed effects	Estimate		CI	p-value	SE	DF
(Intercept)	-0.115	[-0.203, -0.027]		0.011	0.044	103.597
Mismatch	0.275	[0.224, 0.327]		0.000	0.026	10260.365
Strength	0.422	[0.379, 0.465]		0.000	0.022	10262.071
Mismatch x Strength	-0.477	[-0.553, -0.400]		0.000	0.039	10257.275

Table 6. PC4 scores as a function of Mismatch, Strength, and Memory Accuracy

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	-0.137	[-0.235, -0.040]	0.006	0.049	91.302
Mismatch	0.252	[0.184, 0.319]	0.000	0.034	152.507
Accuracy	0.747	[0.589, 0.906]	0.000	0.081	457.972
Strength	0.397	[0.335, 0.460]	0.000	0.032	132.103
Mismatch x Accuracy	-0.488	[-0.692, -0.284]	0.000	0.104	4905.789
Mismatch x Strength	-0.465	[-0.540, -0.390]	0.000	0.038	11923.486
Accuracy x Strength	-0.679	[-0.827, -0.531]	0.000	0.075	3826.939
Mismatch x Accuracy x Strength	0.538	[0.300, 0.777]	0.000	0.122	7308.225

Table 7. PC4 scores and RTs

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	-0.649	[-0.847, -0.450]	0.012	0.024	1.227
PC4 score	0.081	[0.071, 0.091]	0.000	0.005	154.422
Strength	0.170	[0.155, 0.184]	0.000	0.007	115.840
Mismatch	0.140	[0.124, 0.156]	0.000	0.008	142.039
PC4 score x Strength	0.012	[0.001, 0.022]	0.025	0.005	7330.103
PC4 score x Mismatch	-0.028	[-0.041, -0.016]	0.000	0.006	8214.680
Strength x Mismatch	-0.086	[-0.104, -0.069]	0.000	0.009	10176.523
PC4 score x Strength x Mismatch	-0.013	[-0.031, 0.005]	0.169	0.009	10157.134

for that association, and this engagement in memory retrieval may have elicited pupil dilation (c.f., Strength effect in Fig. 1f [↗](#)). Overall, PC4 may therefore reflect an attentional orienting response that, depending on the relative success of memory retrieval and probe identity, may direct attention internally or externally.

PC3 and PC4 scores were differentially associated with RTs, with PC3 showing an inverse association during strong MPEs and PC4 showing positive associations in all conditions. This differential relationship suggests for stronger MPEs (i.e., strong MPEs that were responded to quickly), higher PC3 scores may reflect more rapid resolution of conflict and lower PC4 scores may therefore reflect less attentional orienting. In other cases, when PC4 scores are high, conflict resolution may have taken longer and subsequent orienting responses may not resolve before the onset of the subsequent trial. As such, PC3 and PC4 may have differential effects on cognition and behavioral performance on subsequent trials.

Summary

Following probe onset, detection of a mismatch may elicit engagement of cognitive control (i.e., an increase in frontal theta) to facilitate comparison of mnemonic and sensory evidence. Sustained attention to mnemonic evidence and evidence accumulation during the comparison process may drive an anticipatory pupil response up until the decision; the amplitude of this response may ultimately be impacted by the magnitude of the MPE (PC3). Upon completion of the trial, after a response is made, additional cognitive processes may lead to subsequent changes in pupil size (PC4) that reflect additional changes in attention/arousal.

We next investigated whether, and if so, how control processes interacted with attention/arousal by examining relationships between MPE-related frontal theta, posterior alpha, and pupil responses.

Frontal theta predicts pupillary responses, but not posterior alpha

Altogether, our behavioral, electrophysiological, and pupillary findings provide converging evidence that stronger MPEs were experienced in the strong condition relative to the weak condition. Accordingly, we examined the relationship between MPE-related impacts on frontal theta, posterior alpha, and pupil, focusing on strong trials and the effects of matches and mismatches.

Changes in attention and arousal following MPEs may be elicited by error detection alone or by a separate signal that more control is needed. To investigate this question, we examined whether putative cognitive control signals, assayed by increased frontal theta power, predicted changes in posterior alpha and pupil size (Fig. 2d [↗](#)). First, we observed no relationship between MPE-evoked frontal theta (second Mismatch $\times$ Strength cluster) and posterior alpha (third Mismatch $\times$ Strength cluster) (Fig. 2e [↗](#); main effect of frontal theta: $\beta=0.010$, CI=[-0.012, 0.032], $p=0.369$; frontal theta $\times$ Mismatch interaction: $\beta=-0.023$, CI=[-0.056, 0.011], $p=0.182$). Second, while higher frontal theta on strong mismatches was associated with higher decision-related pupil components — both immediate (Fig. 2f [↗](#); PC3: main effect of frontal theta: $\beta=0.050$, CI=[0.008, 0.092], $p=0.020$) and lagged (Fig. 2g [↗](#); PC4: main effect of frontal theta: $\beta=0.054$, CI=[0.005, 0.104], $p=0.033$) — the strength of this coupling did not differ between matches and mismatches for PC3 (frontal theta $\times$ Mismatch interaction for PC3: $\beta=0.001$, CI=[-0.056, 0.059], $p=0.966$) or PC4 (frontal theta $\times$ Mismatch interaction for PC4: $\beta=0.024$, CI=[-0.055, 0.103], $p=0.544$). Finally, frontal theta predicted mean pupil diameter (Mismatch $\times$ Strength cluster) (Supplementary Fig. 8a [↗](#); main effect of frontal theta: $\beta=0.046$, CI=[0.007, 0.084], $p=0.022$), with no difference in the strength of this relationship between mismatch and match trials (frontal theta $\times$ Mismatch interaction ($\beta=0.038$, CI=[-0.021, 0.097], $p=0.212$)).

Together, these results indicate that the need for cognitive control, as assayed by frontal theta, was associated with immediate and lagged pupil-linked increases in attention and arousal (i.e., PC3, PC4, and mean pupil diameter). Although frontal theta was higher for strong mismatches than strong matches (Fig. 2d [↗](#)), the strength of the relationship between frontal theta and pupillary responses did not differ as a function of MPE detection. Increases in attention as assayed by

posterior alpha were independent of frontal theta control signals. Thus, these outcomes suggest that some changes in attention following MPEs are elicited by error detection alone (i.e., posterior alpha), whereas other changes in attention/arousal (i.e., pupil) are associated with control signals.

Posterior alpha couples with lagged pupillary orienting responses, but not immediate decision-related pupil responses

We next examined whether MPE-driven changes in posterior alpha were associated with pupil-linked attention and arousal responses. There was no relationship between posterior alpha power and PC3 scores (Fig. 2h [↗](#); main effect of posterior alpha: $\beta = -0.005$, $CI = [-0.077, 0.067]$, $p = 0.894$; PC3 score $\times$ Mismatch: $\beta = 0.017$, $CI = [-0.108, 0.141]$, $p = 0.793$). By contrast, PC4 scores and mean pupil diameter were negatively coupled with posterior alpha power (Fig. 2i [↗](#); main effect of posterior alpha on PC4 scores: $\beta = -0.155$, $CI = [-0.229, -0.080]$, $p < 0.001$; Supplementary Fig. 8b [↗](#); main effect of posterior alpha on mean pupil: $\beta = -0.107$, $CI = [-0.182, -0.032]$, $p = 0.007$), indicating that greater posterior alpha suppression (a marker of attentional engagement) predicted larger pupillary orienting responses (a marker of increased arousal and/or attention). There was no PC4 score $\times$ Mismatch interaction ($\beta = -0.053$, $CI = [-0.192, 0.086]$, $p = 0.452$) nor was there a mean pupil $\times$ Mismatch interaction ($\beta = 0.028$, $CI = [-0.112, 0.168]$, $p = 0.687$). As such, greater attentional engagement following memory decisions, as assayed by decreases in posterior alpha, was followed by stronger pupillary orienting responses.

Together, these findings suggest that the magnitude of alpha suppression following memory decisions is associated with lagged orienting pupillary responses, but not immediate decision-related pupillary responses. In other words, the initial MPE-evoked changes in posterior alpha and pupil size (PC3) are independent of each other.

Strong MPE-evoked increases in cognitive control do not explain increases in attention and arousal

While the preceding outcomes suggest that attention, arousal, and control processes are triggered at least partially independently by MPEs, we conducted multiple mediation analyses to further test whether the dynamics of observed cognitive and neural changes following strong MPEs are consistent with or challenge our hypothesis that changes in cognitive control partially drive increases in attention and arousal. In these analyses, we focused on the dynamics in response to strong MPEs (strong mismatch trials) and used RTs as a proxy measure for the trial-level magnitude of the MPE (with shorter RTs reflecting stronger predictions and thus stronger MPEs). Pupil and posterior alpha effects were modeled separately; in each model, we tested whether the trial-level relationships between MPE magnitude (RTs) and changes in attention and arousal (pupil PC3/posterior alpha) were mediated by increases in cognitive control (frontal theta) (Fig. 3a [↗](#)). In the pupil model (Fig. 3b [↗](#)), the credible interval for the effect of RTs on frontal theta included zero, providing no strong evidence for this relationship ($a = -0.052$, 95% credible interval (CrI) = $[-0.137, 0.034]$). There was strong evidence for an effect of frontal theta on PC3 scores ($b = 0.070$, $CrI = [0.002, 0.138]$) and insufficient evidence for frontal theta mediating the relationship between RTs and immediate pupil responses ($a*b = 0.005$, $CrI = [-0.010, 0.027]$). There was strong evidence for a direct effect of RTs on PC3 scores ($c' = -0.214$, $CrI = [-0.295, -0.133]$). In the posterior alpha model (Fig. 3c [↗](#)), there was no strong evidence for an effect of RTs on frontal theta ($a = -0.052$, $CrI = [-0.138, 0.033]$), as in the pupil model. There was strong evidence that higher frontal theta predicted lower posterior alpha ($b = -0.082$, $CrI = [-0.146, -0.018]$) and insufficient evidence for frontal theta mediating the relationship between RTs and posterior alpha ($a*b = 0.009$, $CrI = [-0.004, 0.029]$). There was strong evidence for RTs negatively predicting posterior alpha ($c' = -0.124$, $CrI = [-0.189, -0.058]$). Together, these outcomes suggest that increases in attention and arousal following strong MPEs are not well-explained by increases in cognitive control. Yet, given that these exploratory analyses may be underpowered, caution is warranted in interpreting these null findings.

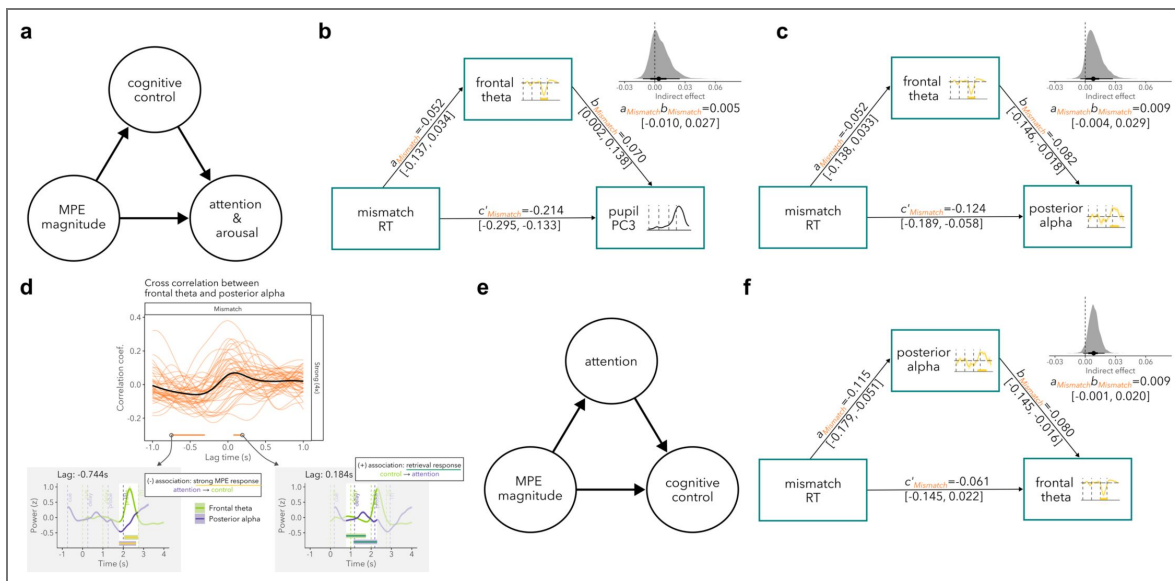


Figure 3. Dynamics between cognitive control, attention, and arousal in response to strong MPEs.

(a) Schematic of hypothesized relationships: stronger MPEs increase attention/arousal and this effect was hypothesized to be partially explained by an increase in cognitive control. (b) A multivariate mediation analysis testing the model diagrammed in (a) with pupil PC3 as the measure of attention/arousal. (c) A multivariate mediation analysis testing the model diagrammed in (a) with posterior alpha as the measure of attention. (d) Cross-correlation analysis showing the correlation between frontal theta time series and lagged posterior alpha time series on strong MPE trials. Gray subplots illustrate the shifted time series from Fig. 1d and Fig. 1e at example lags (circled in gray) for a negative lag timepoint (left) and positive lag timepoint (right). (e) An alternative model: changes in cognitive control following MPEs may be explained by changes in attention, as assayed by posterior alpha. (f) A multivariate mediation analysis testing the model in (e), examining whether posterior alpha mediates the relationship between RTs and frontal theta. Paths in mediation models are labeled with the mean of the posterior and 95% credible intervals. Gray histograms accompanying mediation model diagrams depict the posterior distribution for the indirect effect; black points indicate posterior medians and horizontal lines indicate the 66% and 95% highest density intervals.

Posterior alpha responses to strong MPEs precede increases in frontal theta

To develop further insights into the temporal relationships among the neurocognitive measures, we explored their temporal relationships using cross-correlational analyses with a lag window of 1s. For frontal theta and posterior alpha, there were two significant clusters with dissociable lag patterns. Interpretation of these clusters was guided by the direction of the relationship between frontal theta and posterior alpha and how this direction compared with prior analyses; after identifying candidate cognitive processes associated with a given lag cluster, we then used the lag time to make inferences about which cognitive process was leading or lagging.

There was a cluster with a negative mean correlation and with posterior alpha leading (-0.744 to -0.312 s, mean $r = -0.053$, $CI = [-0.055, 0.052]$). The negative relationship between frontal theta and posterior alpha – greater control and more attention – may primarily reflect cognition in response to the strong MPE, given the differential sign of the primary Mismatch $\times$ Strength interactions in Fig. 1d and Fig. 1e (n.b., at the trial-level in Fig. 2e, there was no Mismatch $\times$ frontal theta interaction on posterior alpha and no main effect of frontal theta on posterior alpha; however, the posterior alpha mediation model (Fig. 3c), wherein the predictors were additionally rescaled, suggests that frontal theta and posterior alpha may be negatively associated at a trial-level following a strong MPE.). The negative lag suggests that an increase in attention, as assayed by posterior alpha, may precede an increase in cognitive control as indexed by frontal theta. This interpretation is consistent with the outcomes in the response-locked trial-level regression models (Supplementary Fig. 4a and Supplementary Fig. 4b), which showed that qualitatively, main effects of Mismatch emerged for posterior alpha before main effects of Mismatch emerged for frontal theta. While suggestive, future work should more directly examine the possibility that posterior alpha-related changes in attention precede and perhaps drive changes in cognitive control as assayed by frontal theta. The second significant cluster is discussed in the next section on *Controlled retrieval and memory retrieval strength*. Cross-correlation analyses of pupil with frontal theta and pupil with posterior alpha did not reveal significant correlation clusters.

Exploratory analyses examining whether changes in posterior alpha explain increases in frontal theta

Given the finding that posterior alpha responses may precede frontal theta responses to strong MPEs, we tested an alternative mediation model in an exploratory analysis to assess the possibility that changes in cognitive control are partially explained by changes in attention as indexed by posterior alpha (Fig. 3e). There was strong evidence for an effect of RTs on posterior alpha (Fig. 3f; $a = -0.115$, $CrI = [-0.179, -0.051]$) and strong evidence for an effect of posterior alpha on frontal theta ($b = -0.080$, $CrI = [-0.145, -0.016]$). There was insufficient evidence for an indirect effect ($a*b = 0.009$, $CrI = [-0.001, 0.020]$). There was also insufficient evidence for a direct effect of RTs on frontal theta ($c' = -0.061$, $CrI = [-0.145, 0.022]$), as in the earlier mediation models (Fig. 3b-c). Bayesian leave-one-out cross-validation showed that the difference in model fit for alpha as a mediator (Fig. 3c; $ELPD = -2842.07$) and theta as a mediator (Fig. 3f) was minimal ($\Delta ELPD = -0.82 \pm 0.41$). As such, the current findings do not provide conclusive evidence for or against the idea that control signals modulate attention, or vice versa.

Summary

In sum, strong MPEs evoked changes in frontal theta (Fig. 1d), posterior alpha (Fig. 1e), and pupil diameter (Fig. 1f). While the changes in frontal theta and posterior alpha were not significantly associated with each other following strong matches and showed no interaction with Mismatch, there was suggestive evidence from the mediation models that they are negatively coupled in response to strong MPEs – with greater control predicting more attention. Changes in frontal theta and posterior alpha were both associated with the magnitude of late pupil responses (PC4). Initial decision-related pupil responses (PC3) were associated with frontal theta but not posterior alpha. These relationships did not differ between Match and Mismatch, suggesting that coupled changes in attention/arousal and control during/following memory decisions are not specific to MPEs. Differences in the patterns of relationships for pupil components and mean pupil diameter highlight the increased sensitivity offered by temporal PCA in identifying pupil

responses to varying cognitive processes engaged at different timepoints. Critically, these outcomes highlight the distinct temporal profiles of the relationship between control and attention/arousal: pupillary assays of increased attention/arousal evoked by memory decisions are associated with cognitive control signals (i.e., PC3, PC4, mean pupil), whereas increases in attention as assayed by posterior alpha only relate to later pupil responses (i.e., PC4 but not PC3 or mean pupil). Moreover, cross-correlation analyses suggested that upregulation of attention, as assayed by posterior alpha, precedes changes in cognitive control. Although this outcome suggests that posterior alpha changes may partially explain changes in frontal theta, mediation analyses did not provide strong evidence for control signals upregulating attention or attention increases upregulating control signals.

2. Controlled retrieval and memory retrieval strength

MPEs reflect the mismatch between a cue-evoked retrieved memory (i.e., a mnemonic prediction) and the future state of the world (i.e. the probe). We next examined how cue-evoked electrophysiological and pupillary markers of processes that precede MPEs — namely cognitive control and memory retrieval strength — relate to MPE responses. First, we examined how cue-evoked differences between strong and weak trials (which presumably reflect, at least in part, differences in the strength of retrieved memories) manifest in frontal theta, posterior alpha, and pupillary signals prior to the onset of the match/mismatch retrieval probe. Next, we identified trial-level indices of pupil-linked cognitive effort and memory retrieval signals. Finally, we examined whether trial-by-trial variability in cognitive control and retrieval strength interact with MPE responses.

Memory retrieval signals in frontal theta, posterior alpha, and pupil

To identify temporal clusters reflecting engagement of cognitive control during memory retrieval and memory retrieval strength, we examined cue-evoked and delay period main effects of Strength. We observed three key findings. First, frontal theta varied by memory strength (main effect of Strength from 0.768 to 1.744s after cue onset; $\beta=0.07$, $CI=[0.03, 0.12]$; $p<0.001$; Fig. 1d), with frontal theta power being higher on weak compared to strong trials. This difference may reflect increased engagement of cognitive control during attempts to retrieve weak associations. Second, for posterior alpha, there were main effects of Strength from 0.128 to 0.232s ($\beta=0.07$, $CI=[0.01, 0.12]$), 0.952 to 2.112s ($\beta=0.11$, $CI=[0.07, 0.15]$); Fig. 1e) after cue onset. For both clusters, posterior alpha power was lower on strong than weak trials, with this cue-evoked alpha suppression possibly reflecting attention to strongly familiar retrieval cues and/or to retrieved contents. Third, pupil diameter paralleled the findings for posterior alpha, showing a main effect of Strength from 0.91 to 2.94s after cue onset ($\beta=-0.12$, $CI=[-0.18, -0.05]$; $p<0.001$; Fig. 1f), with pupil diameter being larger on strong compared to weak trials. As with posterior alpha, this difference could reflect attention to strongly familiar retrieval cues and/or to retrieved contents.

Distinct pupil components reflect cognitive effort and memory retrieval strength

Leveraging temporal PCA, we disentangled cue-evoked pupillary responses and identified two distinct components sensitive to cognitive effort and memory retrieval success.

PC1 scores were higher for stronger memories both at the condition level and the trial level, leading us to interpret PC1 as indicative of the strength of memory retrieval. In particular, PC1 scores were higher for strong compared to weak trials (Table 8; effect of Strength) and higher PC1 scores predicted faster match RTs (Table 9; effect of PC1 score). The relationship between PC1 scores and RTs was attenuated for mismatch trials (Mismatch $\times$ PC1 score interaction), which may be a consequence of the response delay induced by MPEs on mismatch trials (Fig. 1c). PC1 accounted for the most variance (28.81%; Fig. 4a), onset during the cue period, peaked predominantly at the time of probe onset (2.00s, $CI=[1.94, 2.05]$; loading = 0.90, $CI=[0.90, 0.91]$), and had a FWHM of 1.75s ($CI=[1.70, 1.78]$). Altogether, these findings suggest that PC1 tracks graded retrieval success or strength of memory retrieval.

Table 8. PC1 scores as a function of Mismatch and Strength

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	0.038	[-0.036, 0.113]	0.312	0.038	112.302
Mismatch	0.032	[-0.021, 0.085]	0.242	0.027	10269.199
Strength	-0.102	[-0.147, -0.057]	0.000	0.023	10271.457
Mismatch x Strength	0.027	[-0.052, 0.106]	0.506	0.040	10264.248

Table 9. PC1 scores and RTs

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	-0.655	[-0.859, -0.452]	0.012	0.025	1.234
PC1 score	-0.050	[-0.059, -0.040]	0.000	0.005	170.709
Strength	0.201	[0.185, 0.216]	0.000	0.008	104.316
Mismatch	0.159	[0.142, 0.176]	0.000	0.009	136.383
PC1 score x Strength	-0.004	[-0.014, 0.007]	0.479	0.005	9151.269
PC1 score x Mismatch	0.021	[0.008, 0.034]	0.001	0.006	9112.367
Strength x Mismatch	-0.122	[-0.141, -0.104]	0.000	0.009	10189.081
PC1 score x Strength x Mismatch	0.014	[-0.004, 0.033]	0.123	0.009	10218.503

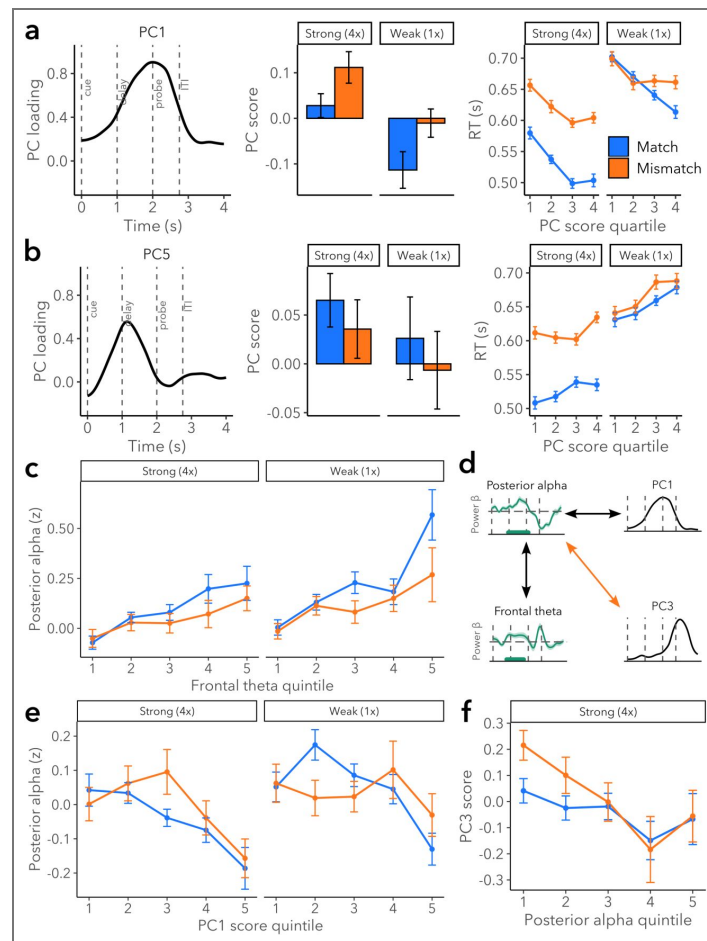


Fig. 4. Pupil components sensitive to memory retrieval strength and cognitive effort and their relationships with frontal theta and posterior alpha.

PC scores for PC1 (a) and PC5 (b) as a function of match/mismatch conditions and memory strength (left) and as related to response times (right). (c) Relationship between frontal theta and posterior alpha during memory retrieval; time windows over which frontal theta and posterior alpha were averaged are underlined in green in (d). (d) Schematic of observed relationships among frontal theta, posterior alpha, PC1 scores, and PC3 scores. Time windows over which frontal theta and posterior alpha were averaged in trial-level analyses are underlined in green and correspond to those displayed in Fig. 1. (e) Relationship between posterior alpha and PC1 during memory retrieval. (f) Relationship between posterior alpha during memory retrieval and PC3 scores. Quintiles and quintiles are shown for visualization purposes only; data were analyzed using continuous measures as described in the Methods.

For PC5, the relationship with the memory strength conditions and RTs indicated that this component was a condition-level marker of cue strength and a trial-level marker of cognitive effort during associative retrieval. PC5 scores were higher for strong than weak trials (Table 10; main effect of Strength), but in contrast to PC1, higher PC5 scores were associated with longer RTs (Table 11; main effect of PC5 scores). The relationship between PC5 scores and RTs was stronger for weak compared to strong pairings (PC5 score $\times$ Strength). Thus, PC5 showed the opposite pattern as PC1, with PC5 scores scaling inversely with trial-level memory strength, suggesting that PC5 reflects a mixture of cue strength (i.e., higher scores for strong cues) and the amount of cognitive effort expended during associative memory retrieval (i.e., inverse relationship with RT). The temporal profile of PC5 is consistent with this interpretation, as PC5 onset during the cue period, peaked during the delay period (1.15s after cue onset, CI=[1.12, 1.18]; loading = 0.56, CI=[0.52, 0.61]; Fig. 4b) and had a FWHM of 0.98s (CI=[0.92, 1.04]). This component accounted for 5.66% of variance.

Posterior alpha (attention) during memory retrieval is coupled with frontal theta (cognitive control) and pupil (attention/arousal)

Having used condition-level memory strength differences to identify temporal clusters associated with associative memory retrieval in frontal theta, posterior alpha, and pupil size (Fig. 1d-f), we then examined whether these electrophysiological and pupillary assays of cognitive control, attention, and arousal during memory retrieval aligned with each other trial-to-trial. In analyses of posterior alpha, we focus on the second Strength cluster (Fig. 1e), as this cluster spanned most of the delay period, when retrieval-related differences were most likely to be present (n.b., mean power for the longest cluster may also have the highest SNR). Similar results were found for the first Strength cluster (Supplementary Fig. 9-10). In addition, since temporal PCA isolated one pupil component related to associative retrieval strength (PC1) and another with cognitive effort during memory retrieval (PC5), pupil analyses focused on these component scores.

Stronger memories may require less cognitive control for bringing memory contents to mind and entail the reinstatement of more content with which to allocate attention. By contrast, weak memories may engage more cognitive control and result in fewer retrieval products to attend to. Consistent with these ideas, the condition-level analyses reported above revealed that strong memories elicited less control (lower frontal theta) and more attention (lower posterior alpha) (Fig. 1d-e). Here, we further tested whether this relationship held on a trial-by-trial basis, with greater control being associated with less attention (higher posterior alpha). Indeed, when analyzing cue-evoked and delay-period signals, there was a main effect of frontal theta on posterior alpha ($\beta=0.132$, CI=[0.104, 0.161], $p<0.001$), indicating that higher frontal theta predicted less posterior alpha suppression. Moreover, a frontal theta $\times$ Strength interaction ($\beta=0.062$, CI=[0.018, 0.106], $p=0.006$; Fig. 4c) revealed that this relationship was stronger for weak compared to strong pairings. There was no frontal theta $\times$ Mismatch interaction ($\beta=-0.031$, CI=[-0.085, 0.023], $p=0.264$) and no frontal theta $\times$ Strength $\times$ Mismatch interaction ($\beta=-0.048$, CI=[-0.122, 0.027], $p=0.212$). Altogether, these results are consistent with frontal theta marking engagement of cognitive control during cued retrieval of weaker memories and that the amount of control engaged during retrieval scales trial-by-trial with less posterior alpha suppression; less alpha suppression may reflect weaker memory reinstatement and fewer retrieval products to allocate attention to.

Posterior alpha and pupil size may both assay attention to memory contents and thus may be inversely related. Indeed, when analyzing cue-evoked and delay-period signals, there were negative relationships between posterior alpha power (second Strength cluster) and both PC1 scores and mean pupil size (main effect of PC1: $\beta=-0.068$, CI=[-0.092, -0.043], $p<0.001$; Fig. 4e; main effect of pupil: $\beta=-0.064$, CI=[-0.091, -0.037], $p<0.001$; Supplementary Fig. 11a). These relationships did not interact with other factors (Strength – PC1 $\times$ Strength interaction: $\beta=0.013$, CI=[-0.023, 0.049], $p=0.472$ or Mismatch – PC1 $\times$ Mismatch interaction: $\beta=-0.012$, CI=[-0.054, 0.030], $p=0.578$; three-way interaction: $\beta=0.033$, CI=[-0.030, 0.096], $p=0.298$; Strength – pupil $\times$ Strength interaction: $\beta=0.000$, CI=[-0.041, 0.041], $p=0.994$; or Mismatch – pupil $\times$ Mismatch interaction: $\beta=-0.028$, CI=[-0.076, 0.020], $p=0.252$; three-way interaction: $\beta=0.039$, CI=[-0.033, 0.112], $p=0.284$).

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	0.081	[-0.353, 0.515]	0.301	0.045	1.137
Mismatch	-0.041	[-0.094, 0.012]	0.127	0.027	10258.839
Strength	-0.090	[-0.134, -0.046]	0.000	0.023	10260.729
Mismatch x Strength	0.045	[-0.033, 0.124]	0.257	0.040	10255.146

Table 10. PC5 scores as a function of Mismatch and Strength

Table 11. PC5 scores and RTs

Fixed effects	Estimate	CI	p-value	SE	DF
(Intercept)	-0.663	[-1.107, -0.219]	0.033	0.036	1.016
PC5 score	0.015	[0.007, 0.022]	0.000	0.004	221.355
Strength	0.215	[0.205, 0.226]	0.000	0.005	10252.177
Mismatch	0.163	[0.151, 0.176]	0.000	0.006	10232.294
PC5 score x Strength	0.010	[-0.001, 0.020]	0.066	0.005	9762.962
PC5 score x Mismatch	-0.008	[-0.021, 0.004]	0.203	0.006	10076.592
Strength x Mismatch	-0.127	[-0.145, -0.108]	0.000	0.009	10232.382
PC5 score x Strength x Mismatch	-0.002	[-0.021, 0.016]	0.796	0.010	10158.572

and there was no relationship between posterior alpha and PC5 (Supplementary Fig. 10 [↗](#)). See Supplementary Fig. 11a [↗](#) for similar results for mean pupil diameter and the second posterior alpha Strength cluster. There were no relationships between frontal theta and PC1 or PC5 scores (Supplementary Fig. 12 [↗](#)). Together, these findings complement the above relationship between cue-evoked and delay-period posterior alpha and frontal theta, indicating that signals of memory retrieval in pupillary responses are associated trial-by-trial with more posterior alpha suppression during memory retrieval; this alpha suppression may reflect attention to and maintenance of retrieved contents.

During memory retrieval, engagement of control precedes changes in attention

The cross-correlation analysis of the frontal theta and posterior alpha time series revealed a significant cluster (a) with a positive correlation and (b) with frontal theta leading (Fig. 3d [↗](#); 0.080 to 0.184s, $r=0.006$, $CI=[0.065, 0.068]$). This positive association – greater control and less attention – may reflect cognition during memory retrieval, given the positive mean β during the delay-period Strength clusters in Fig. 1d [↗](#) and Fig. 1e [↗](#), and the positive trial-level association during those periods (Fig. 4c [↗](#)). The positive lag suggests that at retrieval, cognitive control is first engaged and changes in attention, perhaps toward retrieved contents, follows.

3. MPE responses interact with cognitive processes during mnemonic prediction generation

To understand whether cognitive processes during retrieval (i.e., cognitive control and attention) interact with changes in control, attention, and arousal following MPEs, we investigated whether cue-evoked frontal theta (a marker of cognitive control) predicted subsequent probe-evoked posterior alpha or pupillary MPE responses, and whether posterior alpha or pupil diameter during memory retrieval predicted frontal theta MPE responses.

While there were no relationships between frontal theta during memory retrieval and MPE-specific posterior alpha or pupil responses (Supplementary Fig. 13 [↗](#); no interactions with Mismatch), nor were there effects of posterior alpha or pupil diameter on frontal theta MPE responses (Supplementary Fig. 14a-b [↗](#)), attention during memory retrieval predicted MPE-specific pupil effects. In particular, there was a negative relationship between posterior alpha during memory retrieval (second Strength cluster) and PC3 scores (main effect of posterior alpha: $\beta=-0.097$, $CI=[-0.182, -0.013]$, $p=0.025$; Fig. 4f [↗](#)). This relationship interacted with Mismatch (Mismatch $\times$ posterior alpha interaction: $\beta=-0.111$, $CI=[-0.217, -0.005]$, $p=0.039$), such that the relationship between posterior alpha and PC3 scores was stronger when mnemonic predictions were violated than when they were confirmed. A negative relationship was also observed for posterior alpha and mean pupil size, though no interaction was observed (Supplementary Fig. 14c [↗](#)). By contrast, there was no relationship between posterior alpha and PC4 scores (Supplementary Fig. 14d [↗](#)). These outcomes indicate that greater attentional allocation during memory retrieval, which may occur when memories are stronger and more retrieval products can be reinstated and attended to (Fig. 1e [↗](#); Fig. 4c [↗](#)), predicts the magnitude of immediate pupil responses to MPEs.

4. Learning from MPEs

MPEs similarly enhance recognition of strong and weak mismatches

To investigate whether prediction strength modulates MPE-driven learning, we examined recognition memory for mismatch probes on a subsequent surprise memory test administered after a ~5-min delay (Experiment 1; E1) or a 48-hr delay (Experiment 2; E2). A 48-hr delay was introduced in Experiment 2 to address the possibility of ceiling recognition performance with the ~5-min delay in Experiment 1 and to test whether the effects of prediction strength on learning specifically emerge following overnight consolidation. Contrary to our predictions, subsequent recognition (d') was comparable for strong (E1: 1.81 ± 0.59 ; E2: 1.02 ± 0.52) and weak mismatch probes (E1: 1.86 ± 0.65 ; E2: 1.08 ± 0.60) (E1: $\beta=-0.05$, $CI=[-0.31, 0.21]$, $p=0.89$; E2: $\beta=-0.06$, $CI=[-0.34,$

0.21], $p=0.852$). Recognition of strong (E1: $\beta=0.94$, $CI=[0.68, 1.20]$, $p<0.001$; E2: $\beta=0.52$, $CI=[0.25, 0.80]$, $p<0.001$) and weak (E1: $\beta=0.99$, $CI=[0.73, 1.25]$, $p<0.001$; E2: $\beta=0.58$, $CI=[0.31, 0.86]$, $p<0.001$) mismatch probes was higher than that of familiar items (E1: 0.87 ± 0.48 ; E2: 0.50 ± 0.45), suggesting that MPEs enhanced memory but stronger MPEs do not further boost recognition of prediction-violating items. See Table 1 for hit and false alarm rates and Supplementary Fig. 15 for recognition RTs.

Frontal theta during strong MPEs enhances learning

Finally, to understand whether MPE-driven changes in cognitive control, attention, and/or arousal support learning, we examined whether electrophysiological and pupillary responses to MPEs predict subsequent memory for mismatch probes. In Experiment 1, there was a main effect of mean MPE-related frontal theta (the second Mismatch $\times$ Strength cluster) on subsequent memory (odds ratio (O.R.)=1.186 [1.006, 1.398], $p=0.042$; Fig. 5b) and a significant frontal theta $\times$ Strength interaction (O.R. = 0.745, [0.597, 0.930], $p = 0.009$), indicating that this relationship was attenuated for weak mismatch probes. While frontal theta during strong MPEs partially accounted for recognition memory of strong mismatch probes, frontal theta during weak MPEs did not explain memory for weak mismatch probes. By contrast, MPE-associated changes in posterior alpha and pupil were not associated with subsequent memory (see Supplementary Fig. 16).

Trial-level regression models of frontal theta (Fig. 5c) further revealed main effects of Subsequent Memory from 1.088 to 1.624s ($\beta=-0.15$, $CI=[-0.25, -0.05]$; $p<0.001$), 1.944 to 2.544s ($\beta=-0.24$, $CI=[-0.42, -0.06]$; $p<0.001$), and 2.960 to 3.152s ($\beta=-0.09$, $CI=[-0.15, -0.02]$; $p<0.001$) after cue onset, such that frontal theta was higher for subsequent hits than subsequent misses. There was a significant Subsequent Memory $\times$ Strength interaction from 0.256 to 0.552s after probe onset ($\beta=0.54$, $CI=[0.11, 0.98]$; $p<0.001$) whereby the difference in power between hits and misses was larger for strong compared to weak probes. This analysis revealed that frontal theta partially accounts for subsequent memory regardless of strength condition during select time windows during the delay period and following the probe period. In other words, subsequent memory for strong and weak mismatch probes is partially predicted by the level of cognitive control engaged during mnemonic prediction generation and by increases in cognitive control following a mismatch. Greater frontal theta immediately elicited by strong MPEs during the probe period did not additionally enhance memory for strong mismatch probes above that of weak mismatch probes. Collectively, these findings suggest that cue-evoked and delay-period frontal theta, strength-modulated increases in frontal theta following MPEs, and the magnitude of the post-peak frontal theta responses enhanced learning of probes that violated mnemonic predictions.

Discussion

Features of present experiences often overlap with those of the past. Shared features allow for aspects of current experience to cue memories and enable knowledge about how similar events unfolded in the past to inform predictions about ongoing experience. When ongoing experience unexpectedly diverges from mnemonic predictions, multiple neurocognitive processes may be elicited, fostering adaptation to novel circumstances and facilitating more accurate predictions in the future. Drawing on behavioral, electrophysiological, and pupillary assays, the current study revealed that strong MPEs elicit increases in cognitive control, attention, and arousal, with potential impacts on later cognition and new learning (for a summary, see Fig. 6).

Neurocognitive impacts of MPEs are strength-sensitive

Across three physiological measures – frontal theta, posterior alpha, and pupil size – MPE-related responses were larger when predictions were strong compared to weak. These findings are consistent with the idea that responses to prediction errors are modulated by how *unexpected* the error is^{16,17} and relatedly, how much uncertainty there is when making a decision⁷³. In the current study, weak predictions may give rise to less certainty about an upcoming event (i.e., the probe) such that weak MPEs may be considered an outcome of an *expected* degree of uncertainty following a weak memory cue. On the other hand, more strongly cued predictions – evidenced in

Fig. 5. MPE-driven increases in frontal theta by strong MPEs enhance learning.

(a) Behavioral performance in the surprise recognition task in Experiment 1 and Experiment 2. (b) Higher frontal theta during strong MPEs is associated with a higher likelihood of subsequently remembering the probe ~5 min later (Experiment 1). Points indicate binned values for visualization; smoothed lines reflect model estimates. (c) Trial-level regression models for frontal theta modeling the effects of Subsequent Memory, Strength, and Subsequent Memory $\times$ Strength interactions in Experiment 1.

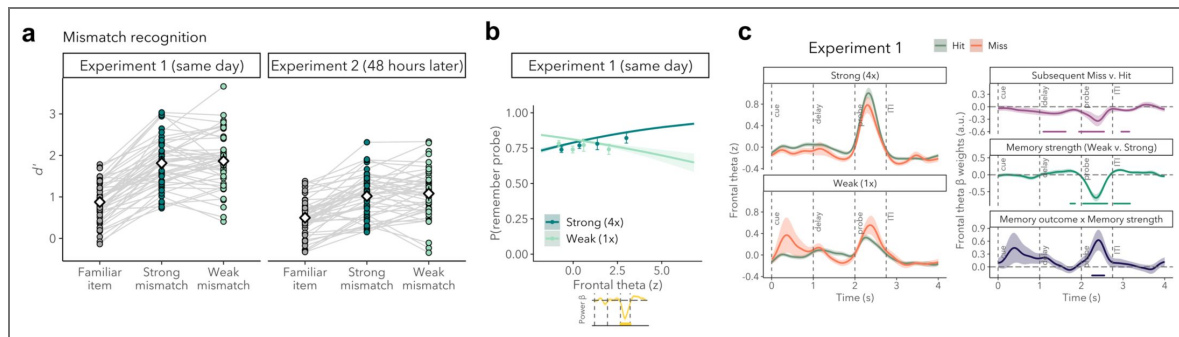
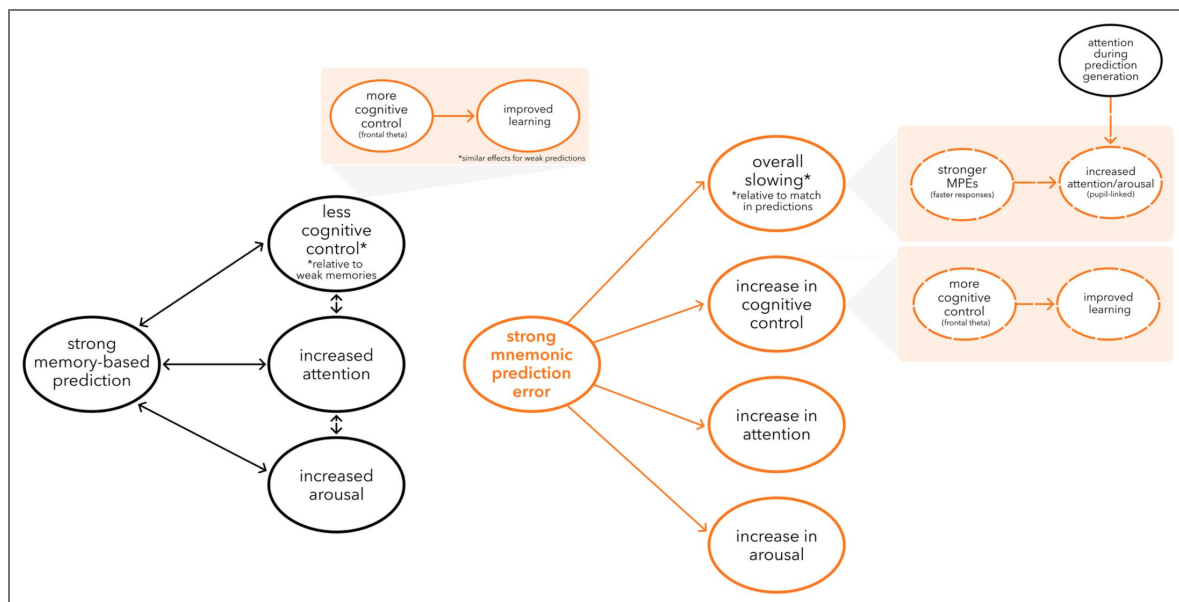


Figure 6. Summary diagram of the observed effects of strong MPEs in contrast to weak MPEs (solid lines in orange), their effects at the trial level (dashed lines in orange), and their trial-level interactions with cognitive processes during mnemonic prediction generation.



the current study by better memory and faster RTs for strong vs. weak pairs – may elicit lower levels of uncertainty; a MPE in this case would be more unexpected and warrant increasing attention to facilitate model updating. Theories of expected and unexpected uncertainty implicate different neuromodulators in tracking different types of uncertainty, with LC norepinephrine release posited to track unexpected uncertainty¹⁷. The present pupil size findings, when interpreted as an indirect assay of noradrenergic LC activity^{14,15}, are consistent with this theory. Mnemonic prediction strength critically modulates demands on control, attention, and arousal.

Attention and control dynamics following strong MPEs

Increases in attention and arousal following strong MPEs, as assayed by posterior alpha and pupil, were not well explained by changes in cognitive control. Exploratory cross-correlation analyses showed that increases in attention in response to strong MPEs, as assayed by posterior alpha, may precede increases in cognitive control. While this result suggests that changes in attention may be elicited first and subsequently upregulate cognitive control, an exploratory formal test of this hypothesis did not provide strong evidence for posterior alpha mediating the effects on frontal theta. Given that these analyses were restricted to strong MPE trials, they may not have been sufficiently powered to detect indirect effects or to detect meaningful differences between these models. Of note, while longer mean RTs for mismatches compared to matches are a behavioral marker of a MPE, within-condition differences in RTs (i.e., between mismatch trials) may reflect more subtle differences in MPE magnitude, with shorter RTs reflecting stronger predictions and thus stronger MPEs. Strong mismatch RTs were used in the mediation models as a proxy measure of MPE magnitude; this estimate may be noisy because RTs in this experiment are likely sensitive to factors independent of mnemonic prediction strength (e.g., preparatory attention (Supplementary Fig. 6 [↗](#)) or memory strength of the mismatch probe). This limitation may have additionally reduced sensitivity to detecting indirect effects. Whether upregulation of attention indeed emerges prior to increases in control, and whether upregulation of attention influences how much control is engaged, remain open questions to be more directly tested in future research.

Memory modes following mnemonic predictions: Open questions and future directions

Notably, for pupil size, the direction of the effect following weak MPEs was the inverse of that observed following strong MPEs, such that there was greater pupil dilation for weak matches than weak mismatches. We propose that this difference may reflect attention being oriented internally to retrieved contents after a weak mnemonic prediction is confirmed, and that increased pupil dilation following weak matches may reflect the probe facilitating pattern completion of the learning episode for the weak association. These findings suggest that the engagement of retrieval^{74,75} or encoding modes^{1,39} following confirmation or violation of mnemonic predictions may be dependent on prediction strength. However, whether increased pupil dilation following the confirmation of weak mnemonic predictions indeed reflects retrieval requires investigation of the mental contents and targets of increased attention following weak matches.

Pupil responses to MPEs: Temporal dynamics and neural interactions

Pupil responses during associative retrieval trials indexed multiple cognitive processes. As observed by others^{7,8}, temporal PCA revealed multiple pupil responses sensitive to task structure. Moreover, we observed strong coupling between PC scores and trial-level behavioral responses, which enabled interpretation of individual components as reflections of distinct cognitive processes. With respect to MPEs, temporal PCA yielded finer-grained insights into the temporal dynamics of pupillary responses. In particular, this analysis isolated two temporally distinct pupil responses to MPEs: a more immediate strength-modulated prediction error response (PC3) and a later attentional orienting response (PC4). By obtaining trial-level PC scores, as opposed to condition- and subject-level scores as in prior work^{7,8}, we were able to relate the multifaceted pupil responses to electrophysiological signals on a trial-by-trial basis. These trial-level analyses

offered greater sensitivity into relationships that the pupil has with frontal theta markers of cognitive control and posterior alpha markers of attention. These analyses highlighted that PC3 scores were predicted by engagement of cognitive control (frontal theta) during memory decisions regardless of prediction error, and that for strong MPEs, PC3 scores scaled with the amount of attention allocated to strong mnemonic predictions during memory retrieval (posterior alpha suppression). The latter result complements previous findings that explicit predictions are needed for the generation of pupil-linked surprise responses⁷⁶. PC4, like PC3, was associated with cognitive control (frontal theta) during memory decisions. In addition, PC4 scores varied with the degree of posterior alpha suppression evoked following memory decisions. Altogether, these findings highlight that initial MPE-related changes in attention and/or arousal assayed separately by pupil and posterior alpha are distinct, but that over time, these measures converge and may subsequently reflect a unitary focus of attention.

Frontal theta and posterior alpha dynamics

Frontal theta and posterior alpha were positively correlated during memory retrieval, with cognitive control preceding attention signals. In contrast, in response to MPEs, there was only weak evidence for a negative relationship between frontal theta and posterior alpha, with exploratory analyses suggesting that posterior alpha was leading. Moreover, frontal theta cognitive control signals appeared to have an earlier influence than posterior alpha on pupil-linked attention and/or arousal (i.e., there was a differential association with PC3 and PC4). Future work should more directly probe the directionality of these effects to understand how control, attention, and arousal signals interact to shape behavior across timescales. Furthermore, direct assays of LC activity, which likely precede cognition-related pupil responses, would provide additional insights into underlying mechanisms driving these responses.

Evidence accumulation during memory-based decisions

Beyond being impacted by MPEs, pupil size also indexed cognitive processes during the generation of mnemonic predictions. In particular, PC1 reflected trial-level mnemonic prediction strength, providing additional support for pupil systems tracking evidence accumulation processes⁷⁷; for the current study, in evidence accumulation from memory. Given that PC1 scores were coupled with posterior alpha during memory retrieval, future work should tease apart whether pupil size and posterior alpha index distinct processes during memory reactivation and memory-based decisions beyond what has previously been observed in perceptual decision-making for parameters that modulate integration of evidence or neuronal noise⁷⁷. Such investigations could also elucidate the commonalities and differences between pupil-linked and posterior alpha-linked attention signals.

Beyond mismatch signals in PFC: Potential interactions with the hippocampus, VTA/SN, and LC

Drawing on frontal theta, the current study characterized the temporal profile of cognitive control following strong MPEs. Yet additional and more direct corroboration of prior MEG findings that vmPFC theta drives hippocampal theta⁴² is needed to resolve questions about their relative timing. Are MPEs detected first by PFC or by hippocampus (e.g., within subfield CA1)? Does prediction strength modulate the magnitude of frontal and hippocampal theta responses similarly? Alternatively, does the strength of MPEs modulate hippocampal theta linearly and frontal theta nonlinearly, such that frontal theta primarily responds to strong MPEs?

Hippocampal-PFC dynamics are a focus of current theories of mismatch detection. From one perspective, PFC is theorized to play a key role in conveying goal information to support *selective* encoding of novel information that is goal-relevant³³. Another model posits that predictions and plans for upcoming decisions are generated by the dorsolateral PFC (dlPFC) and relayed to the hippocampus, which later sends mismatch signals to medial PFC⁷⁸. The latter flow of information to medial PFC, depending on whether the hippocampus is conveying a match or a mismatch in expectations, can further bolster the planned action or enable action switching⁷⁸. The dynamics of

this information flow imply that control signals mediated by medial PFC are evoked following hippocampal theta mismatch signals, rather than with vmPFC theta driving hippocampal theta⁴². Whether and when PFC subregions drive hippocampal activity, and vice versa, remain unclear. Some studies of hippocampal-PFC dynamics during retrieval have shown that signals from PFC initiate memory retrieval processes in the hippocampus/temporal lobe^{79,80}, whereas others demonstrate that hippocampal theta may precede PFC theta during free recall⁸¹. Other studies of cross-regional communication suggest that information flows from hippocampus to PFC via theta oscillations in both humans⁸¹ and non-human animals^{82,83}. With respect to error-driven dynamics, cross-regional communication following MPEs may reflect engagement in memory encoding as opposed to memory retrieval, and these two processes may exhibit differential hippocampal-PFC dynamics. Altogether, these observations highlight critical open questions about (a) the relationship between MPE-driven PFC and hippocampal activity, (b) the importance of theta oscillations⁴⁴ in mediating cross-regional communication, (c) whether hippocampal-PFC dynamics differ during memory encoding vs. retrieval, and (d) whether the temporal dynamics of hippocampal-PFC information flow vary as a function of PFC subregion. Answers to these questions will advance understanding of cross-regional communication upon MPE detection³³.

Given limitations in the spatial resolution of scalp EEG, key questions about other components of models of mismatch detection were not examined in the current study. Future work should further examine the role of the nucleus accumbens, which is theorized to integrate goal information from PFC and MPE signals from the hippocampus and the ventral tegmental area/substantia nigra (VTA/SN)⁸⁴. The VTA/SN are posited to release dopamine in response to MPEs to tag hippocampal memory traces for consolidation³³. Whether dopamine and norepinephrine differentially contribute to learning from MPEs remains unclear. Prior work examining the integrity of LC and SN among older adults implicates the former in episodic memory performance and the latter in working memory⁵⁷. The relative roles of noradrenaline release from LC and dopamine release from VTA/SN³² during MPEs warrant further investigation. Notably, the hippocampus receives more profuse projections from LC than VTA neurons in mice⁵⁶. Moreover, LC neurons influence PFC neuronal firing^{85,86} and stimulation of the LC releases both noradrenaline and dopamine in PFC^{87,88}. PFC neurons also project to LC^{89,90} and PFC neurons can modulate LC activity⁵³. The bidirectional connectivity between LC and PFC raises questions about the temporal dynamics of their interactions following MPEs and how LC modulates PFC-hippocampus interactions.

Future oriented: How MPEs facilitate future behavior

MPEs can facilitate future behavior by increasing the probability that unexpected information is encoded into memory, thereby allowing internal models of the world to be updated and subsequently improve the accuracy of future predictions³². In the current study, the likelihood of learning from strong MPEs was associated with frontal theta magnitude during MPEs. Subsequent memory effects in frontal theta for strong mismatch probes were not selectively observed *during* MPEs, but also prior to and following MPEs; frontal theta during these periods also supported later memory for weak mismatch probes. These findings are consistent with prior work showing that higher peri-stimulus frontal theta boosts memory formation^{51,58,59}. Moreover, these findings complement previous research showing that explicit predictions are required for PE-driven learning to take place¹¹. We show that memory for expectation-violating information, regardless of the overall strength of the prediction, depended on the amount of control engaged during prediction generation. Although frontal theta *during* MPEs was more strongly associated with memory for strong vs. weak mismatch probes, it did not additionally boost memory for strong mismatch probes above that of weak mismatch probes. A richer assay of memory vividness or precision may reveal differential learning effects as a function of prediction strength. Frontal theta increases elicited at the time of the MPE may reflect PFC conveying the goal relevance of surprising information³³, hence greater frontal theta in response to strong vs. weak MPEs. Long-term learning resulting from higher frontal theta in response to strong MPEs may be mediated by hippocampal learning mechanisms and therefore may be further enhanced by consolidation. Given that scalp EEG was not acquired in Experiment 2 (recognition memory test 48hrs later), it

remains unclear whether the subsequent memory effects in frontal theta hold after allowing for memory consolidation. Whether control-associated enhancements in learning are specific to learning from MPEs can be further tested in future work by including a test of subsequent memory for weak match probes as an additional control condition for comparison.

We hypothesized that strong MPE-elicited pupillary size increases, indexed by PC3 and PC4, would support new learning, given theories of LC-hippocampus interactions, the impact of these interactions on memory consolidation¹⁹, prior findings of pupil-linked, PE-related learning effects¹¹, and prior findings that unsigned reward PEs – presumably signaled by the LC-norepinephrine system – enhance memory⁹¹. To our surprise, neither PC3 nor PC4 scores were associated with subsequent recognition memory for mismatch items on the same day (Experiment 1) or after a longer delay that allowed for overnight consolidation (48hrs later; Experiment 2). Whether a more hippocampal-dependent subsequent memory assay reveals pupil-related and consolidation-dependent learning effects remains an open question for future research.

Finally, the magnitude of observed increases in control, attention, and arousal following strong MPEs may be influenced by the decision-making process engaged when making match/mismatch judgments. Not all MPEs necessitate behavioral responses. Whether similar magnitudes in neurocognitive responses and consequences for learning are observed upon detection of an MPE, but in the absence of a decision, remains unclear.

Mnemonic predictions guide everyday behavior. The strength of these predictions varies depending on frequency of prior occurrences²⁰, emotional salience^{92,93}, and attention during learning^{55,94,95}, amongst other factors. Here, we report that violations of *strong* mnemonic predictions elicited increases in cognitive control, attention, and arousal. The level of pupil-linked attention/arousal in response to strong MPEs was influenced by the amount of attention allocated to the mnemonic prediction. Furthermore, the magnitude of cognitive control responses to strong MPEs modulated the likelihood of remembering that information in the future. Altogether, these findings advance understanding of the learning and retrieval rules of the mind and brain and add to a growing literature on how dynamic interactions between knowledge about the past and the present give rise to rapid shifts in cognition that impact learning and future remembering.

Data availability

Data and code for this manuscript are accessible at <https://github.com/alicexue/mpe-attention-cog-ctrl>.

Acknowledgements

This work was supported by the National Institute on Aging R01AG065255 to A.D.W., National Science Foundation GRF DGE-2146755 to A.M.X., and a Stanford Interdisciplinary Graduate Fellowship to A.M.X. The authors would like to thank the members of the Stanford Memory Lab for useful discussions.

Additional files

[Supplemental Figures and Tables](#)

Additional information

Funding

Funder	Grant reference number	Author
HHS NIH National Institute on Aging (NIA)	R01AG065255	Anthony D Wagner
NSF National Science Foundation Graduate Research Fellowship Program (GRFP)	DGE-2146755	Alice M Xue

Author ORCID iDs

Alice M Xue:  <https://orcid.org/0009-0006-3176-8786>

Anthony D Wagner:  <https://orcid.org/0000-0003-0624-4543>

References

1. **Nolden S**, Turan G, Güler B, Günseli E (2024) Prediction error and event segmentation in episodic memory. *Neurosci Biobehav Rev* **157** <https://doi.org/10.1016/j.neubiorev.2024.105533> | PubMed
2. **Cohen JD** (2017) Cognitive control: Core constructs and current considerations. In: Egner T (Ed). *The Wiley Handbook of Cognitive Control* Chichester, UK: John Wiley & Sons, Ltd. pp. 1-28 <https://doi.org/10.1002/9781118920497.ch1>
3. **Zacks JM**, Swallow KM (2007) Event segmentation. *Curr Dir Psychol Sci* **16**:80-84 <https://doi.org/10.1111/j.1467-8721.2007.00480.x> | PubMed
4. **Kurby CA**, Zacks JM (2008) Segmentation in the perception and memory of events. *Trends Cogn Sci* **12**:72-79 <https://doi.org/10.1016/j.tics.2007.11.004> | PubMed
5. **Radvansky GA**, Zacks JM (2017) Event boundaries in memory and cognition. *Curr Opin Behav Sci* **17**:133-140 <https://doi.org/10.1016/j.cobeha.2017.08.006> | PubMed
6. **Ezzyat Y**, Davachi L (2011) What constitutes an episode in episodic memory?. *Psychol Sci* **22**:243-252 <https://doi.org/10.1177/0956797610393742> | PubMed
7. **Clewett D**, Gasser C, Davachi L (2020) Pupil-linked arousal signals track the temporal organization of events in memory. *Nat Commun* **11**:4007 <https://doi.org/10.1038/s41467-020-17851-9> | PubMed
8. **Clewett D**, Huang R, Davachi L (2025) Locus coeruleus activation ‘resets’ hippocampal event representations and separates adjacent memories. *Neuron* **0** <https://doi.org/10.1016/j.neuron.2025.05.013> | PubMed
9. **den Ouden HEM**, Kok P, de Lange FP (2012) How prediction errors shape perception, attention, and motivation. *Front Psychol* **3** <https://doi.org/10.3389/fpsyg.2012.00548> | PubMed
10. **Frank D**, Kafkas A (2021) Expectation-driven novelty effects in episodic memory. *Neurobiol Learn Mem* **183** <https://doi.org/10.1016/j.nlm.2021.107466> | PubMed
11. **Brod G**, Greve A, Jolles D, Theobald M, Galeano-Keiner EM (2022) Explicitly predicting outcomes enhances learning of expectancy-violating information. *Psychon Bull Rev* **29**:2192-2201 <https://doi.org/10.3758/s13423-022-02124-x> | PubMed
12. **Kjaerby C**, et al. (2022) Memory-enhancing properties of sleep depend on the oscillatory amplitude of norepinephrine. *Nat Neurosci* **25**:1059-1070 <https://doi.org/10.1038/s41593-022-01102-9> | PubMed
13. **Strauch C**, Wang C.-A, Einhäuser W, Van der Stigchel S, Naber M (2022) Pupillometry as an integrated readout of distinct attentional networks. *Trends Neurosci* **45**:635-647 <https://doi.org/10.1016/j.tins.2022.05.003> | PubMed
14. **Joshi S**, Gold JI (2020) Pupil size as a window on neural substrates of cognition. *Trends Cogn Sci* **24**:466-480 <https://doi.org/10.1016/j.tics.2020.03.005> | PubMed
15. **Joshi S**, Li Y, Kalwani RM, Gold JI (2016) Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* **89**:221-234 <https://doi.org/10.1016/j.neuron.2015.11.028> | PubMed
16. **Jordan R** (2024) The locus coeruleus as a global model failure system. *Trends Neurosci* **47**:92-105 <https://doi.org/10.1016/j.tins.2023.11.006> | PubMed
17. **Yu AJ**, Dayan P (2005) Uncertainty, neuromodulation, and attention. *Neuron* **46**:681-692 <https://doi.org/10.1016/j.neuron.2005.04.026> | PubMed
18. **Bouret S**, Sara SJ (2005) Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends Neurosci* **28**:574-582 <https://doi.org/10.1016/j.tins.2005.09.002> | PubMed

19. Duszkievicz AJ, McNamara CG, Takeuchi T, Genzel L (2019) Novelty and dopaminergic modulation of memory persistence: A tale of two systems. *Trends Neurosci* **42**:102-114 <https://doi.org/10.1016/j.tins.2018.10.002> | PubMed
20. Chen J, Cook PA, Wagner AD (2015) Prediction strength modulates responses in human area CA1 to sequence violations. *J Neurophysiol* **114**:1227-1238 <https://doi.org/10.1152/jn.00149.2015> | PubMed
21. Klimesch W (2012) A-band oscillations, attention, and controlled access to stored information. *Trends Cogn Sci* **16**:606-617 <https://doi.org/10.1016/j.tics.2012.10.007> | PubMed
22. Peylo C, Hilla Y, Sauseng P (2021) Cause or consequence? Alpha oscillations in visuospatial attention. *Trends Neurosci* **44**:705-713 <https://doi.org/10.1016/j.tins.2021.05.004> | PubMed
23. Carp J, Compton RJ (2009) Alpha power is influenced by performance errors. *Psychophysiology* **46**:336-343 <https://doi.org/10.1111/j.1469-8986.2008.00773.x> | PubMed
24. Pfeffer T, et al. (2022) Coupling of pupil- and neuronal population dynamics reveals diverse influences of arousal on cortical processing. *eLife* **11** <https://doi.org/10.7554/elife.71890> | PubMed
25. Podvalny E, King LE, He BJ (2021) Spectral signature and behavioral consequence of spontaneous shifts of pupil-linked arousal in human. *eLife* **10** <https://doi.org/10.7554/elife.68265> | PubMed
26. Hong L, Walz JM, Sajda P (2014) Your eyes give you away: prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics. *PLoS One* **9**:e91321 <https://doi.org/10.1371/journal.pone.0091321> | PubMed
27. Foote SL, Berridge CW, Adams LM, Pineda JA (1991) Electrophysiological evidence for the involvement of the locus coeruleus in alerting, orienting, and attending. *Prog Brain Res* **88**:521-532 [https://doi.org/10.1016/s0079-6123\(08\)63831-5](https://doi.org/10.1016/s0079-6123(08)63831-5) | PubMed
28. Helfrich RF, Huang M, Wilson G, Knight RT (2017) Prefrontal cortex modulates posterior alpha oscillations during top-down guided visual perception. *Proc Natl Acad Sci U S A* **114**:9457-9462 <https://doi.org/10.1073/pnas.1705965114> | PubMed
29. Kumaran D, Maguire EA (2009) Novelty signals: a window into hippocampal information processing. *Trends Cogn Sci* **13**:47-54 <https://doi.org/10.1016/j.tics.2008.11.004> | PubMed
30. Kumaran D, Maguire EA (2006) An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS Biol* **4**:e424 <https://doi.org/10.1371/journal.pbio.0040424> | PubMed
31. Varga DK, Raykov PP, Jefferies E, Ben-Yakov A, Bird CM (2025) Hippocampal mismatch signals are based on episodic memories and not schematic knowledge. *Proc Natl Acad Sci U S A* **122**:e2503535122 <https://doi.org/10.1073/pnas.2503535122> | PubMed
32. Sinclair AH, Manalili GM, Brunec IK, Adcock RA, Barense MD (2021) Prediction errors disrupt hippocampal representations and update episodic memories. *Proc Natl Acad Sci U S A* **118**:e2117625118 <https://doi.org/10.1073/pnas.2117625118> | PubMed
33. Lisman JE, Grace AA (2005) The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron* **46**:703-713 <https://doi.org/10.1016/j.neuron.2005.05.002> | PubMed
34. Hasselmo ME, Wyble BP (1997) Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behav Brain Res* **89**:1-34 [https://doi.org/10.1016/s0166-4328\(97\)00048-x](https://doi.org/10.1016/s0166-4328(97)00048-x) | PubMed
35. Vinogradova OS (2001) Hippocampus as comparator: role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus* **11**:578-598 <https://doi.org/10.1002/hipo.1073> | PubMed
36. Kumaran D, Maguire EA (2007) Which computational mechanisms operate in the hippocampus during novelty detection?. *Hippocampus* **17**:735-748 <https://doi.org/10.1002/hipo.20326> | PubMed
37. Duncan K, Ketz N, Inati SJ, Davachi L (2012) Evidence for area CA1 as a match/mismatch detector: a high-resolution fMRI study of the human hippocampus. *Hippocampus* **22**:389-398 <https://doi.org/10.1002/hipo.20933> | PubMed

38. **Chen J**, Olsen RK, Preston AR, Glover GH, Wagner AD (2011) Associative retrieval processes in the human medial temporal lobe: hippocampal retrieval success and CA1 mismatch detection. *Learn Mem* **18**:523-528 <https://doi.org/10.1101/lm.2135211> | PubMed
39. **Bein O**, Duncan K, Davachi L (2020) Mnemonic prediction errors bias hippocampal states. *Nat Commun* **11**:3451 <https://doi.org/10.1038/s41467-020-17287-1> | PubMed
40. **Hasselmo ME**, Bodelón C, Wyble BP (2002) A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Comput* **14**:793-817 <https://doi.org/10.1162/089976602317318965> | PubMed
41. **Chen J**, et al. (2013) Human hippocampal increases in low-frequency power during associative prediction violations. *Neuropsychologia* **51**:2344-2351 <https://doi.org/10.1016/j.neuropsychologia.2013.03.019> | PubMed
42. **Garrido MI**, Barnes GR, Kumaran D, Maguire EA, Dolan RJ (2015) Ventromedial prefrontal cortex drives hippocampal theta oscillations induced by mismatch computations. *Neuroimage* **120**:362-370 <https://doi.org/10.1016/j.neuroimage.2015.07.016> | PubMed
43. **Lin J.-J**, et al. (2017) Theta band power increases in the posterior hippocampus predict successful episodic memory encoding in humans. *Hippocampus* **27**:1040-1053 <https://doi.org/10.1002/hipo.22751> | PubMed
44. **Mitchell DJ**, McNaughton N, Flanagan D, Kirk IJ (2008) Frontal-midline theta from the perspective of hippocampal 'theta'. *Prog Neurobiol* **86**:156-185 <https://doi.org/10.1016/j.pneurobio.2008.09.005> | PubMed
45. **Rhodes N**, et al. (2023) Measurement of Frontal Midline Theta Oscillations using OPM-MEG. *Neuroimage* **271** <https://doi.org/10.1016/j.neuroimage.2023.120024> | PubMed
46. **Cavanagh JF**, Frank MJ, Klein TJ, Allen JJB (2010) Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *Neuroimage* **49**:3198-3209 <https://doi.org/10.1016/j.neuroimage.2009.11.080> | PubMed
47. **Cavanagh JF**, Figueroa CM, Cohen MX, Frank MJ (2012) Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cereb Cortex* **22**:2575-2586 <https://doi.org/10.1093/cercor/bhr332> | PubMed
48. **Mas-Herrero E**, Marco-Pallarés J (2014) Frontal theta oscillatory activity is a common mechanism for the computation of unexpected outcomes and learning rate. *J Cogn Neurosci* **26**:447-458 https://doi.org/10.1162/jocn_a_00516 | PubMed
49. **Cavanagh JF**, Frank MJ (2014) Frontal theta as a mechanism for cognitive control. *Trends Cogn Sci* **18**:414-421 <https://doi.org/10.1016/j.tics.2014.04.012> | PubMed
50. **Cavanagh JF**, Cohen MX (2022) Frontal midline theta as a model specimen of cortical theta. In: Gable PA, Miller MW, Bernat EM (Eds). *The Oxford handbook of EEG frequency* **178** Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780192898340.013.9>
51. **Hsieh L.-T**, Ranganath C (2014) Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. *Neuroimage* **85**:721-729 <https://doi.org/10.1016/j.neuroimage.2013.08.003> | PubMed
52. **Badre D** (2025) Cognitive control. *Annu Rev Psychol* **76**:167-195 <https://doi.org/10.1146/annurev-psych-022024-103901> | PubMed
53. **Totah NK**, Logothetis NK, Eschenko O (2021) Synchronous spiking associated with prefrontal high γ oscillations evokes a 5-Hz rhythmic modulation of spiking in locus coeruleus. *J Neurophysiol* **125**:1191-1201 <https://doi.org/10.1152/jn.00677.2020> | PubMed
54. **Schwartz ST**, Yang H, Xue AM, Wagner AD (2025) Attending to remember: Recent advances in methods and theory. *Curr Dir Psychol Sci* <https://doi.org/10.1177/09637214251339452> | PubMed
55. **Chun MM**, Turk-Browne NB (2007) Interactions between attention and memory. *Curr Opin Neurobiol* **17**:177-184 <https://doi.org/10.1016/j.conb.2007.03.005> | PubMed

56. Takeuchi T, et al. (2016) Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature* **537**:357-362 <https://doi.org/10.1038/nature19325> | PubMed
57. Dahl MJ, et al. (2023) The integrity of dopaminergic and noradrenergic brain regions is associated with different aspects of late-life memory performance. *Nat. Aging* **3**:1128-1143 <https://doi.org/10.1038/s43587-023-00469-z> | PubMed
58. Cohen N, et al. (2015) Peri-encoding predictors of memory encoding and consolidation. *Neurosci Biobehav Rev* **50**:128-142 <https://doi.org/10.1016/j.neubiorev.2014.11.002> | PubMed
59. White TP, et al. (2013) Theta power during encoding predicts subsequent-memory performance and default mode network deactivation. *Hum Brain Mapp* **34**:2929-2943 <https://doi.org/10.1002/hbm.22114> | PubMed
60. Brodeur MB, Guérard K, Bouras M (2014) Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLoS One* **9**:e106953 <https://doi.org/10.1371/journal.pone.0106953> | PubMed
61. Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. *J Neurosci* **33**:10235-10242 <https://doi.org/10.1523/jneurosci.0983-13.2013> | PubMed
62. Konkle T, Brady TF, Alvarez GA, Oliva A (2010) Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol Sci* **21**:1551-1556 <https://doi.org/10.1177/0956797610385359> | PubMed
63. Willenbockel V, et al. (2010) Controlling low-level image properties: the SHINE toolbox. *Behav Res Methods* **42**:671-684 <https://doi.org/10.3758/brm.42.3.671> | PubMed
64. Rosenberg M, Noonan S, DeGutis J, Esterman M (2013) Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. *Atten Percept Psychophys* **75**:426-439 <https://doi.org/10.3758/s13414-012-0413-x> | PubMed
65. Gramfort A, et al. (2013) MEG and EEG data analysis with MNE-Python. *Front Neurosci* **7** <https://doi.org/10.3389/fnins.2013.00267> | PubMed
66. Madore KP, et al. (2020) Memory failure predicted by attention lapsing and media multitasking. *Nature* **587**:87-91 <https://doi.org/10.1038/s41586-020-2870-z> | PubMed
67. Kret ME, Sjak-Shie EE (2019) Preprocessing pupil size data: Guidelines and code. *Behav Res Methods* **51**:1336-1342 <https://doi.org/10.3758/s13428-018-1075-y> | PubMed
68. R Core Team (2024) *R: A Language and Environment for Statistical Computing*.
69. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models Using lme4. *J Stat Softw* **67**:1-48 <https://doi.org/10.18637/jss.v067.i01>
70. Barr DJ (2013) Random effects structure for testing interactions in linear mixed-effects models. *Front Psychol* **4** <https://doi.org/10.3389/fpsyg.2013.00328> | PubMed
71. Vuorre M, Bolger N (2018) Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behav Res Methods* **50**:2125-2143 <https://doi.org/10.3758/s13428-017-0980-9> | PubMed
72. Bürkner P.-C (2017) Brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw* **80**:1-28 <https://doi.org/10.18637/jss.v080.i01>
73. Urai AE, Braun A, Donner TH (2017) Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat Commun* **8** <https://doi.org/10.1038/ncomms14637> | PubMed
74. LePage M, Ghaffar O, Nyberg L, Tulving E (2000) Prefrontal cortex and episodic memory retrieval mode. *Proc Natl Acad Sci U S A* **97**:506-511 <https://doi.org/10.1073/pnas.97.1.506> | PubMed
75. Tulving E (2002) Episodic memory: from mind to brain. *Annu Rev Psychol* **53**:1-25 <https://doi.org/10.1146/annurev.psych.53.100901.135114> | PubMed
76. Brod G, Hasselhorn M, Bunge SA (2018) When generating a prediction boosts learning: The element of surprise. *Learn Instr* **55**:22-31 <https://doi.org/10.1016/j.learninstruc.2018.01.013>
77. Keung W, Hagen TA, Wilson RC (2019) Regulation of evidence accumulation by pupil-linked arousal processes. *Nat. Hum Behav* **3**:636-645 <https://doi.org/10.1038/s41562-019-0551-4> | PubMed

78. Numan R (2021) The prefrontal-hippocampal comparator: Volition and episodic memory. *Percept Mot Skills* **128**:2421-2447 <https://doi.org/10.1177/00315125211041341> | PubMed
79. Tomita H, Ohbayashi M, Nakahara K, Hasegawa I, Miyashita Y (1999) Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* **401**:699-703 <https://doi.org/10.1038/44372> | PubMed
80. Rajasethupathy P, et al. (2015) Projections from neocortex mediate top-down control of memory retrieval. *Nature* **526**:653-659 <https://doi.org/10.1038/nature15389> | PubMed
81. Anderson KL, Rajagovindan R, Ghacibeh GA, Meador KJ, Ding M (2010) Theta oscillations mediate interaction between prefrontal cortex and medial temporal lobe in human memory. *Cereb Cortex* **20**:1604-1612 <https://doi.org/10.1093/cercor/bhp223> | PubMed
82. Siapas AG, Lubenov EV, Wilson MA (2005) Prefrontal phase locking to hippocampal theta oscillations. *Neuron* **46**:141-151 <https://doi.org/10.1016/j.neuron.2005.02.028> | PubMed
83. Brockmann MD, Pöschel B, Cichon N, Hanganu-Opatz IL (2011) Coupled oscillations mediate directed interactions between prefrontal cortex and hippocampus of the neonatal rat. *Neuron* **71**:332-347 <https://doi.org/10.1016/j.neuron.2011.05.041> | PubMed
84. Kamiński J, et al. (2018) Novelty-sensitive dopaminergic neurons in the human substantia nigra predict success of declarative memory formation. *Curr Biol* **28**:1333-1343.e4 <https://doi.org/10.1016/j.cub.2018.03.024> | PubMed
85. Sara SJ, Hervé-Minvielle A (1995) Inhibitory influence of frontal cortex on locus coeruleus neurons. *Proc Natl Acad Sci U S A* **92**:6032-6036 <https://doi.org/10.1073/pnas.92.13.6032> | PubMed
86. Neves RM, van Keulen S, Yang M, Logothetis NK, Eschenko O (2018) Locus coeruleus phasic discharge is essential for stimulus-induced gamma oscillations in the prefrontal cortex. *J Neurophysiol* **119**:904-920 <https://doi.org/10.1152/jn.00552.2017> | PubMed
87. Devoto P, Flore G, Saba P, Fà M, Gessa GL (2005) Stimulation of the locus coeruleus elicits noradrenaline and dopamine release in the medial prefrontal and parietal cortex. *J Neurochem* **92**:368-374 <https://doi.org/10.1111/j.1471-4159.2004.02866.x> | PubMed
88. Florin-Lechner SM, Druhan JP, Aston-Jones G, Valentino RJ (1996) Enhanced norepinephrine release in prefrontal cortex with burst stimulation of the locus coeruleus. *Brain Res* **742**:89-97 [https://doi.org/10.1016/s0006-8993\(96\)00967-5](https://doi.org/10.1016/s0006-8993(96)00967-5) | PubMed
89. Arnsten AF, Goldman-Rakic PS (1984) Selective prefrontal cortical projections to the region of the locus coeruleus and raphe nuclei in the rhesus monkey. *Brain Res* **306**:9-18 [https://doi.org/10.1016/0006-8993\(84\)90351-2](https://doi.org/10.1016/0006-8993(84)90351-2) | PubMed
90. Jodo E, Chiang C, Aston-Jones G (1998) Potent excitatory influence of prefrontal cortex activity on noradrenergic locus coeruleus neurons. *Neuroscience* **83**:63-79 [https://doi.org/10.1016/s0306-4522\(97\)00372-2](https://doi.org/10.1016/s0306-4522(97)00372-2) | PubMed
91. Rouhani N, Niv Y (2021) Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife* **10**:e61077 <https://doi.org/10.7554/eLife.61077> | PubMed
92. Anderson AK, Wais PE, Gabrieli JDE (2006) Emotion enhances remembrance of neutral events past. *Proc Natl Acad Sci U S A* **103**:1599-1604 <https://doi.org/10.1073/pnas.0506308103> | PubMed
93. Kensinger EA (2009) Remembering the details: Effects of emotion. *Emot Rev* **1**:99-113 <https://doi.org/10.1177/1754073908100432> | PubMed
94. Mulligan NW (1998) The role of attention during encoding in implicit and explicit memory. *J Exp Psychol Learn Mem Cogn* **24**:27-47 <https://doi.org/10.1037//0278-7393.24.1.27> | PubMed
95. Craik FI, Govoni R, Naveh-Benjamin M, Anderson ND (1996) The effects of divided attention on encoding and retrieval processes in human memory. *J Exp Psychol Gen* **125**:159-180 <https://doi.org/10.1037//0096-3445.125.2.159> | PubMed

Peer reviews

Reviewer #1 (Public review):

This manuscript describes a multi-modal study of associative learning and memory in humans, that combines scalp EEG, pupillometry and behavioral analysis to explore the construct of mnemonic prediction errors (MPEs), in terms of their relationship to attention and cognitive control. Across two pooled studies, participants performed associative memory tasks in which they learned the relationship between a cue word (action verb) and subsequent picture (animate or inanimate) with a strong vs. weak (4 or 1 repetitions) encoding manipulation. At test, participants were encouraged to generate a prediction following the cue word to determine whether the subsequently presented picture was a match or mismatch. The timecourse of pupillary responses during match decisions were decomposed using temporal principal components analysis, which identified 6 distinct and overlapping processes. Some of the components (PC3/PC4) exhibited sensitivity to both the strength and mismatch conditions, as well as behavior (both RT and accuracy) and retrieval success on the subsequent trial. Furthermore, relationships were also observed between pupillary responses (specifically for PC4) and both frontal theta and posterior alpha power measures obtained from scalp EEG in Experiment 2, as well as for frontal theta and subsequent learning from mismatch stimuli (assessed using subsequent memory findings from a surprise recognition test). The authors suggest the findings indicate that MPEs elicit changes in attention, arousal and cognitive control which impact subsequent learning.

Strengths:

This manuscript has many strengths, including a clever study design, thoughtful integration of multiple neurocognitive measures, and a set of rigorous and technically sophisticated analyses, which reveal a large set of relationships among the measures and behavior. The findings demonstrating brain/physiology-behavior relationships are particularly important, in that they point to potential functional consequences of MPEs.

Weaknesses:

The technical proficiency and complexity of the study and analysis also presents a clear limitation and challenge for interpretation. It is likely that readers, even those that are quite knowledgeable about the methods, constructs, and questions being addressed will often struggle (as this reviewer did) to keep the large set of findings in mind and gain understanding of how they all fit together.

Indeed, it seems like there many threads running together in the paper which make it challenging to find the through-line of the key findings. The authors do address some of the key questions motivating the paper in the Introduction, but the results are somewhat ambiguous with regard to the primary question of the study as to whether the detection of MPEs leads to interaction among cognitive control, attention, and arousal. To their credit, the authors tackle this question through both cross-correlation and formal mediation analyses, and summarize these in diagrammatic figures (Figure 3, Figure 6). Yet it is not resolved whether the results represent a clear answer pointing to independence, or rather a lack of statistical power, or ill-resolved formulation of the mediational relationship. In particular, the cross-correlation suggests that posterior alpha suppression in response to MPEs does precede frontal theta, yet this indirect relationship does not explain the variation in trial-by-trial RTs on mismatches. This suggests a potential model misspecification.

In addition to the primary interaction issue mentioned above (between cognitive control, attention & arousal), the Introduction lays out a number of claims: 1) that pupil size will be more sensitive to strong than weak MPEs; 2) that MPE-linked increases in attention (indexed

with posterior alpha suppression) and arousal (indexed with pupil size) will be linked to learning; and 3) MPE learning will vary as a function of prediction strength. Given the focus on learning, it is somewhat surprising that learning is not included in the mediation models. As the authors indicate in the Discussion, the use of trial-by-trial RT variation to drive the mediation model might be problematic, given that the RTs are sensitive to a range of factors beyond mnemonic prediction strength and also are under competing pressures (longer for mismatches than matches, due to surprise-linked slowing, but also faster following stronger rather than weaker mnemonic predictions). Thus, an alternative possibility might be to use trial-by-trial recognition of mismatches as the outcome variable in mediation models rather than trial-by-trial RT as the independent variable.

A large component of the results (Sections 2 and 3) is devoted to analyses of cue-linked pupil and EEG processes that putatively reflect mnemonic predictions (i.e., occurring before picture probes are presented and match/mismatch detection, i.e., MPEs occur). Yet these Results and the subsequent pupillary PCA components (PC1 and PC5) that are elicited are not well-integrated with the primary themes of the paper or the causal hypotheses. One finding that does seem to figure prominently (in that it is mentioned in Abstract, Introduction & Discussion) relates to the amount of attention allocated to the mnemonic prediction generation. Yet this finding is not well emphasized in the Results themselves. Possibly it refers to the negative relationship between posterior alpha during memory retrieval and the magnitude of pupillary PC3 component, described in Section 3. But it was quite challenging to identify amongst the wealth of results described in this Section as well as the others. More generally, the large amount of findings described across all four lengthy Results sections makes it challenging for readers to discern what are the key ones that the authors would like to highlight.

It is recommended that the authors do another pass through the paper to better highlight the most critical findings that they want to emphasize or which are most interpretable from a mechanistic and causal flow perspective and then de-emphasize or move other findings to the Supplemental Materials. Although the authors are to be commended for such a rigorous and comprehensive set of analyses, there are so many of them and findings, that the key points get buried and the reader needs to struggle potentially unnecessarily to identify the key take-away points.

<https://doi.org/10.7554/eLife.109678.2.sa2>

Reviewer #2 (Public review):

Summary:

The authors studied cognitive control and attention in response to mnemonic prediction errors (MPEs): situations in which the external reality violates internal memory-based predictions. The behavioral task first established strong versus weak predictions, and then either confirmed or violated these predictions. The authors examined markers of cognitive control (frontal theta) and attention (posterior alpha suppression, pupil response) while strong and weak predictions were confirmed or violated. They found increased cognitive control (frontal theta) for strong MPEs, which correlated with subsequent memory. Markers of attention (alpha suppression, pupil response) also accompanied strong MPEs but did not correlate with subsequent memory. Pupil response was investigated using an interesting approach that decomposes the response into different components, finding that different components respond earlier or later and show different correlations with MPEs and their strength. The authors also investigated how EEG, reaction time, and pupil responses correlated with one another, providing further insight into the mechanism underlying the response to MPEs. Together, the study points toward multiple control and attention mechanisms involved in MPE response and memory.

Strengths:

The study has a clear behavioral paradigm with multiple measures - behavioral, EEG, and pupillometry that offer an investigation into different aspects of MPE response and memory.

The study is also very comprehensive in looking at multiple phases in processing MPEs: the prediction phase (prior to the violation), the response to MPEs, and subsequent memory of MPEs, all within one study. Specifically, the link between neural mechanisms and subsequent memory is a major advancement, as most prior studies did not include this component. Mechanisms underlying subsequent memory of MPEs are theoretically important, as a primary function of MPEs is to promote learning and memory. As the authors mention, the different neural and pupillary signals are not robustly correlated, suggesting multiple mechanisms underlying MPE detections, which is interesting, offers avenues for future research, and can facilitate a better theory of how MPEs are processed in the brain. Finally, the decomposition of pupil response into different components and their correlation with behavior (RT during match/MPE detection) is interesting.

Weaknesses:

The methods are rigorous, and the data support the claims. The weaknesses are minor and are offered here as avenues for future research.

(1) The relationships the authors find between brain measures and pupil components were largely not specific to mismatches/matches. Thus, the specificity of this relationship is untested.

(2) The results with subsequent memory are important and address a major gap in the field that largely did not relate neural effects of MPE to subsequent memory. However, one major limitation of the study is that the authors did not test memory for matches. I understand the logic of avoiding testing matches. Because matches were repeated more times in the study, it's not a fair comparison and could change participants' overall criterion for old/new decisions. Future research could address this, e.g., by testing weak matches or potentially using a between-subject design.

Comments on revised version.

The authors addressed all my concerns. I appreciate the authors' thoughtful and detailed response.

<https://doi.org/10.7554/eLife.109678.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public review):

Strengths:

This manuscript has many strengths, including a clever study design, thoughtful integration of multiple neurocognitive measures, and a set of rigorous and technically sophisticated analyses, which reveal a large set of relationships among the measures and behavior. The findings demonstrating brain/physiology-behavior relationships are particularly important, in that they point to potential functional consequences of MPES.

We thank the reviewer for noting these strengths of the work along with the below encouragement to revise the manuscript to better highlight the key findings and their implications.

Weaknesses:

The technical proficiency and complexity of the study and analysis also present a clear limitation and challenge for interpretation. As a reader, even those who are quite knowledgeable about the methods, constructs, and questions being addressed will often struggle (as this reviewer did) to keep the large set of findings in mind and gain an understanding of how they all fit together.

Indeed, it seems like there are many threads running together in the paper, which makes it challenging to find the through-line of the key findings, or to understand how they might relate to some pre-existing hypotheses, rather than merely interesting patterns detected in the data. In the Introduction and Discussion, it seems as if the key question is to understand the pathways by which MPEs impact cognition, but this is a rather broad topic, so it is not clear exactly what the authors are aiming at with this question and study design.

As an example, authors operationalize frontal theta power as an index of cognitive control demand, and one of the pathways by which MPEs impact cognition. But this point becomes somewhat circular, since it is not clear how or why the Mismatch x Strength interaction in frontal theta reflects that demand. It would have been better to set this pattern up in the Introduction as a theoretically driven hypothesis, since it currently appears more like a post-hoc interpretation. This is mirrored by how the issue is first brought up in the Introduction, where it states somewhat vaguely: "whether MPEs are followed by an increase in frontal theta... warrants closer examination".

Again, we appreciate the reviewer's thoughtful feedback on where the manuscript can be clearer, especially given the rich set of results it reports. Following the reviewer's guidance, we restructured and revised the Introduction to further motivate the hypotheses that (a) MPEs increase both attention/arousal (grounded in studies of Event Segmentation Theory) and cognitive control (given findings on reward prediction errors and other types of prediction errors), and (b) there are greater increases in these processes triggered by strong compared to weak MPEs. To better link these hypotheses to resulting statistical tests, we note that hypothesis (a) was tested in our trial-level regression models in Fig. 1 by examining main effects of Strength, whereas hypothesis (b) was tested in our models in the Mismatch x Strength interactions. On point (b) and potentially circularity, we note that previous work indicates that frontal theta scales with negative reward prediction errors; as such, we hypothesized that stronger MPEs would elicit more frontal theta, as evidenced by a robust Mismatch x Strength interaction during the probe period.

Later in the results, there are findings relating frontal theta to pupil dilation, posterior alpha suppression and then subsequent memory. It was hard to understand how all the findings might be linked together functionally or conceptually. Are the authors potentially postulating a mediating or mechanistic pathway, in which the MPE leads to increased cognitive control (frontal theta), which then leads to enhanced subsequent memory of those events? If this is the case, then maybe a formal path analysis would be the best way to test or state this hypothesis. It would also be useful to specify more clearly how the pupil components and alpha suppression factor into this mediating path, since it was not clear.

Relatedly, the authors suggest that internal attention and arousal also play relevant roles in this pathway, but these are also not clear. In some cases, it is stated as if this is a

distinct pathway from the cognitive control one, since there is a focus in the results on the independence of frontal theta and posterior alpha, but elsewhere they seem to be treated as two aspects, or distinct steps, within a single pathway. Again, these different threads of the findings were quite challenging for the reader to follow. Pathway analyses, such as with multiple mediation or moderated mediation, could be a useful way to address this question. For example, it seems as if readiness-to-remember is another behavioral outcome (like subsequent memory) that could be used in the search for mediators.

We thank the reviewer for highlighting these ambiguities in the original submission and for the thoughtful encouragement to leverage mediation models to more formally test the hypothesized relationships between MPE magnitude and constructs of control, attention, and arousal. In the revision, we now more clearly hypothesize that the effects of strong MPE-driven increases in attention and arousal might be explained, in part, by cognitive control (as indexed by frontal theta) upregulating attention and arousal. To more explicitly test this model of the relationships between our measures, as recommended by Reviewer #1, we now include multivariate mediation analyses to assess whether, at a trial-level, changes in posterior alpha and the immediate pupil effect PC3 are explained in part by increases in frontal theta. Because changes in posterior alpha following MPEs and PC3 scores did not predict subsequent memory, mediation analyses addressing the hypothesis that our attention/arousal measures mediate the effect of frontal theta on subsequent memory were not conducted. Following insights from a cross-correlation analysis, as recommended by Reviewer #2, we tested an additional model to examine whether the effects of MPE magnitude on frontal theta were explained in part by changes in posterior alpha. Examination of the posterior distributions of the indirect effects did not favor our hypothesized model, nor the alternative model that attention upregulates control. Altogether, these outcomes suggest that increases in control, attention, and arousal following strong MPEs may be elicited independently. Yet, we also note that the current set of experiments may be underpowered for these mediation analyses; future work can further investigate the directionality between these effects. Finally, with respect to the readiness-to-remember findings, they were removed in the interest of space, as recommended by Reviewer #2.

At the minimum, it would be quite helpful to have diagrammatic figures that specify the hypothesized and observed relationships between independent variables (Strength, Mismatch), physiological indices (pupil dilation components, frontal theta, posterior alpha) and key outcome measures (accuracy, RT, next-trial retrieval success, subsequent memory), so that the reader can refer back to them as each component of the analyses is conducted.

To further increase conceptual clarity, we also followed this helpful suggestion, adding diagrammatic figures to illustrate our hypotheses regarding interactions between the effects (Fig. 3a) and to summarize the observed relationships between MPEs, control, attention, and arousal (Fig. 6).

Minor Points:

Many figures had x-axes showing a pupil component or EEG power metric broken down by quartile or quintile. Yet nowhere is it ever explained why this graphical (or analytic?) approach is used and what it reflects, or how it is decided which break down to use (quartile/quintile). If the data are analyzed as a correlation, why is a scatterplot not shown instead?

In the linear mixed effects models, continuous values were used to assess relationships between variables. In the figures, the continuous variables were binned into quartiles or quintiles for ease of visualization. We opted to visualize the data using this approach, rather

than with a scatterplot, given the large number of trials. We updated the figure captions to clarify the approach.

It was surprising that, unlike readiness-to-remember, which was analyzed via logistic regression and odds-ratio, subsequent memory was not analyzed in the same fashion (i.e., as a binary outcome variable predicted by frontal theta), rather than in a reverse chronological one (subsequent memory predicting frontal theta). Historically, it was the case that subsequent memory was analyzed in this manner, but that was before the era in which trial-level linear mixed-effect models were in wide usage, as they are implemented in this study. Thus, the choice seems like a wasted opportunity or a step backwards analytically.

We thank the reviewer for this encouragement and agree with the point. In the revision, we note that the readiness-to-remember results were removed in the interest of space and clarity, as was recommended by Reviewer #2. With respect to the subsequent memory analyses, they are now analyzed via logistic regression.

Reviewer #2 (Public review):

Strengths:

The study has a clear behavioral paradigm with multiple measures - behavioral, EEG, and pupillometry that offer an investigation into different aspects of MPE response and memory.

The study is also very comprehensive in looking at multiple phases in processing MPEs: the prediction phase (prior to the violation), the response to MPEs, and subsequent memory of MPEs, all within one study. Specifically, the link between neural mechanisms and subsequent memory is a major advancement, as most prior studies did not include this component. Mechanisms underlying subsequent memory of MPEs are theoretically important, as a primary function of MPEs is to promote learning and memory. As the authors mention, the different neural and pupillary signals are not robustly correlated, suggesting multiple mechanisms underlying MPE detections, which is interesting, offers avenues for future research, and can facilitate a better theory of how MPEs are processed in the brain. Finally, the decomposition of pupil response into different components and their correlation with behavior (RT during match/MPE detection) is interesting.

We thank the reviewer for noting these strengths of the work along with the below encouragement to revise the manuscript to better highlight the key findings and their implications.

Weaknesses:

The methods are rigorous, and the claims are mostly supported by the data, but there are a few weaknesses or places that could be improved:

(1) The authors conduct PCA analysis to identify different components of the pupillary response to MPE and relate them to behavior. Specifically, the authors identify components PC3 and PC4, which they interpret as related to MPE. However, some parts of the interpretation could be clearer or better justified:

(a) The authors refer to PC4 as "post-decision cognitive processing". But, given that RT was between .5-.7s, and PC3 peaked after more than 1s, wouldn't it be cautious to interpret PC3 as postdecision as well?

Thank you for raising this point. Given that pupil is a relatively sluggish response, it is possible that both components reflect post-decision cognitive processing, even if PC3 peaks

before PC4. Following the reviewer's guidance to adopt more cautious language, we replaced "post-decision" with "post-MPE".

(b) MPEs overall elicit longer RTs in this study, suggesting that long RT is a behavioral marker of MPE. Nonetheless, the authors argue on p. 12: "Altogether, these findings indicate that when stronger mnemonic predictions (as indexed by shorter RTs) were violated." And, PC3 is correlated with shorter RTs for mismatches, meaning that behaviorally, these trials were more similar to matches. Thus, how do the authors interpret shorter versus longer RTs for MPEs, and what processes do these RT reflect?

We thank the reviewer for stressing the need for greater clarity regarding the relationships between RT and the constructs of interest. With respect to RT, we interpret the condition-level difference in RTs between mismatch and match trials as a behavioral marker of an MPE. However, when comparing mismatch trials within a given strength condition to each other (i.e., an analysis at the trial-level), shorter RTs may reflect a stronger prediction, greater certainty that the probe is a mismatch, and therefore the experience of a stronger MPE. Note that while larger PC3 scores were associated with shorter RTs for mismatches (Fig. 2b, right), the mismatch RTs in the largest PC3 quartile were still longer than those on match trials in the corresponding quartile (in other words, there was still a condition-level difference in RTs in the largest PC3 quartile, suggesting that the mismatch trials in this bin are behaviorally still likely to be different from match trials in the corresponding bin).

To clarify these relationships and our interpretation, we modified the referred to text: "(as indexed by shorter strong mismatch RTs)." Moreover, we added text to the Discussion, further delineating our reasoning here and the implications of our findings for understanding the mechanisms giving rise to and triggering by MPEs of varying strengths. This includes adding an explicit summary of the logic and findings that notes that, at the trial-level, shorter RTs may reflect stronger predictions; at the match/mismatch condition-level, longer RTs may reflect the experience of a MPE. For PC3, shorter RTs (trials with a stronger prediction) in the Strong Mismatch condition were associated with a larger pupillary response. The added text notes that "while longer mean RTs for mismatches compared to matches are a behavioral marker of a MPE, within-condition differences in RTs (i.e., between mismatch trials) may reflect more subtle differences in MPE magnitude, with shorter RTs reflecting stronger predictions and thus stronger MPEs. Strong mismatch RTs were used in the mediation models as a proxy measure of MPE magnitude; this estimate may be noisy because RTs in this experiment are likely sensitive to factors independent of mnemonic prediction strength (e.g., preparatory attention (Supplementary Fig. 6) or memory strength of the mismatch probe). This limitation may have additionally reduced sensitivity to detecting indirect effects."

(2) The brain to pupil relationship (p. 13-14): If I understand correctly, this was done on a trial-by-trial basis, but the high temporal resolution allows doing the analysis in a time-resolved manner - does brain activity at a certain time point preceding/following the pupil response correlate with the pupil response? It might be that cognitive control influences attention mechanisms or vice versa (because there is some overlap in the response). Although not testing causality, this temporally resolved correlation would be an interesting way to start probing how signals might influence each other.

Thank you for this suggestion. We now report a cross-correlation analysis (Fig. 3d) that suggests that cognitive control increases precede attention decreases at retrieval, whereas in response to a strong MPE, cognitive control increases follow attention increases. There were no significant clusters for the temporal relationships between frontal theta and pupil, nor for posterior alpha and pupil.

(3) The relationships the authors find between brain measures and pupil components were largely not specific to mismatches/matches. However, are they specific to this task? I

think it would benefit the paper to show that these relationships are potentially specific to making match/mismatch memory decisions, versus, e.g., any stimulus processing. For example, the authors could run the same analyses locked to stimuli in the study phase, anticipating a different pattern, if indeed these findings are specific to the associative memory task.

Many of the associations between our measures indeed did not show an interaction with Mismatch. Due to jitter in the ISI in the study phase, some of the analyses in the retrieval phase cannot be performed in the exact same way for the study phase. We will leave these questions to be addressed in future research. We agree that an important question for future research is to address whether these responses depend on making match/mismatch decisions and now include consideration of this point in the Discussion: “Finally, the magnitude of observed increases in control, attention, and arousal following strong MPEs may be influenced by the decision-making process engaged when making match/mismatch judgments. Not all MPEs necessitate behavioral responses. Whether similar magnitudes in neurocognitive responses and consequences for learning are observed upon detection of an MPE, but in the absence of a decision remains unclear.”

(4) During memory retrieval (i.e., before the probe), the authors find that frontal theta, a marker of cognitive control, was associated on a trial-by-trial basis with more posterior alpha (i.e., less alpha suppression, potentially reflecting less attention), and that this association was stronger for weaker predictions. The authors interpreted this as weaker predictions necessitating more cognitive control, and that more cognitive control was recruited specifically in trials where retrieval included less content (memory reinstatement) to attend to. Generally, cognitive control is recruited to facilitate memory retrieval. If so, one possible interpretation is that this correlation reflects cognitive control effort that has failed to produce enough memory reinstatement. The other possibility is that this correlation reflects more specific retrieval of the correct probe, without retrieval of interfering items (i.e., overall less content). I believe that the former explanation predicts that this correlation would be associated with longer RTs (more difficult decisions), while the latter predicts shorter RTs (easier decisions due to successful retrieval), at least for matches.

Thank you for these insightful comments. Because this analysis is not key to the main questions about MPEs and given both reviewers’ concerns that the manuscript can be overwhelming for the reader given the sheer number of findings reported, we opted to move this point from the main text to the Supplement. However, following the reviewer’s guidance here, we conducted the proposed analyses and tested these alternative accounting by modeling RTs. The results favour the former interpretation:

“Greater control being associated with less attention could reflect failure in controlled retrieval efforts to reinstate sufficient memory evidence of the probe and thus fewer retrieval products to which attention is allocated. Alternatively, greater cognitive control could increase the likelihood of retrieval success, eliciting selective retrieval of the correct probe and inhibition of interfering items. To address these alternatives, we examined how the association between frontal theta and posterior alpha related to the difficulty of a trial, as assayed by RTs. The former failure-of-control account would predict that a stronger positive frontal theta- posterior alpha association would relate to longer RTs, whereas the greater-retrieval-specificity account would predict that a stronger positive association would relate to shorter RTs. In a model predicting RTs as a function of frontal theta, posterior alpha, Strength, Mismatch, and their interactions, we found a two-way frontal theta $\times$ posterior alpha interaction ($\beta=0.016$, $CI=[0.001, 0.031]$, $p=0.033$), such that a stronger positive association between frontal theta and posterior alpha predicted longer RTs. This relationship did not differ as a function of Strength (no frontal theta $\times$ posterior alpha $\times$ Strength interaction: $\beta=-0.015$, $CI=[-0.033, 0.004]$, $p=0.124$), Mismatch (no frontal theta $\times$ posterior alpha $\times$ Mismatch

interaction: $\beta = -0.016$, $CI = [-0.050, 0.018]$, $p = 0.342$), or interact with Strength and Mismatch (no frontal theta $\times$ posterior alpha $\times$ Strength $\times$ Mismatch interaction: $\beta = 0.024$, $CI = [-0.014, 0.062]$, $p = 0.218$). Together, these outcomes support the idea that during memory retrieval, positive coupling between frontal theta and posterior alpha may reflect failure or inefficiency of cognitive control efforts to rapidly accumulate mnemonic evidence to which to attend in support of a memory decision.”

(5) In section 3, the authors found a positive relationship between alpha during memory retrieval and PC3 during MPE. If I understood correctly, this means that less attention during retrieval (less suppression) is correlated with a stronger PC3 response. How do the authors interpret this? Maybe along the same lines as in (5), specifically retrieving the correct information (i.e., less retrieved content to attend to) means a stronger prediction, leading to a stronger MPE, and a stronger MPE response, as reflected by PC3?

We appreciate this comment, as it highlights a need for greater clarity here. The observed relationship was actually negative (Fig. 4f), meaning that more attention during retrieval was associated with a stronger PC3 response. We suspect that the lack of clarity here may be due to the original statement that “there was a positive relationship between posterior alpha suppression during memory retrieval [and PC3 scores]”. To increase clarity, we have modified this statement to “there was a negative relationship between posterior alpha during memory retrieval [and PC3 scores]”. We interpret the negative relationship with posterior alpha (i.e., positive relationship with posterior alpha suppression) to indicate “that greater attentional allocation during memory retrieval, which occurs when memories are stronger and more retrieval products can be reinstated and attended to (Fig. 1e; Fig. 4c), predicts the magnitude of immediate pupil responses to MPEs.”

(6) The results with subsequent memory are important and address a major gap in the field that largely did not relate neural effects of MPE to subsequent memory. However, one major limitation of the study is that the authors did not test memory for matches. I understand the logic of avoiding testing matches. Because matches were repeated more times in the study, it's not a fair comparison, and could change participants' overall criterion for old/new decisions. However, one possibility would have been to test only the weak prediction; this could have given some specificity to the neural subsequent memory findings.

We were indeed concerned about the change in decision criterion and did not include match items for this reason. Nonetheless, this is a useful suggestion that future work could include the weak match items for a better comparison of subsequent recognition memory. We now comment on this in the Discussion: “Whether control-associated enhancements in learning are specific to learning from MPEs can be further tested in future work by including a test of subsequent memory for weak match probes as an additional control condition for comparison.”

(7) The authors nicely characterized the different PC of pupillary MPE response. But, with respect to subsequent memory, they only present pupil size. Unless there is some methodological reason that prevents testing subsequent memory on the PC, I think this will be very informative about the potential mechanisms underlying memory of MPE.

The pupil PCs were not associated with subsequent memory, though there are some interesting trends in the 48-delay condition which could be explored in future work. These findings are now reported in the Supplement (Supplementary Fig. 16).

(8) This paper includes many interesting findings, and I am not sure how they all come together into a cohesive mechanistic understanding of MPE response and subsequent memory. I think the paper would benefit from either a conceptual mechanism figure or,

in the Discussion, have a summary of a proposed mechanism integrating the findings together.

We thank the reviewer for stressing this point, which also was raised by Reviewer #1. To better emphasize the novel contributions of the work and to assist the reader's understanding of the key findings, we: (a) revised the Introduction to more explicitly describe our hypotheses about the relationships among our measures as they relate to MPE responses and subsequent memory; (b) now report mediation models to more directly test these relationships and include diagrams of the hypothesized relationships; and (c) include a schematic figure at the end of the Results to highlight the key mechanistic relationships supported by the data.

(9) Relatedly, the section "Immediate, strength-sensitive neurocognitive impacts of MPEs" does not link the arguments to specific data points, so it's hard to follow which data specifically the authors are interpreting.

The discussion in this paragraph rests on the Strength $\times$ Mismatch interactions observed in Figures 1d-f, summarized in the first sentence of the section. To increase clarity, we changed the title of this section to "Neurocognitive impacts of MPEs are strength-sensitive".

(10) If I understand correctly, the authors did not find improved memory for strong compared to weak MPE. First, I think this behavioral result should be incorporated in the main paper and in the interpretation of the results. Second, given that the neural effects the authors tested either correlated with memory for strong MPE or did not show a relationship with memory, what neural/pupil response could explain memory for weak MPE?

Thank you for raising these points. The behavioral result and the frontal theta trial-level regression model for subsequent memory are now described in the main results and included in the Discussion. As noted in the Discussion, the trial-level regression model indicates pre-probe and post-MPE frontal theta effects may explain memory for weak (and strong) mismatch probes. While the magnitude of probe-period MPE frontal theta is additionally predictive of memory for strong mismatch probes (as indicated by the Subsequent Memory $\times$ Strength interaction in the trial-level regression model in Fig. 5c and the logistic regression in Fig. 5b), frontal theta during this period does not additionally enhance memory for strong mismatch probes above that of weak mismatch probes (Fig. 5a). We added more discussion on why memory for strong and weak mismatch probes in the current experiment did not differ. We also discuss potential directions for future research to further probe mechanisms underlying MPE-driven learning.

Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

It is recommended that the authors determine whether formal path analyses, testing for mediation and moderation, would provide a useful approach from which to better integrate the disparate set of findings and make clear their causal/functional implications.

At a minimum, adding a diagrammatic figure is recommended to visually depict the key components of the study (independent variables, physiological indices, outcome measures) and how they relate to each other both conceptually (ideally in a theoretically hypothesized manner) and in terms of the observed findings. Such a figure will help the reader keep track of the many types of findings and results threads, and with the goal of better organizing the results into a clearer narrative through-line.

Thank you for the suggestions to add mediation analyses and diagrammatic figures. To more formally address our hypothesis that increases in attention/arousal following MPEs are explained in part by increases in cognitive control, we tested two mediation models: one with posterior alpha as the measure of attention and the other with pupil PC3 scores. We now diagram this hypothesis in Fig. 3a. Given the outcomes of the cross-correlation analysis suggested by Reviewer #2, we also tested an additional model where posterior alpha might explain the impact of strong MPEs on frontal theta (now diagrammed in Fig. 3e). We did not find credible evidence for an indirect effect in any of the models, suggesting that strong MPE-driven increases in control, attention, and arousal may be elicited independently. Finally, in a newly added summary diagram (Fig. 6), we highlight the observed effects of strong vs. weak MPEs on RTs, control, attention and arousal; differences in control, attention, and arousal for strong vs. weak predictions preceding the MPE; trial-level mismatch-specific or strong-specific effects on subsequent memory; and mismatch-specific interactions between processes during retrieval and responses to MPEs. We hope these revisions address Reviewer #1's concerns and that these figures help the reader keep track of the key hypothesized relationships and main findings.

Reviewer #2 (Recommendations for the authors):

(1) The relationship between event segmentation and prediction errors has been reviewed recently in two papers (Nolden et al., 2024, Neuroscience & Biobehavioral Reviews; Rouhani et al., 2024, JOCN for a potentially relevant computational model). I wonder if insights from these papers can inform the Introduction/Discussion of the current manuscript.

Thank you for these suggestions. These papers are now incorporated into the Introduction and the Discussion.

(2) Brod et al. (2022, Psych. Bull. Rev.) have previously reported increased pupil dilations for MPE correlating with subsequent memory, specifically for strong prediction errors. I think it's worth including this paper in the Introduction as the finding is highly relevant. The Brod paper might provide more direct evidence of "MPE-related increases in pupil size" than the evidence the authors provide (p. 3).

Thank you for this suggestion. This paper is now incorporated into the Introduction and Discussion.

(3) I'm confused about the temporal analysis: "Trial-level regression analyses were conducted on the frontal theta, posterior alpha, and pupil time series from the associative retrieval test to identify temporal clusters that were sensitive to the factors of Mismatch (i.e., mismatch vs. match probes) and/or Strength (i.e., strong vs. weak associative pairs). For each participant and each time point, a linear regression model testing main effects of Mismatch and Strength, and a Mismatch × Strength interaction was run using R to compute beta weights for each regressor." What regression exactly was run? A separate model for each participant and time point? Across trials, then? Later, the authors mention that beta weights were averaged across participants and t-tests and permutation tests were conducted. However, if the data were averaged, what t-test was conducted? And how was the permutation test conducted? It's also unclear what the authors mean by "the sign of each participant's beta weights" - what sign?

Thank you for raising this point. A regression was run separately for each participant and each time point, across trials: neurocognitive measure ~ Mismatch + Strength + Mismatch: Strength. Beta weights were averaged only for visualization; t-tests were conducted on each set of beta weights (across participants), separately for each time point. We modified the text of the Methods to describe our procedure more clearly.

(4) The associative memory and recognition accuracy data are presented as d' . In addition, the authors should provide hits and false alarms to facilitate a better interpretation of the results.

We now report these outcomes in Table 1 and refer to them in the main text.

(5) The authors argue regarding the frontal theta that "Qualitatively, the main effect of Strength emerged later than the main effect of Mismatch, suggesting that the increase in frontal theta evoked by the probe was more sustained for weak compared to strong trials (or, as a corollary, that the greater control elicited by strong MPEs enabled more rapid resolution of conflict and ultimate choice selection)." It was unclear to me how that stems from the data.

This statement has been removed altogether.

(6) Especially in Figure 1, I think clarity can be improved if the authors would indicate the specific subsection they are referring to in the text, because even within, e.g., 1d, there are different graphs, so mentioning which graph is relevant for which statement would be helpful to the reader.

Thank you for this suggestion for improving the clarity of the manuscript. Fig. 1d-f includes multiple parts because we wanted to show the raw data (the mean time series on the left) as well as the model coefficients (time series on the right). We are hopeful that, given that each subplot is titled and that the main text refers to both the subplots on the left and on the right, this will be clear to the reader as is. We welcome further guidance if this remains a concern.

(7) This seems highly speculative to me: "Elevated PC4 scores on weak hits, where no error was made, may instead reflect retrieval practice and thus internally oriented attention. On these trials, when cue-elicited retrieval may have been weaker, the probe may have provided additional support for pattern completion of the learning episode for that association, and this engagement in memory retrieval may have elicited pupil dilation (c.f., Strength effect in Figure 1f). Overall, PC4 may therefore reflect an attentional orienting response that, depending on the relative success of memory retrieval and probe identity, may direct attention internally or externally." (p.13). In my opinion, the authors make a lot of assumptions about underlying processes. I'd consider removing.

We removed this text.

(8) In Figure 3c, should the x-axis be PC3 quintile? And (c) is not in the figure caption.

Thank you for this note. We addressed these issues.

(9) The carryover effects the authors report are interesting, but they seem detached, and it is unclear how they fit with the additional findings. In a paper that already includes many findings, I'd recommend either integrating better or removing.

Following this guidance, we removed the readiness-to-remember findings in the interest of space and clarity.

<https://doi.org/10.7554/eLife.109678.2.sa0>