JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes.

Nicole M. Long[1], Hongmi Lee[2] and Brice A. Kuhl[1]

[1]*Department of Psychology, University of Oregon 97403*
[2]*Department of Psychology, New York University 10003*

DOI: 10.1523/JNEUROSCI.1850-16.2016

Received: 9 June 2016

Revised: 26 October 2016

Accepted: 1 November 2016

Published: 7 November 2016

Author contributions: N.M.L. and H.L. analyzed data; N.M.L., H.L., and B.A.K. wrote the paper; B.A.K. designed research; B.A.K. performed research.

Conflict of Interest: The authors declare no competing financial interests

We thank Marvin Chun for helpful discussion and feedback. We thank Sam Cartmell for assistance with data collection. This work was supported by NIH Grant NS089729.

Corresponding Author: Nicole Long (niclong@uoregon.edu); Brice Kuhl (bkuhl@uoregon.edu)

Cite as: J. Neurosci 2016; 10.1523/JNEUROSCI.1850-16.2016

Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes.

Running Title: Predictions and hippocampal mismatch detection

Nicole M. Long[1], Hongmi Lee[2], Brice A. Kuhl[1]

Affiliations:

1: Department of Psychology, University of Oregon 97403

2: Department of Psychology, New York University 10003

Corresponding Author: Nicole Long (niclong@uoregon.edu); Brice Kuhl (bkuhl@uoregon.edu)

Conflict of Interest: The authors declare no competing financial interests

Abstract

The hippocampus is thought to compare predicted events with current perceptual input, generating a mismatch signal when predictions are violated. However, most prior studies have only inferred when predictions occur without directly measuring them. Moreover, an important but unresolved question is whether hippocampal mismatch signals are modulated by the degree to which predictions differ from outcomes. Here we conducted a human fMRI study in which subjects repeatedly studied various word-picture pairs, learning to predict particular pictures (outcomes) from the words (cues). Following initial learning, a subset of cues were paired with a novel, unexpected outcome whereas other cues continued to predict the same outcome. Critically, when outcomes changed, the new outcome was either 'near' to the predicted outcome (same visual category as the predicted picture) or 'far' from the predicted outcome (different visual category). Using multi-voxel pattern analysis, we indexed cue-evoked reactivation (prediction) within neocortical areas and related these trial-by-trial measures of prediction strength to univariate hippocampal responses to the outcomes. We found that prediction strength positively modulated hippocampal responses to unexpected outcomes, particularly when unexpected outcomes were close–but not identical–to the prediction. Hippocampal responses to unexpected outcomes were also associated with a tradeoff in performance during a subsequent memory test: relatively faster retrieval of new (updated) associations but relatively slower retrieval of the original (older) associations. Together, these results indicate that hippocampal mismatch signals reflect a comparison between active predictions and current outcomes and that these signals are most robust when predictions are similar, but not identical, to outcomes.

Significance Statement

Although the hippocampus is widely thought to signal 'mismatches' between memory-based predictions and outcomes, previous research has not directly linked hippocampal mismatch signals to neural measures of prediction strength. Here, we show that hippocampal mismatch signals increase as a function of the strength of predictions in neocortical regions. This increase in hippocampal mismatch signals was particularly robust when outcomes were similar, but not identical, to predictions. These results indicate that hippocampal mismatch signals are driven by both the active generation of predictions as well as the similarity between predictions and outcomes.

Introduction

Memories for past experience allow for outcomes to be predicted based on current perceptual experience. Many theoretical perspectives (Gluck & Myers, 1993; Eichenbaum, 2004; Lisman & Grace, 2005; Buckner, 2010; Rolls, 2013; Davachi & DuBrow, 2015) and empirical findings (Kumaran & Maguire, 2006a; Duncan, Curtis, & Davachi, 2009; Chen, Olsen, Preston, Glover, & Wagner, 2011) suggest that the hippocampus plays a critical role in comparing memory-based predictions with perception-based outcomes. When predictions do not match outcomes, the hippocampus is thought to generate mismatch signals that reflect these expectancy violations (Kumaran & Maguire, 2006b; Duncan, Ketz, Inati, & Davachi, 2012; Chen, Cook, & Wagner, 2015). While mismatch signals should criti- cally depend on the active generation of predictions (Kumaran & Maguire, 2007), existing evidence for hippocampal mismatch signals comes from paradigms where prediction strength is inferred but not directly measured (Kumaran & Maguire, 2006b; Chen et al., 2015). Moreover, an important but unresolved question is whether hippocampal mismatch signals are sensitive to the degree of similarity between predictions and outcomes. When predictions are close–but not identical–to outcomes, are mismatch signals relatively weaker or stronger? From a prediction error perspective, mismatch signals may be greater when predictions are relatively farther from outcomes (Fiorillo, Tobler, & Schultz, 2003). However, given the proposed role of the hippocampus in creating distinct representations of simi- lar stimuli (Düzel et al., 2003; Norman, Newman, Detre, & Polyn, 2006; Leutgeb, Leutgeb, Moser, & Moser, 2007; Bakker, Kirwan, Miller, & Stark, 2008; Yassa & Stark, 2011; Hulbert & Norman, 2015; Favila, Chanales, & Kuhl, 2016), hippocampal mismatch signals may *increase* when predictions are relatively closer to outcomes.

Whereas memory-based predictions are thought to be generated by the hippocampus (Eichenbaum & Fortin, 2009; Buckner, 2010; Davachi & DuBrow, 2015; Hindy, Ng, & Turk-Browne, 2016), these predictions are reflected by reactivation in neocortical areas (Kok, Jehee, & de Lange, 2012; Hindy et al., 2016). Memory reactivation has been most extensively documented in visual cortical areas (e.g. Polyn, Natu, Cohen, & Norman, 2005; Kuhl, Rissman, Chun, & Wagner, 2011), but has also been observed in fronto-parietal regions. For example, reactivation within posterior parietal cortex (PPC) reflects detailed information about retrieved content (Kuhl & Chun, 2014; H. Lee, Chun, & Kuhl, 2016). Reactivation within medial prefrontal cortex (mPFC) is thought to be of particular importance when past experience is compared to present experience (Kroes & Fernández, 2012; Schlichting & Preston, 2015; Demblon, Bahri, & D'Argembeau, 2016; Richter, Chanales, & Kuhl, 2016), and connections between mPFC and hippocampus may provide the ideal scaffolding for a prediction generation and comparison system (Preston & Eichenbaum, 2013; Anderson, Bunce, & Barbas, 2015; Rajasethupathy et al., 2015). Collectively, PPC and mPFC are part of the brain's so-called default mode network (DMN) and it has been argued that the DMN plays a central role in memory-based predictions (Bar, 2007, 2009). Indeed, activity patterns distributed across the DMN contain rich information about the contents of memory (Chen, Leong, Norman, & Hasson, 2016).

In the present study, we tested whether and how hippocampal mismatch signals are modulated by the strength and similarity of memory-based predictions. We used an associative memory task in which human subjects learned cue-outcome pairings. Outcomes were either expected or unexpected, and unexpected outcomes were further sub-divided into

'near' and 'far' outcomes based on their similarity to predictions. Motivated by the aforementioned studies, we measured prediction strength by using multi-voxel pattern analysis of the DMN and two sub-regions within the DMN: mPFC and PPC. We first tested whether prediction strength was related to the magnitude of hippocampal outcome responses. We hypothesized that hippocampal outcome responses would increase when predictions were strong but ultimately violated, consistent with a mismatch signal. Second, given the role of the hippocampus in discriminating similar events (Bakker et al., 2008; Yassa & Stark, 2011), we tested whether mismatch signals were stronger when predictions are 'near' to outcomes as compared to 'far' from outcomes. Finally, we assessed whether hippocampal outcome responses were associated with subsequent behavioral expressions of successful memory updating.

Materials and Methods

Subjects

Twenty-three (8 female; mean age = 22.4 years) and 26 (15 female; mean age = 21.2 years) right-handed, native English speakers from the Yale University community participated in two separate experiments. Three subjects were excluded from the second experiment, one due to technical error, one due to user error during scanning, and the third due to failure to follow task instructions, resulting in a final set of 23 subjects included in each of the two experiments. The experiments were identical with the exception of a small difference in the post-scan memory test (see below). We therefore collapsed data across both experiments. All subjects had normal or corrected-to-normal vision. Informed consent was obtained in accordance with the Yale Institutional Review Board.

Materials

Stimuli consisted of 144 words and 252 pictures. Words were verbs with a length between 4 and 11 letters (M = 6.3). Pictures consisted of greyscale photographs (225 x 225 pixels) of famous people (e.g., Steve Martin; *faces*), famous locations (e.g., Sydney Opera House; *scenes*), and common objects (e.g., toothbrush; *objects*). All word-picture pairings and the assignment of words and pictures to conditions were randomized for each subject.

Procedure and design

Acquisition phase. Subjects completed three acquisition rounds during which they encoded word-picture (cue-outcome) pairs (Figure 1). Words were presented directly above each picture. Subjects were instructed to learn these associations in anticipation of a later memory test. No behavioral responses were made during the acquisition rounds. Each acquisition round contained the same 144 associations, with the order and trial structure varying across rounds. During the first two acquisition rounds, which were completed before subjects entered the fMRI scanner, each word-picture association was presented for 2750 ms, followed by a 500 ms fixation cross. The order of presentation of the associations was randomized. The third acquisition round was conducted during a single functional imaging scan. Again, all 144 associations were presented, but during this round, associations were presented in 'mini blocks' grouped by the visual categories of pictures. For example, four associations containing face pictures might

be presented consecutively, followed by four associations containing scene pictures, etc. Within each mini block, a word-picture association was presented for 2500 ms, followed by a fixation cross for 500 ms. Thus, each mini block lasted 12000 ms. Each mini block was followed by a 6000-ms inter-block interval. This interval began with a fixation cross displayed for 700 ms, and was then followed by a series of four randomly oriented (left- vs. right-oriented) arrows. Each arrow was presented for 800 ms and was followed by a fixation cross for 400 ms. A final fixation cross was then presented for 500 ms before the next block began. The motivation for the mini block structure of the third acquisition round was to optimize the use of the functional data for training a pattern classifier to discriminate the three visual categories of the pictures (faces vs. objects vs. scenes).

Updating phase. Following the third acquisition round, subjects began the critical updating phase, which was also conducted while fMRI data were collected. During the updating phase, all of the 144 original cues were presented again. Of the 144 cues, 108 were presented with novel outcomes/pictures (*unexpected* trials) and the remaining 36 were presented with the same outcomes/pictures as during acquisition (*expected* trials). Subjects were instructed that when associations changed (unexpected trials), they should 'update' their memory to reflect the new word-picture association. Importantly, unexpected trials could be further sub-divided according to whether the outcome was from the same visual category as the original outcome (*near* trials) or from a distinct visual category (*far trials*). For example, a 'near trial' would occur if a cue that was originally paired with a scene picture (e.g., Sydney Opera House) was 'updated' with a new scene picture (e.g., ruins of Pompeii; Figure 1). Likewise, a 'far trial' would occur if a cue that was originally paired with a face picture (e.g., Steve Martin) was 'updated' with an object picture (e.g., pacifier; Figure 1).

In contrast to the acquisition phase, the presentation of cues and outcomes was temporally offset in the updating phase. Specifically, each cue (word) was presented alone for 4000 ms and then the outcome (picture) appeared beneath the word for an additional 4000 ms (Figure 1). This temporal offset between cues and outcomes allowed for predictions to be generated and measured before outcomes were displayed. Because our goal was to decode/measure trial-level cue and outcome responses, we did not jitter the presentation of stimuli, which would have potentially confounded measurements with jitter length. Trials were separated by an 8000-ms inter-trial interval. This interval began with a fixation cross displayed for 700 ms, and was then followed by a series of six randomly oriented (left- vs. right-oriented) arrows. Each arrow was presented for 700 ms and was followed by a fixation cross for 400 ms. A final fixation cross was then presented for 700 ms before the next trial began. The updating phase was divided into six rounds. Each round included 24 trials (18 unexpected, 6 expected). The 18 unexpected trials were balanced in terms of the visual categories corresponding to the original and new associations. Specifically, there were 9 different combinations of visual categories for the original vs. new associations (face vs. object, face vs. scene, etc.) and each of these 9 combinations appeared twice per updating round. Likewise, for the 6 expected trials within each round there were an even number of trials corresponding to each of the three visual categories (i.e., 2 face trials, 2 object trials, 2 scene trials).

During the updating phase, subjects were instructed to respond, via button box, whether each outcome (picture)

was new (i.e., had not been encountered during acquisition) or old (i.e., had been encountered during acquisition). Thus, for all unexpected trials, the correct response was 'new,' and for all expected trials the correct response was 'old.' Due to technical error, responses were only recorded from the 6th updating round for the majority of subjects. However, this task was only included in order to ensure subject vigilance, and accuracy from the available data was near ceiling (mean = 93%).

Post-test phase. Following scanning, subjects completed a behavioral post-test that probed memory for each of the 144 associations. On each trial in the post-test, subjects were presented with a cue word and were given a 2- or 3-alternative forced-choice task of selecting the picture that had been presented with the cue word during the updating phase (i.e., the most recent association). The cue words were presented directly above the set of pictures. For cue words that had been paired with the same picture in both the acquisition and updating phases (expected condition), only two choices were included on the test trial: the picture that had been paired with the cue during the acquisition and updating rounds (the target) and a picture from a different association (a lure). For cues that were paired with a new picture during the updating phase (unexpected condition), three alternatives were included on the test trial: the picture that had been paired with the cue during the updating phase (the target), the picture that had been paired with the cue during the acquisition phase (the original association), and a picture from a different association (a lure). Note: lure pictures were randomly drawn from the three visual categories. Following the forced-choice decision, subjects rated their confidence (high/low). The only difference between Experiments 1 and 2 is that in Experiment 2, after subjects selected the 'most recent' association for each cue, they were also asked to select the original picture (for the unexpected condition only). The post-test was self-paced.

fMRI data acquisition

Imaging data were collected on a 3T Siemens Trio scanner at the Anlyan Center at Yale University using a 12-channel head coil. Before the functional imaging, two T1-weighted anatomical scans were collected (in-plane and high-resolution 3D). Functional data were collected using a T2*-weighted gradient EPI sequence; TR = 2000 ms, TE = 25 ms, flip angle = 90°, 34 axial-oblique slices, 224 mm FOV (3.5 × 3.5 × 4 mm). A total of seven functional scans (one acquisition phase scan and six updating phase scans) were collected. The acquisition phase scan consisted of 329 volumes. Each updating phase scan consisted of 197 volumes. The first five volumes from each scan were discarded to allow for T1 equilibration.

fMRI data preprocessing

fMRI data preprocessing was conducted using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Images were first corrected for head motion. High-resolution anatomical images were coregistered to the functional images and segmented into gray matter, white matter, and CSF. Segmented gray matter images were "skull-stripped" and normalized to a gray matter Montreal Neurological Institute template. Resulting parameters were used for normalization of functional images. Functional images were resampled to 3-mm isotropic voxels and smoothed with

a Gaussian kernel (5 mm FWHM). Functional data were then de-trended, high-pass filtered (0.01 Hz), and Z-scored within scan (mean response of each voxel within each scan = 0). Next, data were temporally compressed by averaging over volumes. For the acquisition phase data (i.e., the third acquisition round), each of the 36 mini blocks was treated as a single 'trial' by averaging the 3rd-8th volumes collected after the start of the mini block. This resulted in a total of 36 spatial volumes, with 12 volumes per visual category. For the updating phase, each trial was separated into two components: prediction and outcome. For the prediction component, the 3rd and 4th volumes (4-8s) following cue onset were averaged together. For the outcome component, the 3rd and 4th volumes (4-8s) following outcome onset (or 5th and 6th volumes following cue onset) were averaged together. Thus, each updating phase trial corresponded to a single spatial volume for the prediction component and a single spatial volume for the outcome component.

Pattern classification analyses

Pattern classification analyses were performed using penalized (L2) logistic regression (penalty parameter = 1), implemented via the Liblinear toolbox (Fan, Chang, Hsieh, Wang, & Lin, 2008) and custom MATLAB (RRID:SCR_001622) code. Before pattern classification analyses were performed, an additional round of z-scoring was performed across voxels so that the mean activation within each spatial volume was equal to 0. This additional z-scoring step eliminated trial-level differences in mean univariate activity (Kuhl, Johnson, & Chun, 2013; Kuhl & Chun, 2014).

Subject-specific classifiers were first trained to discriminate face, scene and object trials using data from the acquisition phase. The trained classifiers were then tested on the updating-phase trials, separately for the prediction and outcome components. Classifier performance was assessed in two ways. *Classification accuracy* represented a binary coding of whether or not the classifier successfully 'guessed' the visual category of the original outcome (prediction component) or the visual category of the actual, perceived outcome (outcome component). We used classification accuracy for general assessment of classifier performance (i.e., whether predicted and actual outcomes could be decoded). *Classifier evidence* was a continuous value reflecting the logit-transformed probability that the classifier assigned to the relevant category for each trial. Classifier evidence was used as a trial-specific measure of prediction strength, which was related to univariate activity from the outcome component (see below).

Relationship between decoded prediction strength and univariate responses to outcomes

As described above, each updating-phase trial was decomposed into a prediction component and an outcome component. To test whether univariate responses to outcomes were modulated by prediction strength, linear regression analyses were applied in which univariate responses to outcomes were regressed on trial-by-trial measures of predic- tion strength (from the pattern classifiers). For each subject and each combination of prediction/outcome regions of interest, a total of 12 linear regression analyses were run, reflecting different combinations of trial type (expected, near, far) and visual categories for the original/new images (faces, objects, scenes). For the expected trials, there were three regressions: one for each visual category. For the near trials, there were also three regressions: one for

each visual category. For the unexpected trials, there were six regressions to reflect each possible combination of original and new visual categories: face-scene, face-object, scene-face, scene-object, object-face, and object-scene. We performed separate regression analyses for each visual category condition in order to ensure that any relationships between prediction strength and outcome responses could not be an artifact of differences between visual categories. Resulting t-statistics were averaged across visual category conditions within each trial type, yielding three mean t-statistics per subject: one for the expected trial type, one for the near trial type, and one for the far trial type. Mean t-statistics were then used for group-level analyses.

Regions of interest

Pattern classification analyses were performed on two networks of *a priori* interest: the default mode network (DMN) and the visual network (VisN), as identified from a prior large-scale analysis of fMRI resting-state connectivity (Yeo et al., 2011, Figure 2A). Pattern classification analyses were additionally performed on two sub-regions within the DMN, medial prefrontal cortex (mPFC) and posterior parietal cortex (PPC). We defined these regions using the conjunction of the DMN mask and either medial prefrontal regions or posterior parietal regions as defined in the Anatomical Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). We also defined three *a priori* anatomical regions of interest (ROIs; Figure 3A): hippocampus, the pars triangularis region of left inferior frontal gyrus (LIFG_t) and caudate. LIFG_t and caudate were included as comparison regions given that these regions have previously been implicated in signaling when expectations are violated (Schultz, Dayan, & Montague, 1997; Kawagoe, Takikawa, & Hikosaka, 2004; Daw & Doya, 2006; Daw & Shohamy, 2008) or when mnemonic associations change (Dolan & Fletcher, 1997; Kuhl, Bainbridge, & Chun, 2012). The anatomical masks were created from the AAL atlas. The DMN mask was modified to remove voxels which overlapped with the hippocampal ROI (fewer than ten voxels were removed).

Results

Decoding prediction signals

The primary goal of our study was to measure whether and how neural predictions modulate hippocampal responses to outcomes. To measure neural predictions, we applied pattern classification analyses to the default mode network (DMN). We targeted the DMN because it has been specifically proposed to play a role in memory-based predictions (Bar, 2007, 2009) and because prior applications of pattern classification analyses have revealed robust evidence of memory reactivation within sub-regions of the DMN–particularly within posterior parietal cortex (PPC) and medial prefrontal cortex (mPFC; Euston, Gruber, & McNaughton, 2012; Zeithamova, Dominick, & Preston, 2012; Kuhl & Chun, 2014; Schlichting & Preston, 2015; Richter et al., 2016). For comparison, we also applied pattern classification analyses to voxels within a network of visual regions (VisN) which included areas dedicated to both early visual processing (occipital cortex) and higher level perception (ventral temporal cortex). We predicted that, compared

to the VisN, the DMN would show greater representation of predicted outcomes; in contrast, we predicted that, relative to the DMN, the VisN would show greater representation of perceived outcomes. To first test for evidence of predictions, we trained an L2 logistic regression classifier on acquisition-phase data and tested the classifier on updating-phase data during the cue presentation interval. Decoding of the predicted outcome (the original associate) was well above chance in both the DMN ($t(45) = 7.0, p < .001$) and the VisN ($t(45) = 6.5, p < .001$; Figure 2B), with significantly greater decoding performance in the DMN than in the VisN ($t(45) = 2.3, p = .02$). To assess decoding of the perceived outcome, we again used a classifier that was trained on acquisition-phase data, but now tested the classifier on updating-phase data during the stimulus presentation interval (the outcome). We found reliable decoding of the presented outcome in both the DMN ($t(45) = 15.8, p < .001$) and the VisN ($t(45) = 44.7, p < .001$), but decoding performance was now significantly greater in the VisN than in the DMN ($t(45) = 22.3, p < .001$). A 2 × 2 ANOVA with region (DMN vs. VisN) and decoding target (predicted vs. perceived outcome) yielded a significant interaction ($F(1,45) = 617.5, p < .001$), reflecting the relatively greater contribution of the DMN to representing memory-based predictions and of the VisN to representing perceived outcomes. Within the DMN, decoding of the predicted outcome was well above chance both in mPFC and PPC (mPFC, $t(45) = 5.0, p < .001$; PPC, $t(45) = 9.5, p < .001$).

Prediction strength modulates hippocampal responses to unexpected event outcomes

Having established that memory-based predictions are robustly reflected in DMN activity patterns, we next sought to relate DMN prediction strength (memory reactivation for the original associate) to hippocampal outcome responses. To the extent that hippocampal mismatch signals reflect a comparison between predictions and outcomes, then hippocampal outcome responses should increase when predictions are relatively strong but differ from outcomes. Thus, we expected a positive relationship between prediction strength and hippocampal outcome responses for unexpected trials (near and far), but not for expected trials. Notably, there was no overall difference in hippocampal activation across expected vs. unexpected trials ($t(45) = .07, p = .95$), indicating that associative novelty alone (without accounting for prediction strength) did not modulate hippocampal activity. To index prediction strength, we used classifier evidence for memory reactivation, which yielded a continuous, trial-by-trial value where higher values index stronger predictions (see *Methods*; Gershman, Schapiro, Hupbach, & Norman, 2013; Kuhl et al., 2013). Importantly, whereas predictions were decoded using brain volumes acquired 3-4 TRs after the onset of the *cue* (word), outcome responses were based on brain volumes acquired 3-4 TRs after the onset of the *outcome* (picture), which was 5-6 TRs after cue onset. This allowed for separation of cue- and outcome-evoked activity.

Subject-specific linear regression analyses were applied in which classifier evidence for the predicted outcome (prediction strength) was the independent variable and hippocampal univariate activity in response to the actual outcome was the dependent measure. Using classifier evidence from the DMN to index predictions, there was a significant positive relationship between prediction strength and hippocampal responses to unexpected outcomes ($t(45) = 2.6, p = .01$; Figure 3B, top panel), but not to expected outcomes ($t(45) = .35, p = .73$). Thus, hippocampal responses to unexpected outcomes increased as a function of DMN prediction strength, consistent with the idea that

hippocampal mismatch signals depend on the active generation of predictions. Prediction strength within VisN did not modulate hippocampal responses to unexpected ($t(45) = .86, p = .39$) or expected ($t(45) = 1.9, p = .06$) outcomes (Figure 3B, bottom panel). A region (DMN, VisN) × trial type (expected, unexpected) repeated measures ANOVA revealed a significant interaction ($F(1,45) = 4.2, p = .046$). We also tested whether prediction strength within the hippocampus was correlated with hippocampal outcome responses, but did not find a significant relationship for expected ($t(45) = -.32, p = .75$) or unexpected ($t(45) = -1.3, p = .18$) trials.

We next considered two subregions of the DMN: mPFC and PPC. For mPFC, prediction strength was positively related to hippocampal responses to unexpected outcomes ($t(45) = 2.4, p = .02$), but not expected outcomes ($t(45) = -1.5, p = .15$), with a significant difference in the relationship for unexpected vs. expected trials ($t(45) = 2.6, p = .01$; Figure 3B, second panel). For PPC, prediction strength was positively related to hippocampal responses to unexpected outcomes ($t(45) = 2.0, p = .049$), but not expected outcomes ($t(45) = 1.5, p = .13$); however, there was no significant difference in the relationship for unexpected vs. expected trials ($t(45) = .33, p = .75$; Figure 3B, third panel). The relatively greater effect of trial type for mPFC was confirmed by a significant region (mPFC, PPC) × trial type (expected, unexpected) interaction (repeated measures ANOVA, $F(1,45) = 6.1, p = .017$).

Similarity between predictions and outcomes modulates hippocampal responses

The preceding analyses confirm our first hypothesis–that hippocampal responses to unexpected outcomes are modulated by the strength of neural predictions. Our second question was whether this relationship varies as a function of the similarity between predictions and outcomes. Specifically, does the relationship between prediction strength and hippocampal outcome response vary for near vs. far unexpected trials? Using the DMN to index predictions, we found that the difference between the relationships (near vs. far) trended toward significance ($t(45) = 1.7, p = .10$; Figure 3C, top panel). There was a significant, positive relationship between prediction strength and hippocampal outcome responses for near trials ($t(45) = 2.7, p = .01$), but not far trials ($t(45) = 1.3, p = .21$). For VisN, prediction strength was not significantly related to hippocampal outcome responses for near ($t(45) = 1.0, p = .32$) or far trials ($t(45) = .16, p = .87$), nor was there a significant difference between near vs. far trials ($t(45) = .88, p = .38$). However, a region (DMN, VisN) × trial type (near, far) repeated measures ANOVA did not reveal a significant interaction ($F(1,45) = .60, p = .44$).

For mPFC, the relationship between prediction strength and hippocampal outcome response was significantly more positive for near than far trials ($t(45) = 3.0, p = .005$; Figure 3C, second panel), and was positive and highly significant for near trials ($t(45) = 3.2, p = .003$), but not far trials ($t(45) = .28, p = .78$). Qualitatively, the results for PPC were similar, but attenuated (Figure 3C, third panel): the relationship between PPC prediction strength and hippocampal outcome response was not significantly more positive for near compared to far trials ($t(45) = 1.5, p = .13$), but there was a significant positive relationship for near ($t(45) = 2.3, p = .03$), but not far trials ($t(45) = .30, p = .77$). A region (mPFC, PPC) × trial type (near, far) repeated measures ANOVA revealed a significant main effect of trial type ($F(1,45) = 8.1, p = .007$), reflecting the stronger relationship for near than far trials, with no interaction ($F(1,45) = 2.3, p = .14$).

To supplement our findings above, we also ran several control analyses. First, we confirmed that prediction strength only modulated outcome responses *after outcomes actually appeared* –in other words, that the relationship between prediction strength and outcome response was temporally offset. To confirm this, we re-ran the regression analyses where prediction signals were again time-locked to the cue period (TRs 3-4 post-cue onset), but we now systematically varied the time point of the 'outcome,' starting with time points before the outcome appeared through time points after the outcome appeared. Specifically, for each TR (1-8), we ran a separate regression analysis where a single TR was used to index the outcome response. Because TRs 1-4 correspond to volumes before an outcome-related fMRI response should peak, there should not be any outcome-related effects during these TRs. Indeed, relationships between prediction strength and unexpected 'outcomes' only emerged in TRs after the onset appeared (Figure 3D). As an additional test of the temporal relationship between prediction and outcome responses, we also ran a control analysis where we measured 'prediction strength' during the putative outcome window (TRs 5 and 6 post-cue onset) and measured the hippocampal 'outcome response' during the putative prediction window (TRs 3 and 4). With this 'switching' of the windows, there was no relationship between 'prediction strength' and hippocampal 'outcome response' for any of the prediction ROIs (DMN, mPFC, PPC, VisN) for either expected (p's $> .4$) or unexpected (p's $> .45$) trials.

Finally, we ran a new analysis where instead of assessing the relationship between prediction strength and outcome response via linear regression, we divided outcome responses according to three equally sized bins of prediction strength: high, medium, and low. This allowed us to address the relationship between prediction strength and outcome response via repeated measures ANOVA. Qualitatively, and statistically, this analysis approach yielded virtually identical results (Figure 3E). In particular, hippocampal outcome responses varied as a function of prediction strength (high, medium, low) and trial type (expected, near, far trials) in both DMN and mPFC (DMN, $F(4,180) = 2.7$, $p = .03$; mPFC, $F(4,180) = 4.4$, $p = .002$), but not PPC or VisN (PPC, $F(4,180) = .75$, $p = .56$; VisN; $F(4,180) = .97$, $p = .43$). Specifically, for DMN and mPFC, hippocampal outcome responses tended to decrease as a function of prediction strength when outcomes were expected, and to increase when outcomes were near, but unexpected.

Outcome responses in fronto-striatal regions

We have shown that hippocampal mismatch signals are modulated by DMN prediction strength and that this relationship is particularly strong when considering prediction strength within mPFC. While we had an *a priori* interest in outcome responses within the hippocampus, for comparison purposes we also considered outcome responses in two additional regions which have previously been implicated in memory updating: the pars triangularis region of left inferior frontal gyrus (LIFG_t) and caudate. Caudate has been shown to respond to expectancy violations (Schultz et al., 1997; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) and LIFG_t has been shown to respond when episodic memory associations change (Dolan & Fletcher, 1997; Kuhl et al., 2012). Based on these prior findings, we expected that both LIFG_t and caudate would exhibit outcome responses that were related to prediction strength. Of critical interest, however, was whether LIFG_t and/or caudate would be sensitive to the similarity between predictions and outcomes (near vs. far trials).

For both $LIFG_t$ and caudate, there tended to be a positive relationship between mPFC prediction strength and outcome response for unexpected trials and a negative relationship for expected trials (Figure 4). For both regions, this difference in relationship for expected vs. unexpected trials was significant ($LIFG_t$: $t(45) = 2.4$, $p = .02$; caudate $t(45) = 2.2$, $p = .03$). Thus, as with hippocampus, $LIFG_t$ and caudate responses were differentially modulated by mPFC prediction strength as a function of whether outcomes were expected vs. unexpected. However, whereas the relationship between mPFC prediction strength and hippocampal outcome response was significantly stronger for near trials than far trials, there was no significant difference between near vs. far trials for either $LIFG_t$ or caudate ($LIFG_t$, $t(45) = .32$, $p = .75$; caudate, $t(45) = 1.0$, $p = .31$). In fact, for caudate, the relationship between mPFC prediction strength and outcome response was significant for far trials ($t(45) = 2.1$, $p = .04$), but not near trials ($t(45) = .48$, $p = .64$). The relatively greater sensitivity of hippocampus to the type of unexpected outcome was confirmed by region × trial type (near vs. far) interactions: hippocampus vs. $LIFG_t$ ($F(1,45) = 6.2$, $p = .02$), and hippocampus vs. caudate ($F(1,45) = 26.8$, $p < .0001$). Thus, although all three regions (hippocampus, $LIFG_t$, caudate) were sensitive to the difference between unexpected vs. expected outcomes, the hippocampus was uniquely sensitive to differences in similarity among unexpected outcomes.

Post-hoc, voxel-wise whole brain analyses (uncorrected threshold of $p = .001$, 5 voxel extent threshold) confirmed the above findings. A contrast of unexpected vs. expected trials revealed significant clusters in hippocampus, caudate, and IFG (right-lateralized). However, for the comparison of near vs. far trials, a cluster was observed in hippocampus, but not IFG or caudate. Likewise, the comparison of near vs. far trials did not reveal clusters in the ventral striatum, midbrain, or insula–regions that have previously been associated with prediction error, mismatch, and/or expectancy violation signals (Berns, McClure, Pagnoni, & Montague, 2001; Axmacher et al., 2010; Preuschoff, Quartz, & Bossaerts, 2008; Lisman & Grace, 2005; D'Ardenne, McClure, Nystrom, & Cohen, 2008).

Mismatch signals and subsequent memory performance

As a final question, we asked whether mismatch signals in the hippocampus reflect an adaptive learning mechanism. Specifically, we tested whether greater hippocampal responses to unexpected outcomes were associated with better performance on a subsequent memory test. The subsequent memory test, which was conducted after fMRI scanning, probed subjects' memories for the most recent cue-outcome associations. For expected trials, associations never changed and, therefore, the 'most recent' association was also the original association. For unexpected trials, responses on the post-test were divided into 'successful updating' trials (when subjects selected the most recent association) and 'failed updating' trials (when subjects selected the original association; Kuhl et al., 2012).

Subjects selected the most recent associate (successful updating) on the majority of trials (collapsed across high and low confidence, expected trials: M = 92.6%, SD = 7.9%; near trials: M = 72.7%, SD = 19.2%; far trials: M = 69.5%, SD = 18.1%). For near and far trials, 'failed updating' occurred when subjects selected the original association (near: M = 16.3%, SD = 15.5%; far: M = 17.9%, SD = 15.8%). Subsequent memory analyses (i.e., comparing hippocampal activation for subsequent 'successful updating' vs. 'failed updating') were complicated by the relatively low rate of failed updating trials–particularly when also controlling for visual category condition. That

said, we ran two subsequent memory analyses to test for relationships between hippocampal outcome responses and subsequent memory accuracy. In one version, we included all subjects in the analysis and removed conditions, for each subject, that contained an empty cell (either an empty 'successful udating' or 'failure to update' cell). However, there was no reliable subsequent memory for either near trials ($t(34) = -0.1, p = .92$) or far trials ($t(39) = .71, p = .48$). In a second version, we only included subjects that had at least one trial in each cell of the design. This resulted in the inclusion of only 16 of 46 subjects. Again, subsequent memory effects were not observed for near trials ($t(15) = -.54, p = .59$) or far trials ($t(15) = .64, p = .53$).

Given the generally high performance and relatively low number of failed updating trials, we focused instead on potential relationships between hippocampal outcome responses and reaction times (RTs) on the subsequent memory test. We predicted that greater hippocampal outcome responses would be associated with faster RTs ('better' memory). Mean RTs across successful updating trials (high and low confidence combined) were: expected = 2379 ms (SD = 619 ms), near = 3716 ms (SD = 1196 ms), and far = 3679 ms (SD = 941 ms). To test for trial-level relationships between hippocampal outcome responses and memory performance, we ran linear regression analyses where the predictor variable was univariate hippocampal activity measured in response to the outcome (during the updating phase) and the dependent measure was RT during the post-scan memory test. For each subject, separate linear regression analyses were run for each trial type (expected, near, far) and visual category conditions, resulting in 12 separate regressions. Resulting t-statistics were then averaged across visual category conditions. Two subjects who did not complete the post-test were excluded from the analysis.

Consistent with the idea that hippocampal mismatch signals reflect an adaptive mechanism, there was a significant negative relationship between hippocampal outcome responses on unexpected (near and far) trials and post-test RTs ($t(43) = 2.7$, $p < .01$; Figure 5A). There was no significant relationship between hippocampal outcome responses and RTs for expected trials ($t(43) = 1.5, p = .14$), nor was there a difference in the relationship for near vs. far trials ($t(43) = .24$, $p = .81$). To compliment this regression analysis, we also ran separate linear mixed-effects models for unexpected and expected trials to assess the relationship between hippocampal outcome response and post-test RTs. The full model included hippocampal activity and visual category as fixed effects. An intercept for subjects and by-subject random slopes for all fixed effects were included as random effects. Again, we observed a significant main effect of hippocampal outcome activity on post-test RTs for unexpected trials ($\chi^2 = 24.3$, $p < .001$), but now also observed a significant effect for expected trials ($\chi^2 = 12.7$, $p = .01$). We also tested whether post-test RTs were related to prediction strength (as opposed to outcome response) during the updating phase by re-running the regression analyses with DMN prediction strength replacing outcome response. To be clear, for this analysis we were relating reactivation of the *original association* to subsequent retrieval of the *most recent association*. However, DMN prediction strength did not predict post-test RTs (t's < 1, p's > .4). Likewise, there was no relationship between prediction strength and post-test RTs when predictions were indexed by mPFC, PPC or VisN (p's > .4). Thus, although neocortical prediction strength was related to hippocampal outcome responses, and hippocampal outcome responses predicted subsequent memory performance, neocortical prediction strength did not, on its own, predict subsequent memory performance.

To the extent that hippocampal outcome responses are related to successful memory *updating*, better memory for the most recent (new) association may come at the expense of retaining the original (old) association (Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014). We were able to test this idea using data from Experiment 2, as this Experiment (but not Experiment 1) included a two-stage post-test where 'stage 1' probed memory for the new association (as described above) and 'stage 2' probed memory for the original association (see *Methods*). Toward this end, we re-ran the linear regression analyses using Experiment 2 data only. In one set of analyses, the predictor variable was hippocampal outcome responses and the dependent measure was RTs during subsequent successful retrieval of the new association ('stage 1' performance). In a separate set of analyses, the dependent measure was instead RTs during subsequent successful retrieval of the original association ('stage 2' performance). Importantly, we only considered trials for which subjects successfully selected the most recent ('stage 1') *and* the original association ('stage 2'). We also separately considered relationships for near vs. far trials. A 2 × 2 repeated measures ANOVA with factors of trial type (near, far) and memory association (new, original) revealed a significant main effect of memory association ($F(1,20) = 6.4$, $p = .02$) as well as a significant interaction ($F(1,20) = 4.8$, $p = .04$). The main effect of memory association reflected the fact that greater hippocampal outcome responses were associated with relatively *faster* RTs during retrieval of new associations and relatively *slower* RTs during retrieval of original associations (Figure 5B). In other words, hippocampal outcome responses signaled a tradeoff between memory for the new vs. original associations. The interaction reflected the fact that this difference was relatively stronger for near trials compared to far trials, complimenting our finding that hippocampal mismatch signals were relatively more robust for near than far trials. While these analyses are based only on subsequent retrieval speed, and not retrieval accuracy, they are consistent with the idea that hippocampal outcome responses reflect an adaptive mechanism.

Discussion

Here, we tested whether hippocampal responses to unexpected outcomes are sensitive to the strength of neural predictions and to the similarity between predictions and outcomes. We quantified prediction strength by measuring cue-evoked memory reactivation within the default mode network (DMN) and two sub-regions within the DMN: medial prefrontal cortex (mPFC) and posterior parietal cortex (PPC). We report three main findings. First, prediction strength was positively related to hippocampal responses to unexpected outcomes–but not expected outcomes–consistent with the proposed role of the hippocampus as a mismatch detector (Kumaran & Maguire, 2006b; Duncan et al., 2012). Second, hippocampal outcome responses were sensitive to the similarity between predictions and outcomes, particularly when considering predictions derived from mPFC. More specifically, hippocampal outcome responses increased with prediction strength to a greater degree when predictions were similar to outcomes (near trials) compared to when predictions were dissimilar to outcomes (far trials). Finally, hippocampal responses to unexpected outcomes were associated with subsequent behavioral expressions of memory updating, with greater hippocampal outcome responses predicting relatively faster retrieval of new associations and relatively slower retrieval of older associations.

Compared to prior reports of hippocampal mismatch signals (Kumaran & Maguire, 2006a, 2006b; Duncan et al., 2009, 2012; Chen et al., 2015), a critical–and novel–advantage of our experimental approach is that we directly measured, on a trial-by-trial basis, neural prediction strength. Although the hippocampus is known to act as a novelty detector (Stern et al., 1996; Strange, Fletcher, Henson, Friston, & Dolan, 1999; Ranganath & Rainer, 2003), the mismatch signal is not thought to simply reflect associative novelty, but to reflect a comparison between an *actively generated prediction* and a new, but unexpected associative outcome (Kumaran & Maguire, 2007). Consistent with this perspective, we found that hippocampal outcome responses increased as a function of prediction strength for unexpected trials, but not expected trials. In other words, hippocampal outcome responses did not simply scale with prediction strength, but were instead sensitive to whether predictions matched outcomes. This dissociation was observed when predictions were derived from DMN, mPFC, or PPC, but the dissociation was most compelling for mPFC predictions where a significant, positive relationship was observed for unexpected trials and a numerically negative relationship was observed for expected trials. Importantly, we did not find differences in hippocampal outcome responses to unexpected vs. expected outcomes when prediction strength was not taken into account, confirming that hippocampal outcome responses reflected a combination of strong predictions and a violation of those predictions.

We also found that hippocampal outcome responses were sensitive to the similarity between predictions and outcomes. Considering predictions derived from DMN, mPFC, or PPC we found a significant relationship between prediction strength and hippocampal outcome response for 'near' outcomes but not 'far' outcomes. Again, this dissociation was particularly evident when considering mPFC predictions, with a significantly stronger relationship for near trials compared to far trials. Thus, hippocampal mismatch signals were greatest when outcomes were close–but not identical–to predictions. Potentially, this finding is related to the proposed role of the hippocampus in disambiguating similar stimuli (Leutgeb et al., 2007; McNaughton & Morris, 1987; Marr, 1971; O'Reilly & McClelland, 1994; Yassa & Stark, 2011). Recently, we have shown that hippocampal activity patterns become differentiated when stimuli are highly overlapping (Favila et al., 2016). However, whether hippocampal mismatch signals are directly related to disambiguation of hippocampal activity patterns remains an open question and such a relationship could take multiple forms. On the one hand, when similar representations are successfully disambiguated or pattern separated, this may facilitate mismatch detection (I. Lee, Hunsacker, & Kesner, 2005). On the other hand, detection of 'near' mismatches may drive the hippocampus into a 'pattern separating' state (Duncan et al., 2012) that results in disambiguation. In either case, there may be a direct relationship between hippocampal mismatch detection and the disambiguation of similar stimuli. One caveat, however, in relating the present findings to prior evidence of hippocampal disambiguation of similar events (Bakker et al., 2008; Favila et al., 2016; Hulbert & Norman, 2015) is that our 'near' condition only referred to stimuli from a common visual category, whereas prior studies have considered stimuli with much stronger perceptual overlap. Thus, an informative follow-up to the present work would be to consider the relationship between prediction strength and hippocampal outcome responses across a wider range of similarities between predictions and outcomes (Lacy, Yassa, Stark, Muftuler, & Stark, 2011; Duncan et al., 2012).

As a point of comparison, we also considered outcome responses in caudate and $LIFG_t$. As with hippocampal

outcome responses, we found that caudate and $LIFG_t$ outcome responses tended to increase with mPFC prediction strength when outcomes were unexpected, but to decrease with prediction strength when outcomes were expected. The negative relationship for expected trials, which was marginally significant for both caudate and $LIFG_t$ (Figure 4), potentially reflects a form of repetition suppression that occurs when outcomes match actively-held predictions (Miller, Li, & Desimone, 1991; Wiggs & Martin, 1998; Henson & Rugg, 2003; Grill-Spector, Henson, & Martin, 2006; Meyer & Olson, 2011; Klein-Flügge, Barron, Brodersen, Dolan, & Behrens, 2013; Boorman, Rajendran, OReilly, & Behrens, 2016). On the other hand, the tendency for activity to increase as a function of prediction strength on unexpected trials is consistent with prior evidence relating caudate responses to expectancy violations (Schultz et al., 1997; Daw & Doya, 2006; Daw & Shohamy, 2008) and $LIFG_t$ activity to changes in mnemonic associations (Dolan & Fletcher, 1997; Kuhl et al., 2012). However, in contrast to the hippocampus, neither caudate or $LIFG_t$ were sensitive to the similarity between predictions and outcomes (Figure 4), with no differences in the strength of relationships between prediction strength and outcome responses for near vs. far trials. Indeed, the hippocampus was significantly more sensitive to the difference between near vs. far trials than either caudate or $LIFG_t$. Thus, while hippocampus, caudate, and $LIFG_t$ may each play a role in updating mnemonic associations, our findings point to a qualitative difference across these regions, with hippocampus uniquely sensitive to the similarity between predictions and outcome.

By considering performance on the post-scan memory test, we were able to assess whether hippocampal outcome responses reflected an adaptive learning mechanism. While we did not find relationships between hippocampal outcome responses and subsequent memory accuracy, we did find relationships with subsequent reaction times. Namely, greater hippocampal outcome responses on unexpected trials predicted faster reaction times during *subsequent retrieval* of the new (updated) association. Interestingly, we observed a significantly different relationship when we considered subsequent retrieval of the older (original) associations, with greater hippocampal outcome responses tending to predict *slower* retrieval of the older associations. These findings are consistent with the idea that hippocampal responses to unexpected outcomes reflect an adaptive tradeoff that biases memory toward new, relevant associations and away from older, irrelevant associations (Kim et al., 2014).

To measure memory-based predictions, we targeted the DMN based on theoretical proposals that the DMN actively represents memory-based predictions (Bar, 2007, 2009) and recent evidence that activity patterns in the DMN and its subregions reflect the contents of memory retrieval (Chen et al., 2016; Kuhl & Chun, 2014; Richter et al., 2016). Our findings strongly reinforce both of these points, while also providing new insight into the functional significance of predictions carried by the DMN. Although memory reactivation has most typically been studied in visual cortical areas (e.g. Polyn et al., 2005; Kuhl et al., 2011), we observed a double dissociation between the DMN and the visual network (VisN), with stronger reactivation (predictions) in the DMN than the VisN and stronger decoding of outcomes (perception) in the VisN than the DMN. Moreover, we found that predictions derived from the DMN and its subregions–but not predictions from VisN–were related to hippocampal outcome responses. These dissociations between the DMN and the VisN add to prior evidence of functional dissociations in reactivation across visual and fronto-parietal regions (Kuhl et al., 2013). Our findings also specifically highlight mPFC as an important

'prediction region' in relation to hippocampal outcome responses. The current mPFC findings are highly consistent with prior evidence that mPFC represents older memories in relation to new learning (Richter et al., 2016) and that mPFC interacts with the hippocampus during memory updating (van Kesteren, Fernández, Norris, & Hermans, 2010; Zeithamova et al., 2012).

Finally, while our findings suggest that predictions within the DMN may be particularly relevant to hippocampal mismatch signals, it is important to emphasize that the predictions we decoded from DMN activity patterns were presumably triggered by hippocampal pattern completion processes (O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001; Staresina, Henson, Kriegeskorte, & Alink, 2012; Rolls, 2013; Hindy et al., 2016). Thus, the hippocampus may play a critical role both in generating predictions and comparing predictions to outcomes (Hasselmo & Wyble, 1997; Lisman & Grace, 2005; Kumaran & Maguire, 2007; Chen et al., 2015). That said, when we directly decoded prediction strength from the hippocampus, it was not related to univariate hippocampal outcome responses. On the one hand, this null result may simply reflect the difficulty of decoding predictions from the hippocampus (Mack & Preston, 2016). Alternatively, it is possible that the DMN plays an important role in transforming or processing hippocampal predictions before they are fed back to the hippocampus. While future work will be required to tease apart these possibilities–which would benefit from methods with more precise temporal resolution–the present work establishes an important relationship between the predictions carried by the DMN and mismatch signals within the hippocampus.

Acknowledgments We thank Marvin Chun for helpful discussion and feedback. We thank Sam Cartmell for assistance with data collection. This work was supported by NIH Grant NS089729.

References

Anderson, M. C., Bunce, J. G., & Barbas, H. (2015). Prefrontal–hippocampal pathways underlying inhibitory control over memory. *Neurobiology of learning and memory*.

Axmacher, N., Cohen, M. X., Fell, J., Haupt, S., Dümpelmann, M., Elger, C. E., … Ranganath, C. (2010). Intracra- nial eeg correlates of expectancy and memory formation in the human hippocampus and nucleus accumbens. *Neuron*, *65* (4), 541–549.

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. (2008). Pattern separation in the human hippocampal ca3 and dentate gyrus. *Science*, *319* (5870), 1640–1642.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11* (7), 280–289.

Bar, M. (2009). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364* (1521), 1235–1243.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The journal of neuroscience*, *21* (8), 2793–2798.

Boorman, E. D., Rajendran, V. G., OReilly, J. X., & Behrens, T. E. (2016). Two anatomically and computationally distinct learning signals predict changes to stimulus-outcome associations in hippocampus. *Neuron*.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, *61*, 27–48.

Chen, J., Cook, P. A., & Wagner, A. D. (2015). Prediction strength modulates responses in human area ca1 to sequence violations. *Journal of neurophysiology*, *114*(2), 1227–1238.

Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (2016). Shared experience, shared memory: a common structure for brain activity during naturalistic recall. *bioRxiv*, 035931.

Chen, J., Olsen, R. K., Preston, A. R., Glover, G. H., & Wagner, A. D. (2011). Associative retrieval processes in the human medial temporal lobe: Hippocampal retrieval success and ca1 mismatch detection. *Learning & Memory*, *18*(8), 523–528.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). Bold responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267.

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: the role of prediction and context. *Trends in Cognitive Sciences*, *19*(2), 1–8.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593-620.

Demblon, J., Bahri, M. A., & D'Argembeau, A. (2016). Neural correlates of event clusters in past and future thoughts: How the brain integrates specific episodes with autobiographical knowledge. *NeuroImage*, *127*, 257–266.

Dolan, R., & Fletcher, P. (1997). Dissociating prefrontal and hippocampal function in episodic memory encoding. *Nature*, *388*(6642), 582–585.

Duncan, K., Curtis, C., & Davachi, L. (2009). Distinct memory signatures in the hippocampus: intentional states distinguish match and mismatch enhancement signals. *The Journal of Neuroscience*, *29*(1), 131–139.

Duncan, K., Ketz, N., Inati, S. J., & Davachi, L. (2012). Evidence for area ca1 as a match/mismatch detector: a high-resolution fmri study of the human hippocampus. *Hippocampus*, *22*(3), 389–398.

Düzel, E., Habib, R., Rotte, M., Guderian, S., Tulving, E., & Heinze, H.-J. (2003). Human hippocampal and parahippocampal activity during visual associative recognition memory for spatial and nonspatial stimulus configurations. *The Journal of Neuroscience*, *23*(28), 9439–9444.

Eichenbaum, H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*(1), 109–120.

Eichenbaum, H., & Fortin, N. J. (2009). The neurobiology of memory based predictions. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1183–1191.

671 Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and
672 decision making. *Neuron*, *76*(6), 1057–1070.
673 Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). Liblinear: A library for large linear
674 classification. *The Journal of Machine Learning Research*, *9*, 1871-1874.
675 Favila, S. E., Chanales, A. J., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation
676 prevents interference during subsequent learning. *Nature Communications*, *7*.
677 Fiorillo, C., Tobler, P., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine
678 neurons. *Science*, *299*(5614), 1898-1902.
679 Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory
680 misattribution. *The Journal of Neuroscience*, *33*(20), 8590 - 8595.
681 Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory.
682
683 *Hippocampus*, *3*(4), 491–516.
684
685 Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific
686 effects. *Trends in Cognitive Sciences*, *10*(1), 14–23.
687 Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus:
688
689 simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1–2), 1–34. Henson,
690 R., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural
691 priming. *Neuropsychologia*, *41*(3), 263-70.
692
693 Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive
694 coding in visual cortex. *Nature Neuroscience*, *19*(5).
695 Hulbert, J., & Norman, K. (2015). Neural differentiation tracks improved recall of competing memories following
696 interleaved study and retrieval practice. *Cerebral Cortex*, *25*(10), 3994–4008.
697 Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Reward-predicting activity of dopamine and caudate neuronsa
698 possible mechanism of motivational control of saccadic eye movement. *Journal of Neurophysiology*, *91*(2),
699 1013–1024.
700
701 Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based
702 prediction error. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24).
703 Klein-Flügge, M. C., Barron, H. C., Brodersen, K. H., Dolan, R. J., & Behrens, T. E. J. (2013). Segregated encoding of
704 reward–identity and stimulus–reward associations in human orbitofrontal cortex. *The Journal of Neuroscience*,
705 *33*(7), 3202–3211.
706 Kok, P., Jehee, J. F., & de Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual
707 cortex. *Neuron*, *75*(2), 265–270.
708 Kroes, M. C., & Fernández, G. (2012). Dynamic neural systems enable adaptive, flexible memories. *Neuroscience &*
709 *Biobehavioral Reviews*, *36*(7), 1646–1666.
710 Kuhl, B. A., Bainbridge, W. A., & Chun, M. M. (2012). Neural reactivation reveals mechanisms for updating
711 memory. *The Journal of Neuroscience*, *32*(10), 3453–3461.

712 Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal
713 cortex. *The Journal Of Neuroscience*, *34*(23), 8051–8060.
714 Kuhl, B. A., Johnson, M. K., & Chun, M. M. (2013). Dissociable neural mechanisms for goal-directed versus
715 incidental memory reactivation. *The Journal of Neuroscience*, *33*(41), 16099–16109.
716 Kuhl, B. A., Rissman, J., Chun, M., & Wagner, A. (2011). Fidelity of neural reactivation reveals competition
717 between memories. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(14),
718 5903-5908.
719
720 Kumaran, D., & Maguire, E. A. (2006a). The dynamics of hippocampal activation during encoding of overlapping
721 sequences. *Neuron*, *49*(4), 617–629.
722 Kumaran, D., & Maguire, E. A. (2006b). An unexpected sequence of events: mismatch detection in the human
723 hippocampus. *PLoS Biol*, *4*(12), e424.
724 Kumaran, D., & Maguire, E. A. (2007). Match–mismatch processes underlie human hippocampal responses to
725 associative novelty. *The Journal of Neuroscience*, *27*(32), 8517–8524.
726 Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T., & Stark, C. E. (2011). Distinct pattern separation related transfer
727 functions in human ca3/dentate and ca1 revealed using high-resolution fmri and variable mnemonic similarity.
728 *Learning & Memory*, *18*(1), 15–18.
729 Lee, H., Chun, M. M., & Kuhl, B. A. (2016). Lower parietal encoding activation is associated with sharper information and
730 better memory. *Cerebral Cortex*, bhw097.
731 Lee, I., Hunsacker, M., & Kesner, R. (2005). The role of hippocampal subregions in detecting spatial novelty.
732
733 *Behavioral Neuroscience*, *119*, 145–53.
734
735 Leutgeb, J., Leutgeb, S., Moser, M., & Moser, E. (2007). Pattern Separation in the Dentate Gyrus and CA3 of the
736
737 Hippocampus. *Science*, *315*(5814), 961.
738
739 Lisman, J., & Grace, A. (2005). The hippocampal-vta loop: Controlling the entry of information into long-term
740 memory. *Neuron*, *46*, 703–13.
741 Mack, M. L., & Preston, A. R. (2016). Decisions about the past are guided by reinstatement of specific memories in the
742 hippocampus and perirhinal cortex. *NeuroImage*, *127*, 144–157.
743 Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society B:*
744 *Biological Sciences*, *262*(841), 23-81.
745 McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a
746 distributed memory system. *Trends in Neurosciences*, *10*, 408-415.
747 Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex.
748
749 *Proceedings of the National Academy of Sciences of the United States of America*, *108*(48), 19401–19406. Miller, E. K., Li,
750 L., & Desimone, R. (1991). A neural mechanism for working and recognition memory in inferior
751 temporal cortex. *Science*, *254*(5036), 1377–1379.
752
753 Norman, K. A., Newman, E., Detre, G., & Polyn, S. M. (2006). How inhibitory oscillations can train neural networks and
754 punish competitors. *Neural Computation*, *18*, 1577–1610.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus*, *4*(6), 661–682.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311 - 345.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*, 1963–1966.

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764–R773.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *The Journal of neuroscience*, *28*(11), 2745–2752.

Rajasethupathy, P., Sankaran, S., Marshel, J. H., Kim, C. K., Ferenczi, E., Lee, S. Y., ... others (2015). Projections from neocortex mediate top-down control of memory retrieval. *Nature*, *526*, 653-659.

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*(3), 193–202.

Richter, F. R., Chanales, A. J., & Kuhl, B. A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, *124*, 323–335.

Rolls, E. T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Front. Syst. Neurosci*, *7*(74), 10–3389.

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current opinion in behavioral sciences*, *1*, 1–8.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Staresina, B. P., Henson, R. N., Kriegeskorte, N., & Alink, A. (2012). Episodic reinstatement in the medial temporal lobe. *The Journal of Neuroscience*, *32*(50), 18150–18156.

Stern, C. E., Corkin, S., Gonzalez, R. G., Guimaraes, A. R., Baker, J. R., Jennings, P. J., ... Rosen, B. R. (1996, Aug). The hippocampal formation participates in novel picture encoding: Evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(16), 8660-8665.

Strange, B., Fletcher, P., Henson, R., Friston, K., & Dolan, R. (1999). Segregating the functions of human hippocam- pus. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(7), 4034–4039.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., ... Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289.

van Kesteren, M. T., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceed- ings of the National Academy of Sciences of the United States of America*, *107*(16), 7550–7555.

Wiggs, C. L., & Martin, A. (1998, Apr). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, *8*(2), 227-33.

Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … others (2011). The orga- nization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, *106*(3), 1125–1165.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179.

Figure 1 **Experimental paradigm.** Subjects first completed three acquisition rounds during which they learned word-picture (cue-outcome) associations. The trial structure for the first two acquisition rounds is shown above; the third acquisition round had a different trial structure and timing parameters (see Methods). Acquisition rounds 1 and 2 were collected prior to fMRI scanning; acquisition round 3 was conducted during fMRI scanning. Following the acquisition rounds, subjects began the scanned updating phase. Subjects again studied words paired with pictures, however presentation of the words (cues) and pictures (outcomes) was now separated by 4 s. There were three trial types during the updating phase: expected (an outcome from the acquisition phase was repeated during the updating phase), near (an outcome from the updating phase was replaced with a new picture from the same visual category as the original outcome), and far (an outcome from the acquisition phase was replaced with a new picture from a different visual category as the original outcome). Pattern classifiers were trained to discriminate visual categories using data from the 3rd acquisition round and were separately tested on cue and outcome components from the updating-phase.

Figure 2 **Decoding predicted and perceived outcomes**. **(A)** Pattern classification analyses were applied to two networks: the default mode network (DMN, blue) and the visual network (VisN, orange). **(B)** Classification accuracy for predicted outcomes was greater in the DMN than the VisN (left panel) whereas classification accuracy for perceived outcomes was greater in the VisN than the DMN (right panel). Error bars represent standard error of the mean. * $p < .05$.

Figure 3 **Relationship between prediction strength and hippocampal outcome responses. (A)** We assessed prediction strength (reactivation of the original associate) in the default mode network (DMN), two sub-regions of the DMN [medial prefrontal cortex (mPFC) and posterior parietal cortex (PPC, which included lateral and medial portions, but only lateral PPC is shown here)], and the visual network (VisN). **(B-C)** Both figures show results from linear regressions relating classifier evidence for the predicted outcome to univariate hippocampal response amplitude to the actual outcome. Each row corresponds to a different 'prediction region,' as shown in (A). **(B)** The relationship between prediction strength and hippocampal outcome responses for expected and unexpected trials. **(C)** The relationship between prediction strength and hippocampal outcome responses for the two sub-types of unexpected trials: near and far trials. **(D)** Prediction strength measured during TRs 3-4 was used to predict the hippocampal outcome response for each trial type (expected, unexpected- near, unexpected-far) at each TR. The vertical grey line indicates the time point at which outcomes were shown. Asterisks denote significant one-way ANOVAs across trial type (expected, near, far) at each TR. Note: significant effects of trial type only occurred after outcomes were shown. **(E)** Hippocampal outcome activation (y-axis) binned according to prediction strength (x-axis), separately for each of the outcome conditions. Significant interactions between condition (expected, unexpected- near, unexpected-far) and prediction strength (low, medium, high) were observed for DMN and mPFC (p's $< .05$). Error bars represent standard error of the mean. ~$p < .10$, * $p < .05$, ** $p < .01$.

Figure 4 **Relationship between mPFC prediction strength and outcome responses in LIFG$_t$ and caudate.** Linear regression analyses were applied in which mPFC prediction strength was used to predict univariate outcome responses in LIFG$_t$ and caudate. Relationships are separately shown for expected and unexpected trials (left column), and for the unexpected trial sub-types: near and far trials (right column). LIFG$_t$ and caudate were each sensitive to the difference between expected vs. unexpected trials, but not to the difference between near vs. far unexpected trials. Error bars represent standard error of the mean. $\tilde{}p < .10, *p < .05$.

Figure 5 **Relationships between hippocampal outcome responses and reaction times during post-scan memory test**. Linear regressions were used to test for relationships between hippocampal outcome responses and reaction times (RTs) on a subsequent memory test (correct trials only). **(A)** For unexpected trials, there was a negative relationship between hippocampal outcome responses and subsequent RTs for retrieval of new (updated) associations indicating that greater hippocampal outcome responses corresponded to faster subsequent retrieval of the new associations. **(B)** Using data from Experiment 2 only, linear regression analyses tested for relationships between hippocampal outcome responses and subsequent RTs for retrieval of the new association (left panel) and the original association (right panel). Greater hippocampal outcome responses were associated with relatively faster RTs for subsequent retrieval of new associations but relatively slower RTs for retrieval of original associations (main effect of association: $F(1,20) = 6.4$, $p = .02$). This difference was stronger for near than far trials, as reflected by a significant interaction between association (new, original) and trial type (near, far; $F(1,20) = 4.8$, $p = .04$). Error bars represent standard error of the mean. $*p < .05, ** p < .01$.