Published in final edited form as:
Neuroimage. 2016 January 1; 124(0 0): 323–335. doi:10.1016/j.neuroimage.2015.08.051.

Predicting the integration of overlapping memories by decoding mnemonic processing states during learning

Franziska R. Richter[a], **Avi J. H. Chanales**[a], and **Brice A. Kuhl**[a,b]

[a] Department of Psychology, New York University, 10003 NY, NY

[b] Center for Neural Science, New York University, 10003 NY, NY

Abstract

The hippocampal memory system is thought to alternate between two opposing processing states: encoding and retrieval. When present experience overlaps with past experience, this creates a potential tradeoff between encoding the present and retrieving the past. This tradeoff may be resolved by memory *integration*—that is, by forming a mnemonic representation that links present experience with overlapping past experience. Here, we used fMRI decoding analyses to predict *when*—and establish *how*—past and present experiences become integrated in memory. In an initial experiment, we alternately instructed subjects to adopt encoding, retrieval or integration states during overlapping learning. We then trained across-subject pattern classifiers to 'read out' the instructed processing states from fMRI activity patterns. We show that an integration state was clearly dissociable from encoding or retrieval states. Moreover, trial-by-trial fluctuations in decoded evidence for an integration state during learning reliably predicted behavioral expressions of successful memory integration. Strikingly, the decoding algorithm also successfully predicted specific instances of spontaneous memory integration in an entirely independent sample of subjects for whom processing state instructions were not administered. Finally, we show that medial prefrontal cortex and hippocampus differentially contribute to encoding, retrieval, and integration states: whereas hippocampus signals the tradeoff between encoding vs. retrieval states, medial prefrontal cortex actively represents past experience in relation to new learning.

Keywords

integration; MVPA; reinstatement; episodic memory; hippocampus; medial prefrontal cortex

1 INTRODUCTION

The hippocampal memory system is thought to alternate between two opposing processing states: encoding and retrieval. The idea of opposing encoding and retrieval states is central

Corresponding author: Franziska R. Richter, who is now at the Department of Psychology, University of Cambridge, Cambridge, UK. frr28@cam.ac.uk..

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest: None

to computational models of episodic memory (Hasselmo, Bodelón, & Wyble, 2002; O'Reilly & McClelland, 1994) and is supported by experimental evidence across levels of analysis. For example, encoding and retrieval are associated with distinct electrophysiological activity states in rodents (Douchamps et al., 2013; Hasselmo et al., 2002; Kunec, Hasselmo, & Kopell, 2005; Siegle & Wilson, 2014) and humans (Rizzuto et al., 2006), and human fMRI studies have identified distinct activity patterns corresponding to encoding and retrieval (Donaldson et al., 2001; Duncan, Tompary, & Davachi, 2014; Eldridge et al., 2005). However, the opposition between encoding and retrieval states poses an important problem whenever new learning overlaps with past experience (O'Reilly & McClelland, 1994). In such cases, the overlap can trigger retrieval of past experience (Kuhl et al., 2010), creating a potential tradeoff between remembering the past and encoding the present. Indeed, understanding the tradeoff between encoding and retrieval states during the learning of overlapping experiences has been of central interest to computational models of episodic memory (O'Reilly & McClelland, 1994).

One way to avoid a tradeoff between encoding and retrieval is by *integrating* present experience into existing memories of past experience. For example, a present conversation with a friend may trigger the retrieval of a past conversation with that friend; integration achieves a balance between remembering this past conversation and encoding the present conversation by allowing the present conversation to be incorporated into an existing representation of the past conversation. Memory integration has important behavioral consequences: it can allow for novel inferences concerning the relationship between temporally discrete events (Preston & Eichenbaum, 2013; Zeithamova, Dominick, & Preston, 2012), it has been associated with reduced interference-related forgetting (Anderson & McCulloch, 1999), and it can facilitate new learning (Schlichting & Preston, 2014; Tse et al., 2007). But when and how do memories become integrated?

Neuroimaging studies have shown that memory integration occurs ‘online’— that is, during new learning (Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012; Zeithamova & Preston, 2010) —and that it is related to the reactivation of past experience during new learning (Zeithamova et al., 2012). Behavioral studies have shown that subtle manipulations of learning context can influence the probability that present experience will be integrated with past experience by altering the relative balance between encoding and retrieval states (Duncan, Sadanand, & Davachi, 2012). Intuitively, integration requires avoiding a processing state that is either ‘pure encoding’ or ‘pure retrieval’ as both extremes would prevent present experience from being *related to* past experience. Here, we asked whether memory integration is associated with a processing state during learning that can be discriminated from encoding and retrieval states based on neural activity patterns. To the extent that mnemonic processing states can be ‘read out’ from neural activity patterns, can these read-outs be used to predict the specific experiences that will become integrated in memory?

We conducted a human fMRI experiment in which subjects learned initial (old) associations followed by overlapping (new) associations. During learning of the new associations we provided instructions that alternately biased subjects’ processing toward encoding of current experience (the new association), retrieval of past experience (the old association), or integration of past with present. We used pattern classification analyses to test whether the processing states (encoding, retrieval, integration) elicited discriminable (i.e., decodable) patterns of neural activity. As noted above, prior evidence indicates that encoding and retrieval are associated with distinct profiles of neural activity (Donaldson et al., 2001; Douchamps et al., 2013; Duncan et al., 2014; Eldridge et al., 2005; Hasselmo et al., 2002; Kunec et al., 2005; Rizzuto et al., 2006; Siegle & Wilson, 2014), but these studies have not directly attempted to read out processing states from patterns of neural activity on individual learning trials. More importantly, prior studies have not tested whether an 'integration state' can be discriminated from encoding and/or retrieval states.

Critically, to the extent that an integration state could be discriminated from encoding/ retrieval states, we sought to validate this result by relating it to behavior. To this end, we derived from our pattern classifier the strength of evidence for an integration processing state *on each learning trial* and then asked whether that evidence predicted performance on a subsequent (and un-anticipated) behavioral integration test. As an even stronger validation step, we also collected data in an additional sample of subjects that completed the same learning paradigm except that we did not instruct/bias processing states during learning. Rather, any fluctuations in processing states were completely subject-driven. This allowed us to test whether a classifier that was trained on data from the first sample of subjects ('instructed subjects') would successfully transfer to the second sample of subjects ('uninstructed subjects'). That is, could we predict specific instances of memory integration in the uninstructed subjects based on what the classifier learned from the instructed subjects? This allowed us to test whether spontaneous memory integration is associated with a pattern of neural activity that generalizes across subjects.

In separate analyses, we also measured (again, using decoding methods) the degree to which older memories were reactivated during new learning and tested whether reactivation predicted memory integration (Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012; Zeithamova et al., 2012). Finally, although our primary analyses were based on whole-brain pattern classification analyses, we also report targeted, secondary analyses that compared regions that have previously been implicated in memory integration—i.e., medial prefrontal cortex (MPFC) and the hippocampus (Shohamy & Wagner, 2008; van Kesteren et al., 2013; Zeithamova et al., 2012)—in order to clarify their respective contributions to memory integration.

2 METHODS

2.1 Participants

Twenty-one subjects (17 female; mean age = 23.04) participated in the 'instructed' version of the experiment and another 8 (6 female; mean age = 21.13) participated in the 'uninstructed' version. Two additional subjects (one instructed, one uninstructed) were excluded due to technical errors. Of the 21 instructed subjects, one was excluded only from analyses related to the direct association test (see below) due to an error saving the data. Subjects were recruited from the New York University community, were 18–35 years of age, right-handed, native English speakers, had normal or corrected-to-normal vision, and had no history of neurological disorders. Informed consent was obtained according to procedures approved by the New York University Committee on Activities Involving Human Subjects. Subjects received payment for their involvement in the study.

2.2 Materials

Stimuli consisted of 144 words and 288 pictures. Word length ranged from 3 to 11 letters (*M* = 5.95). The pictures consisted of photographs of famous people (e.g., Tom Cruise; *faces*), famous locations (e.g., Taj Mahal; *scenes*), and common objects (e.g., wrench; *objects*). All word-picture pairings and the assignment of words and pictures to conditions were randomized for each subject.

2.3 Procedures

Both the ‘instructed’ and the ‘uninstructed’ versions of the experiment consisted of four phases: acquisition, new learning, direct association test, and integration test. Only the acquisition and new learning phases were conducted during fMRI scanning. As detailed below, the only differences between the instructed and uninstructed versions were in (a) the instructions and trial timing during the new learning phase and (b) the order of the direct association and integration tests.

Acquisition and new learning phases—Subjects completed 8 fMRI scan runs, with each run consisting of an acquisition round followed by a new learning round. None of the materials (words or pictures) repeated across scan runs. In acquisition rounds, subjects studied associations between words and pictures. Pictures were drawn from three visual categories: faces, scenes, or objects. Each trial (4s) consisted of a word presented directly above a picture. After presentation of the word-picture pair there was an 8s inter-trial interval (ITI) which included a fixation cross followed by presentation of three single-digit numbers and then another fixation cross. For each number that was presented, subjects were required to indicate via button-press whether it was odd or even. This task was included in order to reduce continued rehearsal of the pairs during the ITI. There were a total of 18 trials in each acquisition round and the procedures for this round were identical across the instructed and uninstructed versions of the experiment.

After each acquisition round, subjects completed a new learning round. For the instructed version, a screen first instructed subjects to “Get Ready” (10s), followed by a reminder of the shape-to-instruction mappings (8s) and a fixation cross (4s). For the uninstructed version, a “Get Ready” screen (6s) was followed by a fixation cross (4s). For both versions of the experiment, the new learning round began immediately after the fixation cross. In each new learning round, all of the words from the immediately preceding acquisition round were presented again, but were paired with a new picture. The ‘new’ picture presented with each word was always from a different category than the ‘old’ picture that had appeared with that word in the acquisition round. In the instructed version, each word-picture pair was presented for 2s and was followed by a shape cue (square, circle, or triangle), which remained on the screen for 6s. The shape cue instructed subjects to either rehearse the old association only (*retrieve* condition), rehearse the new association only (*encode* condition), or rehearse and try to link the new and old associations (*integrate* condition). The assignment of shape cues to instructions was counterbalanced across participants. Subjects had time to memorize the shape-to-instruction assignment prior to entering the scanner. In the uninstructed version, each word-picture pair was presented for 4s and was not followed by a shape cue. Rather, subjects were simply instructed to learn each pair for a later test (equivalent to the encoding condition for the instructed subjects). For both the instructed and uninstructed versions of the experiment, there was an 8s ITI between trials, identical to the acquisition phase (including the odd/even task).

Post Tests—Upon exiting the scanner, subjects completed one post-test that measured memory for the previously studied word-picture pairs (direct association test) and another that measured the ability to 'remember' *across* pairs (integration test).

In the direct association test, each trial presented subjects with a word directly above 6 picture options. One of the pictures had previously been paired with the given word (target) and the other 5 pictures had been paired with a different word (alternatives). Each trial *either* tested memory for an old pair (from the acquisition rounds) or a new pair (from the new learning rounds). Old and new trials were pseudo-randomly intermixed, but the distinction between old and new trials was not explicitly relevant to subjects: their task was simply to choose which picture, from the set of 6 options, had previously been paired with the word. For trials that tested an old pair, all 5 alternative pictures were old pictures from the same visual category as the target. Likewise, for trials that tested a new pair, all 5 alternative pictures were new pictures from the same visual category as the target. Subjects had 6s to select the target picture via mouse click on each trial.

The integration test assessed subjects' ability to link pictures that shared a common word cue. On each trial of the integration test, subjects were presented with a picture from the new learning phase and attempted to 'remember' the old picture that shared a word cue with the new picture, despite never having studied the old and new pictures together and not having been warned that their memory would be tested in this way. Each trial consisted of two steps. In the first step (*category memory*) participants were presented with three category labels (face, scene, object) beneath the new picture and had to select the visual category to which the corresponding old picture belonged. Subjects had 4s to make the category choice via mouse click. Immediately after their response (or when 4s elapsed) subjects were shown a set of 4 pictures that tested their *item memory* (1 target + 3 alternatives). Here, subjects were required to select the specific old picture (target) that was indirectly associated with the new picture (i.e., the old picture that shared a word cue with the new picture). The 3 alternatives were always from the same category as the target picture and were drawn from the set of old pictures. The item memory step was included irrespective of whether subjects selected the correct category label in the prior step. The specific pictures displayed during the item step were independent of the participants' accuracy on the category step. That is, if a subject selected 'face' at the category step, but the target was in fact a 'scene,' then the subject would be shown 4 'scenes' (i.e., pictures from the correct category) at the item step. Subjects had 3 s to choose the correct picture via mouse click. Note: the time limits placed on the integration test were challenging, but were intended to reduce the probability that subjects would 'solve' these trials by separately recalling individual pairs as opposed to recalling pre-existing integrated representations.

In the instructed version of the experiment, the direct association test preceded the integration test; while this means that the direct association test may have, in some way, influenced performance on the integration test, this should only have served as a source of noise that would work against our ability to predict performance on the integration test. For the uninstructed version, because we were specifically interested in predicting performance on the integration test (in order to replicate a finding from the instructed version), we reversed the order and conducted the integration test before the direct association test to reduce any potential influence that the direct association test might have on integration test performance.

2.4 fMRI acquisition

fMRI scanning was performed on the 3T Siemens Allegra head-only scanner at the Center for Brain Imaging at New York University using a Siemens head coil. Structural images were collected using a T1-weighted protocol (256 × 256 matrix, 176 1-mm sagittal slices). Functional images were acquired parallel to the anterior commissure–posterior commissure axis using a single-shot EPI sequence (repetition time = 2 s; echo time = 30 ms; field of view = 192 × 240 mm, flip angle = 82 degrees, bandwidth = 4,165 Hz/px and echo spacing = 0.31 ms). For all functional scanning, we obtained 35 contiguous oblique-axial slices (3 × 3 × 3-mm voxels) per volume. Field map and calibration scans were used to improve functional-to-anatomical image co-registration.

Acquisition and new learning rounds occurred in alternation, with each fMRI scan (block) consisting of one round of acquisition followed by one round of new learning. In the instructed version of the experiment a total of 268 volumes were collected (8m 36s) during each block. Of the 268 volumes, the first 5 were discarded, the next 108 corresponded to the acquisition round, the next 11 included a momentary break and a reminder of the mapping of shapes to instructions (5 volumes for a "Get Ready" screen, 4 volumes for instructions, 2 volumes for a fixation cross), and the final 144 corresponded to the new learning round. For the uninstructed participants a total of 226 volumes were collected in each block (7m 32s). Of the 226 volumes, the first 5 were discarded, the next 108 corresponded to the acquisition round, the next 5 included a momentary break between the acquisition and new learning rounds (3 volumes for a "Get Ready" screen, 2 volumes for a fixation cross), and the final 108 corresponded to the new learning round. Note: fewer volumes separated the acquisition and new learning phases for uninstructed participants than instructed participants, as there were no shape-to-instruction mappings of which to remind the uninstructed subjects. Likewise, the length of the new learning round was shorter for the uninstructed subjects because the instruction cues were not presented.

2.5 fMRI preprocessing

Data preprocessing and analysis was performed using SPM8 (Wellcome Department of Cognitive Neurology, London, United Kingdom), FSL (FMRIB's Software Library, Oxford, United Kingdom) and custom Matlab (The MathWorks, Natick, MA) routines. Preprocessing procedures involved corrections for head motion, coregistration of functional to anatomical images (using a registration procedure that aligned both functional and anatomical images to a calibration scan), an unwarping procedure, normalization to the

Montreal Neurological Institute (MNI) gray matter template, and spatial smoothing using a 5-mm full-width/half-maximum Gaussian kernel. We chose to smooth the data, using a moderate kernel, to benefit across-subject decoding; however, we did not expect this to compromise within-subject decoding (Kamitani & Sawahata, 2010).

2.6 Pattern classification analyses

Pattern classification analyses were applied to 'raw' (unmodeled) fMRI data. All pattern classification analyses were performed using sparse multinomial logistic regression implemented with the Princeton Multi-Voxel Pattern Analysis Toolbox (http://www.pni.princeton.edu/mvpa) and custom Matlab routines.

2.7 fMRI preprocessing for pattern classification analyses

In addition to the standard fMRI preprocessing steps, several additional preprocessing steps were applied to the fMRI data before pattern classification analyses were performed. Functional data were high-pass filtered (0.01 Hz), detrended, and z-scored within scan. All statistical analyses and inferences were based on fMRI data that were temporally compressed so that each trial corresponded to a single spatial pattern. For trials in the acquisition phase, the 3rd and 4th volumes (4-8s post word-picture pair onset) were averaged. For trials in the new learning phase, volumes 4-6 were averaged for the instructed version and volumes 3-5 were averaged for the uninstructed version. A different time window was used for the uninstructed subjects simply to account for the fact that no instruction cue was shown; thus, the window either corresponded to 4-10s post instruction onset (instructed version) or 4-10s post trial onset (uninstructed version). A wider temporal window was used for trials in the new learning phase than the acquisition phase to account for the fact that retrieval and integration processes should take longer to unfold (during new learning) than encoding processes (during acquisition). The temporal windows and averaging used here were selected *a priori* based on our previous studies (Kuhl & Chun, 2014; Kuhl, Johnson, & Chun, 2013; Kuhl et al., 2011) and were therefore not 'optimized' to find the effects of interest. After a single spatial pattern was obtained for each trial and only relevant trials were selected, additional z-scoring was performed (again, as in our previous work). First, z-scoring was performed *across all voxels* within each volume (i.e., mean response for each volume on each trial = 0), which had the effect of expressing the activity of a given voxel on a given trial *relative to activity in other voxels*. Second, z-scoring was performed, for each voxel, across all trials within each phase. For example, the mean response for each voxel within the acquisition phase would equal 0. This had the effect of expressing the activity of a given voxel on a given trial relative to the response of that voxel on other trials from the same phase.

Decoding mnemonic processing states—Decoding of mnemonic processing states was performed using across-subject pattern classification. There were two motivations for using across-subject classification. First, we cued different mnemonic processing states using shapes, and the assignment of shape to instruction was fixed within subjects but counterbalanced across subjects. Thus, performing classification across-subjects deconfounded shape and processing instruction. Second, we were specifically interested in whether a classifier trained using 'instructed' subjects would generalize to a set of ‘uninstructed’ subjects. Decoding of processing states was performed using three-way classifiers (encode vs. retrieve vs. integrate) as well as using separate pairwise classifiers (encode vs. retrieve, encode vs. integrate, retrieve vs. integrate).

For the instructed subjects, across-subject classification used leave-one-subject-out cross-validation. Specifically, the pattern classifier was trained on data from 20 of the 21 subjects, and tested on data from the held-out subject. This was repeated iteratively until each subject's data was ‘held-out’ once. As an example, for the three-way classification of processing state, the classifier would be trained on a total of 20 (subjects) * 144 (trials per subject) = 2,880 total trials and tested separately on each of 144 trials for the held-out subject. For the uninstructed subjects, across-subject classification was performed by training the classifier on all 21 of the instructed subjects (3,024 trials) and testing the classifier on each trial for each of the uninstructed subjects.

As an intuitive measure of classifier performance, we report classification accuracy for decoding of processing states, where chance accuracy was either 33.33% (for three-way classification) or 50% (for two-way classification). Note: classification accuracy was not relevant for the uninstructed version, since there was no correct (instructed) processing state in these subjects. For all analyses in which classifier-based evidence was used to predict behavioral performance, classifier evidence was defined as the log odds of the classifier output. More specifically, if x represents the classifier output corresponding to a given condition on a given trial, classifier evidence was calculated as: log[x/(1-x)]. This log transformation step was included in order to correct for non-normality in the distribution of raw classifier output.

Decoding reactivation—To test for reactivation of old associations during the new learning phase, subject-specific classifiers were trained to learn visual category information (face vs. scene vs. object) based on trials in the acquisition phase and were then tested on each trial in the new learning phase. As with the process-based classifiers, classifier evidence was defined as the log odds of the classifier's output. For each trial, one visual category corresponded to the new picture, one category corresponded to the old picture, and one category served as a baseline (neither old nor new). To obtain a measure of reactivation, classifier evidence for the baseline category was subtracted from classifier evidence for the category of the old picture. Thus, if classifier evidence corresponding to the old picture was greater than evidence for the baseline picture, this produced a positive reactivation value (Kuhl, Bainbridge, & Chun, 2012; Polyn et al., 2005). If classifier evidence corresponding to the old picture was equal to evidence for the baseline category, the reactivation value would be 0. All of the reactivation-based analyses we report are based on these continuous measures of the strength of reactivation.

2.8 Anatomical brain masks

Pattern classification analyses were restricted to specific brain regions using standard-space anatomical masks. The anatomical masks or regions of interest (ROIs) were created using the Anatomical Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). A ‘whole brain’ mask was created that included all of prefrontal cortex, posterior parietal cortex, temporal cortex, occipital cortex, and hippocampus. For all of our core analyses, we used this whole brain mask. However, because a secondary aim was to characterize how individual brain areas contribute to memory integration, we also divided the whole brain mask into twelve sub-regions according to the AAL labels: inferior frontal gyrus, middle frontal gyrus, superior frontal gyrus, medial prefrontal cortex (including anterior cingulate cortex), orbitofrontal cortex, inferior parietal lobule (including AAL regions corresponding to angular and supramarginal gyri), superior parietal lobule, medial parietal cortex, lateral temporal cortex, ventral temporal cortex, occipital cortex, and hippocampus. We favored an approach using relatively broad ROIs as opposed to searchlight analyses because (a) here, it was not of interest to localize effects to highly specific anatomical coordinates but, instead, to characterize information at the level of broad anatomical regions, and (b) the greater anatomical specificity afforded by searchlight analyses carries a cost in the form of more stringent corrections needed for multiple comparisons.

2.9 Statistical analyses

Statistical analyses were performed using R, SPSS and Matlab. We report results from paired-sample and independent sample t-tests, repeated measures ANOVA, and logistic regression. T-tests were two-tailed except in the following situations where there were obvious directional predictions: (a) when pattern classification accuracy was compared to chance, (b) when pattern classifier evidence was compared to a ‘baseline level’ and (c) when a statistical test was an internal replication of another result.

3 RESULTS

3.1 Behavioral measures of associative memory

After completing all of the acquisition and new learning rounds in the scanner (**Figure 1A-B**), subjects completed the direct association and integration tests. The direct association test allowed us to assess whether processing instructions during the new learning phase influenced participants’ subsequent memory for the old pairs and/or new pairs. Trials from the direct association test were scored as ‘correct’ if subjects selected the target picture from the set of 6 choices within the time limit (see Methods). The percentages of correct trials as a function of pair type (old vs. new pairs), processing instruction (retrieve, encode, integrate) and experiment (instructed vs. uninstructed) are reported in Table 1. Because subjects sometimes failed to make a response in the allotted time (old pairs, instructed subjects: M = 5.2%; old pairs, uninstructed subjects: M = 5.6%; new pairs, instructed subjects: M = 12.9%; new pairs, uninstructed subjects: M = 6.3%), a true measure of ‘chance performance’ was not available. However, relative to a conservative chance estimate of 16.6% (because 6 picture options were available to choose from), accuracy for both the old and new pairs was above chance in each instruction condition for the instructed subjects (t_{19}'s > 2.3, p's < .05) as well as for the uninstructed subjects (t_7's > 3.3, p's < .05). For the instructed subjects, processing instructions significantly influenced memory for the new pairs ($F_{2,38}$ = 10.23, p < .005, Greenhouse-Geisser corrected), but not for the old pairs (F < 1). For the new pairs, accuracy was higher in the encode condition (M = 38.4%) than integrate condition (M = 34.1%; t_{19} = 2.13, p = .047), and accuracy in the integrate condition was, in turn, higher than in the retrieve condition (M = 24.3%; t_{19} = 2.74, p = .01). Thus, processing instructions—

which appeared immediately *after* new pairs were presented—significantly influenced later memory for the new pairs but did not influence memory for the previously studied old pairs.

The integration test probed memory for associations *across overlapping pairs* (associations that were never directly studied; **Figure 1C**). Each integration test trial consisted of two steps: a category-level decision and an item-level decision. Table 2 displays the percentage of correct trials for each step, as well as the percentage of trials where both steps were correct, as a function of processing instruction (retrieve, encode, integrate) and experiment (instructed vs. uninstructed). Because subjects sometimes failed to make a response in the allotted time (category decision, instructed subjects: M = 6.4%; category decision, uninstructed subjects: M = 3.5%; item decision, instructed subjects: M = 19.7%; item decision, uninstructed subjects: M = 6.3%), a true measure of chance performance was not available. However, relative to a chance estimate of 33.3% for the category decision (because 3 options were available), performance was above chance for each instruction condition in the instructed subjects (t_7's > 3.7, p's < .005) and for the uninstructed subjects (t_7 = 3.22, p = .01). For the item decision, relative to a conservative chance estimate of 25% (because 4 options were available), accuracy was above chance for the instructed subjects in the integrate condition (t_{20} = 3.35, p = .003) and encode condition (t_{20} = 2.37, p = .03), but not the retrieve condition (t_{20} = .22, p = .83); accuracy for the uninstructed subjects trended toward being above 25% (t_7 = 1.61, p = .15). Notably, accuracy for the item decision was higher if the category decision was correct vs. incorrect (timed out or error): for instructed subjects, this difference was highly significant (M = 39.0% vs. M = 25.3%, t_{20} = 5.30, p = . 00003), and a similar trend was observed for uninstructed subjects (M = 34.9% vs. M = 26.8%, t_7 = 2.00, p = .09). Thus, although the category and item decisions were fully independent in terms of the task structure, accuracy of the category decision was predictive of accuracy of the item decision.

For the instructed subjects, processing instructions had a modest, non-significant influence on integration test accuracy at the category level ($F_{2,40}$ = 1.43, p = .25; **Figure 1D**), but a robust influence on accuracy at the item level ($F_{2,40}$ = 10.32, p = .0002) and on the probability of selecting the correct category *and* item ($F_{2,40}$ = 7.89, p = .001; **Figure 1D**). Across all three measures, accuracy was numerically greatest in the integrate condition and lowest in the retrieve condition. That is, although the integrate and retrieve conditions each required that subjects ‘think back’ to the old association immediately *after* presentation of the new association, the integrate condition yielded better performance on the subsequent integration test (category and item: M = 22.1%) than did the retrieve condition (category and item: M = 13.6%; t_{20} = 3.16, p = .005). Thus, simply thinking back to the old pair after seeing the new pair was not sufficient to produce the same level of performance on the integration test that was observed with the integrate instruction. On the other hand, accuracy was only modestly higher for the integrate than encode conditions (category and item: t_{20} = 1.13, p = .27), indicating that the difference between these conditions was more subtle.

3.2 Decoding mnemonic processing states

Having established that processing instructions influenced behavioral performance, we next tested whether processing states could be decoded from fMRI activity patterns (for the instructed subjects). That is, could we classify whether a given trial was associated with an encode, retrieve, or integrate instruction? To test this, we used fMRI activity patterns from the period immediately following the instruction cue (see Methods). Notably, classification was performed using leave-one-subject-out cross validation, where the classifier was trained, on each cross validation fold, to ‘learn’ the mapping of instruction condition to fMRI activity patterns based on 20 out of 21 subjects and then tested on each trial from the held-out subject (**Figure 2A**). The motivation for using across-subject classification was two-fold: (1) this approach would yield classifiers that could potentially predict processing states in new, independent subject samples, and (2) because our paradigm conveyed instructions via shape cues (**Figure 1B**)—and because the mapping of instruction to shape cue was fixed within subjects but counterbalanced across subjects—an across-subject approach avoided the possibility of simply decoding the shape that subjects were shown (i.e., circle vs. triangle vs. square). Thus, by design, our classifier could only succeed in decoding processing states to the extent that instructions elicited activity patterns that were consistent *across subjects.*

Using a (near) whole brain mask consisting of prefrontal, parietal, temporal, and occipital cortex, as well as the hippocampus (see Methods), three-way classification of processing instructions (encode vs. retrieve vs. integrate; chance accuracy = 33.3%) was significantly above chance (M = 40.6%, SD = 5.2%; t_{20} = 6.46, $p < 1.4 \times 10^{-06}$, one-tailed t-test; **Figure 2B**). Notably, the whole-brain classifier significantly out-performed each of the sub-regions that comprised the mask (t_{20}`s > 2.1, p's < .05; **Figure 2C**), indicating that the classifier made use of broadly distributed information. The distribution of voxels maximally active for each processing state can be seen in Supplementary Figure 1.

Because we were specifically interested in the distinction between an integration state vs. encoding/retrieval states, we also separately tested pairwise classifiers: encode vs. retrieve, integrate vs. encode, and integrate vs. retrieve. Using the whole-brain mask, classification accuracy was above chance for each pair of classifiers (t_{20}`s > 3.2, p's < .005, one-tailed t-tests) and classification accuracy did not significantly differ across the three pairwise classifiers ($F_{2,40}$ = 1.98, p = .15). Thus, each of the three processing instructions elicited distinct, and broadly distributed, neural activity patterns that generalized across subjects.

3.3 Predicting memory outcomes by decoding processing states

The preceding results indicate that we were able to decode the processing states subjects engaged on individual trials in the new learning phase. Of particular importance, integrate trials could be discriminated from encode/retrieve trials based on the neural activity patterns they evoked. We next asked whether we could use classifier-derived evidence for processing states to *predict* when individual memories would be integrated. In other words, when a classifier ‘detected’ strong evidence for integration during a particular new learning trial, did this correspond to higher accuracy on the corresponding trial in the post-scan integration test? As noted above, processing instructions influenced behavioral performance on the integration test; thus, the critical question is whether the classifier could predict performance on the integration test when controlling for the instructions that subjects actually received on each trial. Specifically, did variability in the strength of classifier evidence *within each instruction condition* relate to performance on the integration test?

To test for a relationship between classifier-derived evidence for integration and performance on the integration test, we applied logistic regression analyses for each of the instructed subjects wherein integration success (a binary measure based on post-test performance) was regressed upon classifier evidence from each trial during the new learning phase. Classifier evidence was derived from the three-way (encode vs. retrieve vs. integrate) whole-brain classifier. A total of nine regression analyses were run for each subject, reflecting separate analyses for each combination of form of evidence (encode evidence, retrieve evidence, integrate evidence) and instruction condition (encode trials, retrieve trials, integrate trials). The resulting beta values were then averaged *across instruction conditions* resulting in three mean beta values per subject that reflected the strength of the relationship between classifier evidence for each processing state and performance on the integration test. Critically, because the regression analyses were always separately run *within each instruction condition*, the regression was unbiased by (i.e., controlled for) any differences in performance that were related to the actual instructions. The mean beta values for each subject were then compared to a test value of o (i.e., no relationship). Although our prediction was that *integrate evidence* would predict performance on the integration test, for comparison we also tested whether encode and/or retrieve evidence predicted performance on the integration test.

One caveat for this analysis is that there were multiple ways in which ‘success’ on the integration test could be defined. Specifically, the integration post-test consisted of two steps: a category-level decision and an item-level decision (**Figure 1C**). Because the mean percentage of trials with category + item level accuracy was relatively low, and resulting bin sizes for correct trials were therefore quite small for some subjects, we chose to define successful integration as trials on which subjects made accurate category-level responses, regardless of whether or not subjects selected the specific item correctly. With this definition, a mean of 45.3% of the trials were associated with successful integration and 54.7% with unsuccessful integration. Although this division of trials was ‘blind’ to accuracy at the item-level, item-level accuracy was, as noted above, much higher when subjects were accurate relative to inaccurate at the category level.

There was a significant, positive relationship between classifier-based evidence for integration and performance on the integration test (t_{20} = 2.26, p = .04; **Figure 3A**). Thus, even when removing the effect that instructions had on behavior, decoded evidence for an integration state during new learning predicted that overlapping events would be integrated in memory. Importantly, performance on the integration test was not predicted by classifier evidence for a retrieval state (t_{20} = −1.38, p = .18) or an encoding state (t_{20} = −0.39, p = .72). [Indeed, integrate evidence better predicted subsequent integration performance than did retrieve evidence (t_{20} = 2.21, p = .04)]. Likewise, classifier evidence for an integration state did not predict performance (either positively or negatively) for the old associations or new associations, as measured by the direct association test (t_{19}`s < 1.4, p's > .2; see Supplementary Figure 2B-C). Thus, there was a selective relationship between classifier-derived evidence for an integration state and performance on the integration test.

It is notable that, although integration putatively requires that past experience be retrieved/ reactivated during new learning, evidence for a retrieval state tended to negatively predict performance on the integration test (**Figure 3A**). Presumably, this is because a ‘pure’ retrieval state comes at the expense of successfully encoding present experience at all, let alone the relationship between past and present (integration). Indeed, retrieve evidence *negatively* predicted subsequent memory for the new associations, as measured by accuracy on the direct association test ($t_{19} = -3.97$, $p = .0008$). Thus, classifier-derived evidence for a retrieval state clearly reflected situations in which present experience was not effectively encoded.

The preceding analyses demonstrate that, among the instructed sample of subjects, classifier-derived evidence for an integration state predicted subsequent performance on the integration test. Because this relationship was evident when controlling for the instructions subjects received, we believe the classifier did not learn to decode processing instructions, per se, but instead learned to decode the *processing states elicited by the instructions*. If so, then the classifier should also succeed in identifying processing states when instructions are altogether *absent*. To test this idea, we trained a new classifier using the whole-brain masks for the entire sample of instructed subjects ($n = 21$) and then applied the trained classifier to each trial for each of the uninstructed subjects ($n = 8$). As described above (see Methods), the uninstructed sample of subjects completed a nearly identical experiment, with the critical difference being that processing states were not manipulated during new learning for the uninstructed subjects. Instead, subjects were simply instructed to try to remember each pair that they studied (equivalent to the encoding condition for the instructed subjects). After exiting the scanner, subjects completed an un-anticipated integration test (the format of the integration test was identical for the instructed and uninstructed subjects).

As with the instructed subjects, we used a three-way classifier (encode vs. retrieve vs. integrate) to derive evidence for each of the three processing states. Classifier-derived evidence for each state was then used as a predictor variable in subject-specific logistic regression analyses. Strikingly, we again found that greater classifier evidence for integration during new learning was associated with better performance on the critical integration test ($t_7 = 1.95$, $p = .046$; one-tailed t-test; **Figure 3B**). As before, integrate evidence *better predicted* performance on the integration test than did retrieve evidence ($t_7 = 3.01$; $p < .01$, one-tailed t-test). Thus, the classifier was clearly successful in identifying spontaneous, subject-driven fluctuations in mnemonic processing states.

3.4 Individual differences in processing states

The preceding section assessed the relationship between trial-level evidence for an integration state and subsequent performance on the integration test (i.e., within-subject analyses). A complimentary question is whether individual differences in processing states were correlated with performance on the integration test (i.e., between-subject analyses). To this end, we tested whether participants that showed more evidence for integration during new learning (i.e., a higher percentage of trials labeled by the classifier as ‘integrate trials’) also exhibited better performance on the subsequent integration test. For this analysis, we combined the instructed and uninstructed samples of subjects to increase statistical power.

As shown in **Figure 3C** there was a marginally significant correlation between the percentage of trials in the new learning phase that were labeled by the classifier as ‘integrate’ and the percentage of trials in the integration test with accurate category-level responses (*r* = .34, *p* = .076; Figure 3C). Although the low number of integration test trials with accurate category + item memory precluded within-subject analysis of classifier evidence as a function of subsequent category + item memory (because of very low bin sizes), the low bin sizes were not problematic for across-subject analyses. Indeed, the across-subject correlation between the percentage of trials labeled as integrate and the percentage of integration test trials with accurate category- *and* item-level responses (‘category + item accuracy’) was highly significant (*r* = .53, *p* = .003; **Figure 3D**). Thus, individual differences in the degree to which an integration state was engaged during new learning (as indexed by the pattern classifier) were related to individual differences in performance on the integration test. Note: from here forward, we use category + item accuracy for all across-subject correlations, but comparisons of correlations based on category-level vs. category + item-level accuracy and other, related correlation analyses can be found in Supplementary Figure 3.

Because there were a handful of subjects that had relatively high category + item accuracy, we also tested for a within-subject relationship between integrate evidence and integration test performance, based on category + item accuracy, in this sub-sample of subjects with more favorable bin sizes. As can be seen in **Figure 3E**, when combining the instructed and uninstructed subjects, there was a ~10% gap in category + item accuracy between the 5 highest performing subjects (mean accuracy = 39.9%) and the remaining 24 subjects (mean accuracy = 13.4%). We thus repeated the within-subject logistic regression analysis for these ‘high-performing’ subjects (*n* = 5). For this analysis, ‘successful’ integration was defined as trials associated with accurate category + item memory and ‘unsuccessful’ integration as all other trials. Indeed, within this sub-sample, there was a very robust trial-level (within-subject) relationship between classifier-based evidence for integration and performance on the integration test (t_4 = 7.97, *p* = .001; **Figure 3E**). Thus, for those subjects with sufficiently high category + item accuracy, decoded integration evidence clearly predicted category + item accuracy on the integration test.

3.5 Reactivation of older memories during new learning

All of the above process-based decoding analyses were orthogonal to the specific content that subjects were remembering (i.e., whether subjects were processing faces, scenes, or objects). However, prior studies have found that successful memory integration can be predicted by measuring the *content* of the memory system during new learning: specifically, by measuring the degree to which older memories are reactivated during new learning (Wimmer & Shohamy, 2012; Zeithamova et al., 2012). Motivated by this prior work we applied decoding analyses to measure and quantify reactivation of older memories during the new learning phase. We used fMRI data collected during learning of the old pairs (acquisition rounds) to train subject-specific pattern classifiers to decode the visual category information (face vs. scene vs. object) and then tested these classifiers on data from the new learning phase (Kuhl et al., 2011; Polyn et al., 2005). On each new learning trial, one of the three visual categories corresponded to the old picture, one corresponded to the new picture, and one functioned as a baseline (neither old nor new). To obtain a measure of reactivation, the classifier evidence for the baseline category was subtracted from classifier evidence for the category of the old picture (**Figure 4A**) (Kuhl et al., 2012).

Using the whole brain mask, significant reactivation of the old picture was observed in each of the three instruction conditions (t_{20}`s > 3.8, p's < .001, one-tailed t-tests), but the strength of reactivation was strongly modulated by instruction condition ($F_{2,40} = 8.92$, $p < .001$; **Figure 4B**). Statistically, reactivation was comparable in the integrate and retrieve conditions ($t_{20} = -0.61$; $p = .55$), with both conditions eliciting stronger reactivation than the encode condition (t_{20}'s > 3.8, p's < .005).

We also measured reactivation for the uninstructed subjects. Although these subjects were never explicitly told to retrieve old items, we nonetheless saw modest evidence for reactivation (**Figure 4C**). Notably, for both the instructed and uninstructed subjects, the whole brain classifiers tended to perform *worse* than some of the individual sub-regions (which contrasted with the processing state classifier, see **Figure 2C**).

We next tested whether trial-by-trial variability in classifier-derived evidence for reactivation of older memories during new learning predicted performance on the subsequent integration test. This analysis was very similar to the regression analysis described above relating evidence for an integration state to performance on the integration test, with the only difference being that we changed the predictor variable (instead of decoded evidence for processing state, here we used decoded evidence for reactivation). As before, subject-specific logistic regression analyses were performed separately for each instruction condition in order to control for task-instructions. Resulting beta values were then averaged across instruction conditions to produce a single beta value per subject. Here, we also combined data across the instructed and uninstructed samples, in order to increase sensitivity.

Using trial-by-trial reactivation strength derived from the whole brain mask, reactivation did not predict performance on the integration test ($t_{28} = 0.26$, $p = .80$; see Supplementary Figure 2D-F for this and related analyses). However, there was a modest, but significant correlation between individual differences in reactivation strength and individual differences in performance on the integration test ($r = .39$, $p = .03$; Supplementary Figure 4). Additionally, there was a significant, positive relationship between trial-by-trial variability in reactivation strength and the strength of classifier evidence for integration, as measured by subject-specific, trial-level correlations between these two forms of classifier evidence (instructed and uninstructed samples combined and controlling for instruction condition among the instructed subjects; mean z-transformed correlation = 0.037, $t_{20} = 3.20$, $p = .003$; Supplementary Figure 5).

We also tested for a relationship between reactivation and memory on the subsequent direct association test (i.e., the old association and new associations), again using subject-specific logistic regression analyses that combined across instructed and uninstructed subjects and controlled for instruction condition among the instructed subjects. Consistent with prior evidence (Kuhl et al., 2010), we found a positive relationship between trial-by-trial fluctuations in reactivation strength during new learning and subsequent memory for corresponding old associations (t_{27} = 2.29, p = .03; Supplementary Figure 2E). That is, if old associations were reactivated during new learning, they were more likely to be subsequently remembered. Likewise, there was a robust across-subject correlation between the strength of reactivation and subsequent memory for old associations (r = .65, p < .001; Supplementary Figure 4). Subsequent memory for the new associations was not related to trial-by-trial variability in reactivation of old associations during new learning (t_{27} = .26, p = .80; Supplementary Figure 2F) nor to across-subject differences in reactivation strength (r = .17, p = .39; Supplementary Figure 4).

3.6 Processing states in regions of a priori interest

Previous research has indicated that the medial prefrontal cortex (MPFC) and hippocampus (HIPP) are particularly important for memory integration (Benoit, Szpunar, & Schacter, 2014; Schlichting & Preston, 2015; van Kesteren et al., 2013; Zeithamova & Preston, 2010; Zeithamova et al., 2012). Additionally, integration has been linked to reactivation of older memories in ventral temporal cortex (VTC) during new learning (Zeithamova et al., 2012). We therefore conducted several follow-up analyses targeting these three regions of a priori interest in order to better characterize their respective mechanistic contributions to memory integration.

As a first step, we compared pairwise processing state classification accuracy across the sub-regions. An ANOVA with factors of sub-region (MPFC vs. HIPP vs. VTC) and state pair (encode vs. retrieve, retrieve vs. integrate, encode vs. integrate) revealed a significant interaction, ($F_{4,80}$ = 2.57, p = .04; **Figure 5A**). Of particular interest was the dissociation between MPFC and HIPP: in MPFC classification was numerically highest— and only above chance—for encode vs. integrate (t_{20} = 2.50, p = .01, one-tailed t-test), whereas in HIPP, classification was numerically highest and only above chance for encode vs. retrieve (t_{20} = 4.52, p = .0001, one-tailed t-test). When specifically considering MPFC vs. HIPP, the interaction between region and state pair was significant ($F_{2,40}$ = 4.17, p = .02), confirming that these regions were differentially signaling subjects' mnemonic processing states. A complementary analysis comparing the similarity (correlation) of activation patterns across processing states and across sub-regions revealed similar results (Supplementary Figure 6).

The fact that HIPP did not distinguish integrate trials from either encode or retrieve trials (t_{20}'s < 1.2 p's > .12) is notable given that the hippocampus has previously been implicated in memory integration (Preston & Eichenbaum, 2013; Schlichting, Zeithamova, & Preston, 2014; Shohamy & Wagner, 2008; Zeithamova, Schlichting, & Preston, 2012). One possibility is that HIPP may have been poorly suited to the across-subject decoding approach that we used. To address this concern, we re-ran the processing state decoding analyses in the hippocampus *within-subjects* (i.e., using leaving one scan out cross-validation). The results were nearly identical to the across-subject results: encode vs. retrieve classification was significantly above chance (t_{20} = 2.66, p = .008, one-tailed t-test), but integrate trials could not be distinguished from either encode or retrieve trials (t_{20}'s < 1.1, p's > .15, one-tailed t-tests). Thus, the within- and between- subject analyses each indicated that hippocampal activity patterns did not differentiate an integration state from encoding or retrieval states.

Next, for each of the same sub-regions, we asked whether classifier-derived processing state evidence predicted performance on the integration test. The analyses were identical to the whole brain version (**Figure 3A**), with the exception that here we combined evidence from the instructed and uninstructed samples in order to increase sensitivity. That is, although there were separate procedures for performing classification for the instructed and uninstructed samples, we pooled the beta values produced by the subject-specific logistic regression analyses (total n = 29). The relationship between classifier-derived integration evidence and integration test performance was marginally significant in MPFC (t_{28} = 1.90, p = .068), and significant in VTC (t_{28} = 2.21, p = .04), but not significant in HIPP (t_{28} = 0.19, p = .85). Integration test performance was also *negatively* predicted by retrieve evidence in MPFC (t_{28} = −2.21, p = .04), and by encode evidence in VTC (t_{28} = −2.54, p = .02). We did not observe, for any of the sub-regions, significant correlations between individual differences in the percentage of trials in the new learning phase labeled by the classifier as ‘integrate’ and the percentage of trials in the integration test with accurate category + item (or category only) level responses (r's < .1; see Supplementary Figure 3).

3.7 Reactivation in regions of a priori interest

Our final analyses focused on reactivation within the three sub-regions of interest. As noted above, VTC reactivation of older memories during new learning has previously been associated with successful memory integration (via across-subject correlation analysis; Zeithamova et al., 2012). However, despite considerable evidence implicating MPFC in integration (Benoit et al., 2014; van Kesteren et al., 2013; Zeithamova & Preston, 2010; Zeithamova et al., 2012), prior studies have not directly probed reactivation within MPFC in relation to integration.

In VTC, we observed significant reactivation across all instruction conditions (t_{20}'s > 3.8, p's < .005) and reactivation was strongly modulated by instruction condition ($F_{2,40}$ = 9.01, p = .0006; **Figure 6A**), with the lowest degree of reactivation for encode trials. In MPFC, reactivation was significant across all conditions (t_{20}'s > 2.4, p's < .05) and there was a marginally significant effect of instruction on reactivation ($F_{2,40}$ = 2.86, p = .07; **Figure 6A**). In HIPP, reactivation was robust only in the retrieve condition (retrieve: t_{20} = 4.22, p = .0004; others: t_{20}'s < 1.1, p's > .29), and there was a significant effect of instruction on reactivation ($F_{2,40}$ = 5.05, p = .01). A targeted ANOVA comparing reactivation in HIPP vs. MPFC across retrieve vs. integrate trials revealed a significant interaction ($F_{1,20}$ = 4.78, p = .04), reflecting relatively greater reactivation in HIPP for retrieve than integrate trials and relatively greater reactivation in MPFC for integrate than retrieve trials. Thus, compared to HIPP, MPFC played a greater role in representing older memories when there was a demand to integrate past with present.

We next tested whether sub-region reactivation was related to behavioral performance on the integration test. We first performed within-subject logistic regression analyses that related trial-by-trial fluctuations in reactivation to category-level accuracy on the integration test, combining data from the instructed and uninstructed samples and controlling for instruction condition for the instructed subjects (identical to the whole-brain version of this analysis described above). A significant positive relationship was observed for VTC (t_{28} = 2.14, p = .04; **Figure 6B**), but not for MPFC (t_{28} = −.71, p = .49) or HIPP (t_{28} = −0.09, p = .93). We also tested whether individual differences in mean reactivation during new learning correlated with mean accuracy on the integration test (category + item). For this analysis (**Figure 6C**) we again combined the instructed and uninstructed samples. Significant positive correlations were observed for MPFC (r = 0.57, p = .001) and VTC (r = 0.48, p = .008), but not HIPP (r = .25, p = .19). For both VTC and MPFC, the relationship between reactivation and integration was qualitatively similar for instructed vs. uninstructed samples (**Figure 6C**: black vs. green trend lines).

4 DISCUSSION

Here, we sought to ‘read out’ mnemonic processing states from patterns of fMRI activity acquired during the learning of overlapping events so that we could predict when events would be integrated in memory. In an initial set of ‘instructed’ subjects, we explicitly biased mnemonic processing states and used across-subject decoding analyses to discriminate encoding, retrieval, and integration states. Validating these decoding results, we found that decoded evidence for an integration state during overlapping learning predicted performance on a subsequent test of memory integration. Strikingly, we found that our decoding algorithm could also successfully predict integration in a new set of ‘uninstructed’ subjects whose processing states were not biased in any way. Finally, we compared how several regions of *a priori* interest contributed to memory integration. We found that medial prefrontal cortex (MPFC) and hippocampus (HIPP) differentially signaled subjects’ mnemonic processing states. Namely, activity patterns in MPFC were relatively more diagnostic of an integration state whereas HIPP activity patterns were relatively more diagnostic of the tradeoff between encoding vs. retrieval states. Complementing this dissociation, we found that in MPFC—but not HIPP—older memories were reactivated in service of integration.

4.1 Decoding mnemonic processing states

Computational models, behavioral studies, electrophysiological recordings, and neuroimaging data all support the idea that the memory system fluctuates between distinct processing states and that these states are reflected in profiles of neural activity. For example, fMRI studies have shown that encoding and retrieval processes differentially modulate activity in several brain regions (Donaldson et al., 2001; Duncan et al., 2014; Eldridge et al., 2005). Likewise, electrophysiological recordings in humans (Rizzuto et al., 2006) and rodents (Douchamps et al., 2013; Hasselmo et al., 2002; Kunec et al., 2005; Siegle & Wilson, 2014) have identified distinct neural correlates of encoding vs. retrieval. Our study builds on these findings in two ways. First, in addition to considering encoding vs. retrieval states, we also considered integration as a potentially distinct state of the memory system. Second, we applied machine-learning algorithms in order to *decode* trial-level fluctuations in mnemonic processing states from distributed neural activity patterns. Several aspects of this decoding-based approach are notable for methodological reasons.

In the context of memory research, a growing number of fMRI studies have applied decoding analyses to read out *what* human subjects are remembering (e.g., Kuhl et al., 2011; Polyn et al., 2005). There are, however, several examples where pattern classifiers have been used—as in the present study—to decode cognitive processes or operations that are thought to generalize across the content or stimuli that subjects are seeing or remembering (e.g., McDuff, Frankel, & Norman, 2009; Poldrack, Halchenko, & Hanson, 2009; Rissman, Greely, & Wagner, 2010). Interestingly, we found that decoding of processing states was significantly better when considering whole brain activity patterns compared to any of the individual sub-regions within the whole brain mask. This suggests that individual sub-regions carried non-redundant information and that processing states were related to broadly distributed information (see **Figure 2D**). In contrast, reactivation-based decoding (a form of content decoding) was as robust or better in several of the sub-regions (e.g., VTC) compared to the whole brain mask.

Another important feature of our decoding approach is that it was applied *across subjects*. Thus, our classifiers could only succeed in decoding processing states to the extent that mappings between these states and neural activity patterns generalized across subjects. We anticipated that this across-subject decoding approach would be possible based on prior examples of across-subject decoding of cognitive processes (Mitchell et al., 2004; Poldrack et al., 2009; Rissman et al., 2010). Indeed, across-subject generalizability was a critical feature of the present approach as it allowed us to test whether a classifier trained on data from the set of instructed subjects would transfer to the set of uninstructed subjects. The fact that we observed transfer from the instructed to the uninstructed sample is striking because the very experimental manipulation that was used to train the classifier in the instructed subjects (i.e., the instructions) was absent in the uninstructed subjects. While this meant that there was no ‘correct’ label for each trial for the uninstructed subjects, we were able to validate the classifier's predictions by relating these predictions to performance on the subsequent integration test. Using this approach, it would therefore be possible to test for spontaneous memory integration in other unconstrained learning contexts or to compare the relative strength of integration across individuals. One potential advantage of such an approach is that it can isolate integration processes that occur *at the time of learning* (Shohamy & Wagner, 2008) as opposed to integration that might occur ‘offline’ during periods of rest or sleep following learning (Kumaran & McClelland, 2012).

4.2 Integration as a distinct processing state

Using whole-brain, across-subject decoding analyses, we found that integration was clearly discriminable from encoding and retrieval states. Critically, we validated these classification results by showing that decoded evidence for an integration state *positively predicted* performance on the integration test. In contrast, decoded evidence for encoding or retrieval states did not (positively) predict performance on the integration test. In fact, evidence for a retrieval state *negatively predicted* subsequent memory for new associations. Importantly, because we controlled for the instructions subjects received on each trial, these relationships between decoded processing state evidence and performance on the subsequent memory tests cannot be explained in terms of subjects following or not following instructions.

What might have contributed to the classifier's ability to detect an integration state? There are multiple sources of information the classifier may have exploited, including: working memory load, cognitive control demands, relational processing demands, an abstract intention ‘to integrate,’ etc. While we cannot tease apart these possibilities, they are not mutually exclusive. Indeed, it seems likely that different regions contributed different forms of information, which is consistent with our observation that the whole brain classifier out-performed all of the sub-regions. Thus, ‘integration’ may well be supported by a set of sub-processes. Of critical interest here, however, was to capture the broader processing state that is associated with memory integration.

It is also notable that, when considering pairwise classification of the three processing states, we did not observe a significant difference in accuracy across the three pairs (encode vs. retrieve, encode vs. integrate, retrieve vs. integrate). This is consistent with the idea that integration requires a processing state that is qualitatively distinct—or at least some sub-processes that are qualitatively distinct—from encoding and retrieval. If integration were simply ‘in between’ encoding and retrieval states, we would have expected an integration state to be more confusable with encoding/retrieval states (as was seen in HIPP). When considering candidate subprocesses that contribute to integration (e.g., working memory, cognitive control), it is also clear that integration does not lie ‘in between’ encoding and retrieval.

Although decoding of processing states was clearly most accurate when using the whole brain classifier (**Figure 2C**), we compared classification accuracy across several sub-regions of *a priori* interest (MPFC, HIPP, and VTC) in order to better understand how these regions contribute to memory integration. When considering pairwise classification (encode vs. retrieve, retrieve vs. integrate, encode vs. integrate) across the three sub-regions, we observed a significant pair-by-region interaction. In other words, these regions differentially signaled the three processing states. Whereas HIPP strongly distinguished between encoding and retrieval states, it did not differentiate either of these states from integration (VTC was qualitatively similar to HIPP). In contrast, classification accuracy in MPFC was numerically highest, and only above chance, for encode vs. integrate trials. Although direct classification of retrieve vs. integrate trials was not successful in MPFC, there was a marginally significant relationship between classifier-derived evidence for integration in MPFC and performance on the integration test (**Figure 5C**). Moreover, MPFC evidence for retrieval *negatively* predicted integration test performance. Collectively, these results indicate that activity patterns in MPFC were somewhat more diagnostic of an integration state than were activity patterns in HIPP, consistent with the idea that these regions make dissociable contributions to memory integration (Zeithamova & Preston, 2010). That said, considered on their own—and in relation to whole brain activity patterns—MPFC activity patterns were only weakly diagnostic of an integration state.

The fact that HIPP was selectively sensitive to the comparison of encode vs. retrieve trials is consistent with proposals that the hippocampus alternates between opposing encoding and retrieval states (Buzsáki 1989; Carr & Frank, 2012; O'Reilly & McClelland, 1994). Integration, in contrast, was representationally ‘in between’ encoding and retrieval states in the hippocampus. However, it is important to emphasize that these data do not argue against a role for the hippocampus in integration; rather, they clarify what that role might be (particularly in relation to MPFC). One possibility is that the hippocampus alternates between encoding and retrieval states but that these alternations occur on the order of hundreds of milliseconds (Douchamps et al., 2013; Hasselmo et al., 2002; Kunec et al., 2005; Rizzuto et al., 2006; Siegle & Wilson, 2014). This rapid alternation could allow for near-simultaneous encoding and retrieval (Kemere et al., 2013; Paulsen & Moser, 1998); however, these alternations would not be visible at the level of fMRI time scales. Alternatively, encoding and retrieval may not be categorically distinct states, and integration may reflect a processing state that is somewhere along an encoding-retrieval continuum (Carr & Frank, 2012). In either case, integration would elicit hippocampal activity patterns (as measured by fMRI) that would be representationally in between (and therefore confusable with) encoding and retrieval states, as seen here.

In contrast to HIPP, the pattern of decoding results in MPFC is not easily explained in terms of a single encoding-retrieval dimension. Indeed, MPFC did not successfully distinguish between encoding and retrieval states (**Figure 5A**). Rather, the demand to integrate past with present may require a qualitatively different form of processing that relies on MPFC. Our reactivation-based decoding results, which are considered in the following section, provide additional insight into how MPFC contributes to integration.

4.3 Relationship between reactivation and integration

Whole-brain analyses revealed that reactivation was present across all instruction conditions, but was markedly greater for retrieve and integrate trials than encode trials. These data provide confirmation that subjects successfully modulated internal representations of past experience in-line with instructions. Notably, trial-level variability in whole-brain reactivation was not predictive of performance on the integration test. However, there was a significant trial-level relationship between decoded evidence for an integration state and decoded evidence for reactivation (i.e., between evidence from the process-based and reactivation-based classifiers). This correlation is consistent with the idea that reactivation was a component of integration.

As can be seen in **Figure 4C**, reactivation was more robust in several of the sub-regions than in the whole brain mask. When specifically considering the sub-regions of *a priori* interest (MPFC, HIPP, VTC), we observed informative differences across the sub-regions. In VTC, there was robust evidence for reactivation across all instruction conditions, with greater reactivation for retrieve/integrate trials than encode trials (**Figure 6A**). Moreover, trial-level variability and individual differences in VTC reactivation during new learning predicted performance on the memory integration test (**Figures 6B-C**), consistent with prior evidence (Zeithamova et al., 2012).

Prior studies, however, have not specifically tested for reactivation within MPFC. Indeed, one question raised by prior studies (e.g., see Benoit et al., 2014) is whether MPFC supports integration by biasing reactivation in posterior regions (Schlichting & Preston, 2015) or by actively representing multiple events in an integrated manner. We found clear evidence for MPFC reactivation across all trials, with reactivation numerically greatest for integrate trials (**Figure 6A**). While we did not observe a trial-by-trial relationship between MPFC reactivation and integration test performance, there was a robust across-subject relationship between MPFC reactivation and performance on the integration test (**Figure 6C**). Collectively, these findings clearly indicate that MPFC reactivated past experience during new learning, but provide mixed evidence as to whether MPFC reactivation during new learning contributes to memory integration.

In HIPP, evidence for reactivation was robust during retrieve trials, but absent during integrate trials. Moreover, neither trial-level nor across-subject variability in HIPP reactivation predicted performance on the integration test. Thus, as with the process-based decoding analyses, the reactivation-based decoding analyses suggest that MPFC and HIPP differentially contribute to memory integration. Namely, relative to HIPP, MPFC plays a more important role in actively representing past experience during new learning. More generally, these data are consistent with evidence that prefrontal cortex allows for active representation of multiple memories (Bor et al., 2003; Hernández et al., 2010; Siegel, Warden, & Miller, 2009), particularly when task demands involve relating individual memories to one another.

4.4 Summary

Here, we used a novel methodological approach to determine how and when memory integration occurs. We show that integration involves a processing state of the memory system that is distinct from encoding/retrieval states and is reflected in broadly distributed neural activity patterns. Moreover, by decoding the processing states on individual learning trials, we were able to reliably predict behavioral expressions of memory integration. We show that this approach is flexible and powerful and also provides important new insight into the relative contributions of specific brain regions to memory integration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Supported by National Institutes of Health (R01NS089729) to B.A.K. and a Postdoctoral Fellowship from the German Academic Exchange Service (Deutscher Akademischer Austauschdienst) to F.R.R.

References

Anderson MC, McCulloch KC. Integration as a general boundary condition on retrieval-induced forgetting. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1999; 25(3): 608. Retrieved from Google Scholar.

Benoit RG, Szpunar KK, Schacter DL. Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(46):16550–5. doi:10.1073/pnas.1419274111. [PubMed: 25368170]

Bor D, Duncan J, Wiseman RJ, Owen AM. Encoding strategies dissociate prefrontal activity from working memory demand. Neuron. 2003; 37(2):361–7. [PubMed: 12546829]

Buzsáki G. Two-stage model of memory trace formation: A role for “noisy” brain states. Neuroscience. 1989; 31(3):551–70. [PubMed: 2687720]

Carr MF, Frank LM. A single microcircuit with multiple functions: State dependent information processing in the hippocampus. Current Opinion in Neurobiology. 2012; 22(4):704–8. doi:10.1016/j.conb.2012.03.007. [PubMed: 22480878]

Donaldson DI, Petersen SE, Ollinger JM, Buckner RL. Dissociating state and item components of recognition memory using fmri. NeuroImage. 2001; 13(1):129–42. doi:10.1006/nimg.2000.0664. [PubMed: 11133316]

Douchamps V, Jeewajee A, Blundell P, Burgess N, Lever C. Evidence for encoding versus retrieval scheduling in the hippocampus by theta phase and acetylcholine. Journal of Neuroscience. 2013; 33(20):8689–704. doi:10.1523/JNEUROSCI.4483-12.2013. [PubMed: 23678113]

Duncan K, Sadanand A, Davachi L. Memory's penumbra: Episodic memory decisions induce lingering mnemonic biases. Science (New York, N.Y.). 2012; 337(6093):485–7. doi:10.1126/science.1221936.

Duncan K, Tompary A, Davachi L. Associative encoding and retrieval are predicted by functional connectivity in distinct hippocampal area CA1 pathways. Journal of Neuroscience. 2014; 34(34): 11188–98. doi:10.1523/JNEUROSCI.0521-14.2014. [PubMed: 25143600]

Eldridge LL, Engel SA, Zeineh MM, Bookheimer SY, Knowlton BJ. A dissociation of encoding and retrieval processes in the human hippocampus. Journal of Neuroscience. 2005; 25(13):3280–6. doi:10.1523/JNEUROSCI.3420-04.2005. [PubMed: 15800182]

Hasselmo ME, Bodelón C, Wyble BP. A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. Neural Computation. 2002; 14(4):793–817. doi:10.1162/089976602317318965. [PubMed: 11936962]

Hernández A, Nácher V, Luna R, Zainos A, Lemus L, Alvarez M, Romo R. Decoding a perceptual decision process across cortex. Neuron. 2010; 66(2):300–14. doi:10.1016/j.neuron.2010.03.031. [PubMed: 20435005]

Kamitani Y, Sawahata Y. Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. NeuroImage. 2010; 49(3):1949–52. doi:10.1016/j.neuroimage.2009.06.040. [PubMed: 19559797]

Kemere C, Carr MF, Karlsson MP, Frank LM. Rapid and continuous modulation of hippocampal network state during exploration of new places. PloS One. 2013; 8(9):e73114. doi:10.1371/journal.pone.0073114. [PubMed: 24023818]

Kuhl BA, Chun MM. Successful remembering elicits event-specific activity patterns in lateral parietal cortex. Journal of Neuroscience. 2014; 34(23):8051–60. doi:10.1523/JNEUROSCI.4328-13.2014. [PubMed: 24899726]

Kuhl BA, Bainbridge WA, Chun MM. Neural reactivation reveals mechanisms for updating memory. Journal of Neuroscience. 2012; 32(10):3453–61. doi:10.1523/JNEUROSCI.5846-11.2012. [PubMed: 22399768]

Kuhl BA, Johnson MK, Chun MM. Dissociable neural mechanisms for goal-directed versus incidental memory reactivation. Journal of Neuroscience. 2013; 33(41):16099–109. doi:10.1523/JNEUROSCI.0207-13.2013. [PubMed: 24107943]

Kuhl BA, Rissman J, Chun MM, Wagner AD. Fidelity of neural reactivation reveals competition between memories. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(14):5903–8. doi:10.1073/pnas.1016939108. [PubMed: 21436044]

Kuhl BA, Shah AT, DuBrow S, Wagner AD. Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. Nature Neuroscience. 2010; 13(4):501–6. doi:10.1038/nn.2498. [PubMed: 20190745]

Kumaran D, McClelland JL. Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. Psychological Review. 2012; 119(3):573–616. doi:10.1037/a0028681. [PubMed: 22775499]

Kunec S, Hasselmo ME, Kopell N. Encoding and retrieval in the CA3 region of the hippocampus: A model of theta-phase separation. Journal of Neurophysiology. 2005; 94(1):70–82. doi:10.1152/jn.00731.2004. [PubMed: 15728768]

McDuff SG, Frankel HC, Norman KA. Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. Journal of Neuroscience. 2009; 29(2):508–516. Retrieved from Google Scholar. [PubMed: 19144851]

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. Learning to decode cognitive states from brain images. Machine Learning. 2004; 57(1-2):145–175.

O'Reilly RC, McClelland JL. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. Hippocampus. 1994; 4(6):661–82. doi:10.1002/hipo.450040605. [PubMed: 7704110]

Paulsen O, Moser EI. A model of hippocampal memory encoding and retrieval: GABAergic control of synaptic plasticity. Trends in Neurosciences. 1998; 21(7):273–8. [PubMed: 9683315]

Poldrack RA, Halchenko YO, Hanson SJ. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychological Science. 2009; 20(11):1364–72. doi: 10.1111/j.1467-9280.2009.02460.x. [PubMed: 19883493]

Polyn SM, Natu VS, Cohen JD, Norman KA. Category-specific cortical activity precedes retrieval during memory search. Science. 2005; 310(5756):1963–6. doi:10.1126/science.1117645. [PubMed: 16373577]

Preston AR, Eichenbaum H. Interplay of hippocampus and prefrontal cortex in memory. Current Biology. 2013; 23(17):R764–73. doi:10.1016/j.cub.2013.05.041. [PubMed: 24028960]

Rissman J, Greely HT, Wagner AD. Detecting individual memories through the neural decoding of memory states and past experience. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(21):9849–54. doi:10.1073/pnas.1001028107. [PubMed: 20457911]

Rizzuto DS, Madsen JR, Bromfield EB, Schulze-Bonhage A, Kahana MJ. Human neocortical oscillations exhibit theta phase differences between encoding and retrieval. NeuroImage. 2006; 31(3):1352–8. doi:10.1016/j.neuroimage.2006.01.009. [PubMed: 16542856]

Schlichting ML, Preston AR. Memory reactivation during rest supports upcoming learning of related content. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(44):15845–50. doi:10.1073/pnas.1404396111. [PubMed: 25331890]

Schlichting ML, Preston AR. Memory integration: Neural mechanisms and implications for behavior. Current Opinion in Behavioral Sciences. 2015; 1:1–8. doi:10.1016/j.cobeha.2014.07.005. [PubMed: 25750931]

Schlichting ML, Zeithamova D, Preston AR. CA1 subfield contributions to memory integration and inference. Hippocampus. 2014 doi:10.1002/hipo.22310.

Shohamy D, Wagner AD. Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. Neuron. 2008; 60(2):378–89. doi:10.1016/j.neuron.2008.09.023. [PubMed: 18957228]

Siegel M, Warden MR, Miller EK. Phase-dependent neuronal coding of objects in short-term memory. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(50): 21341–6. doi:10.1073/pnas.0908193106. [PubMed: 19926847]

Siegle JH, Wilson MA. Enhancement of encoding and retrieval functions through theta phase-specific manipulation of hippocampus. Elife. 2014; 3:e03061. [PubMed: 25073927]

Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER, Morris RG. Schemas and memory consolidation. Science. 2007; 316(5821):76–82. doi:10.1126/science.1135935. [PubMed: 17412951]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Joliot M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage. 2002; 15(1):273–89. doi:10.1006/nimg.2001.0978. [PubMed: 11771995]

van Kesteren MT, Beul SF, Takashima A, Henson RN, Ruiter DJ, Fernández G. Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. Neuropsychologia. 2013; 51(12):2352–9. doi:10.1016/j.neuropsychologia.2013.05.027. [PubMed: 23770537]

Wimmer GE, Shohamy D. Preference by association: How memory mechanisms in the hippocampus bias decisions. Science. 2012; 338(6104):270–3. doi:10.1126/science.1223252. [PubMed: 23066083]

Zeithamova D, Preston AR. Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. Journal of Neuroscience. 2010; 30(44):14676–84. doi: 10.1523/JNEUROSCI.3250-10.2010. [PubMed: 21048124]

Zeithamova D, Dominick AL, Preston AR. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron. 2012; 75(1):168–79. doi: 10.1016/j.neuron.2012.05.010. [PubMed: 22794270]

Zeithamova D, Schlichting ML, Preston AR. The hippocampus and inferential reasoning: Building memories to navigate future decisions. Frontiers in Human Neuroscience. 2012; 6:70. doi:10.3389/fnhum.2012.00070. [PubMed: 22470333]

Highlights

- Memory retrieval, encoding, and integration elicit distinct fMRI activity patterns
- By decoding memory states during learning, subsequent behavior can be predicted
- Decoded evidence for an integration state selectively predicts across-event memory
- Reactivation of older memories is related to, but dissociable from, integration
- Medial prefrontal cortex and hippocampus differentially signal memory states

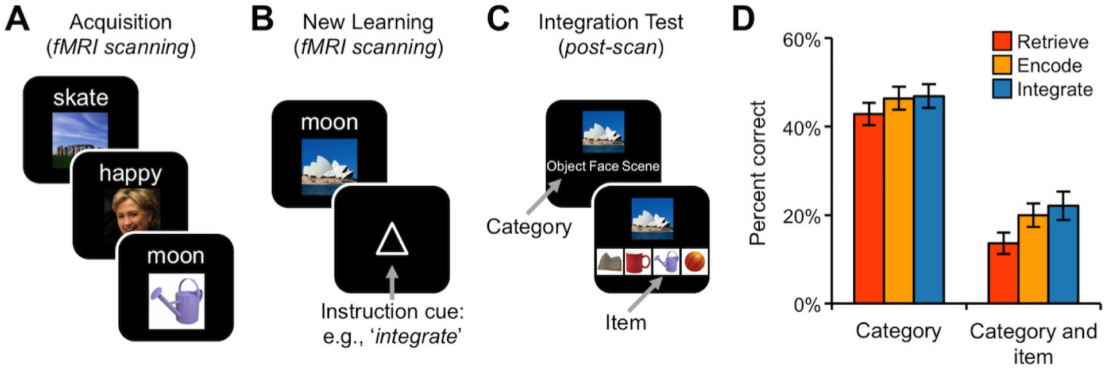

Figure 1.

Experimental design and behavioral results. (**A**) During acquisition rounds (8 total), subjects studied word-picture pairs (*old pairs*; 4s each). Pictures were drawn from three categories: faces, scenes, objects. (**B**) Each acquisition round was followed by a new learning round in which words from the immediately preceding acquisition round were paired with new pictures (*new pairs*, 2s each). After each word-picture pair disappeared, a shape cue (6s) instructed participants to: *encode* the new pair, *retrieve* the old pair, or *integrate* the old and new pairs. (**C**) After all of the acquisition/new learning rounds, subjects completed a surprise integration test. On each trial, a picture from the new learning rounds was presented and subjects attempted to remember the corresponding old picture (i.e., the picture that shared the same word cue). The integration test consisted of two steps: first participants indicated the category of the old picture (object, face, or scene; 4s maximum), and then subjects indicated the specific old picture from a set of 4 choices (all from the same visual category; 3s maximum). (**D**) Instructions during new learning (encode, retrieve, integrate) significantly influenced accuracy in selecting the specific picture ($F_{2,40}$ = 7.89, *p* = .001); there was a similar but non-significant pattern for the category-level decision ($F_{2,40}$ = 1.43, *p* = .25). Error bars correspond to standard error of the mean.

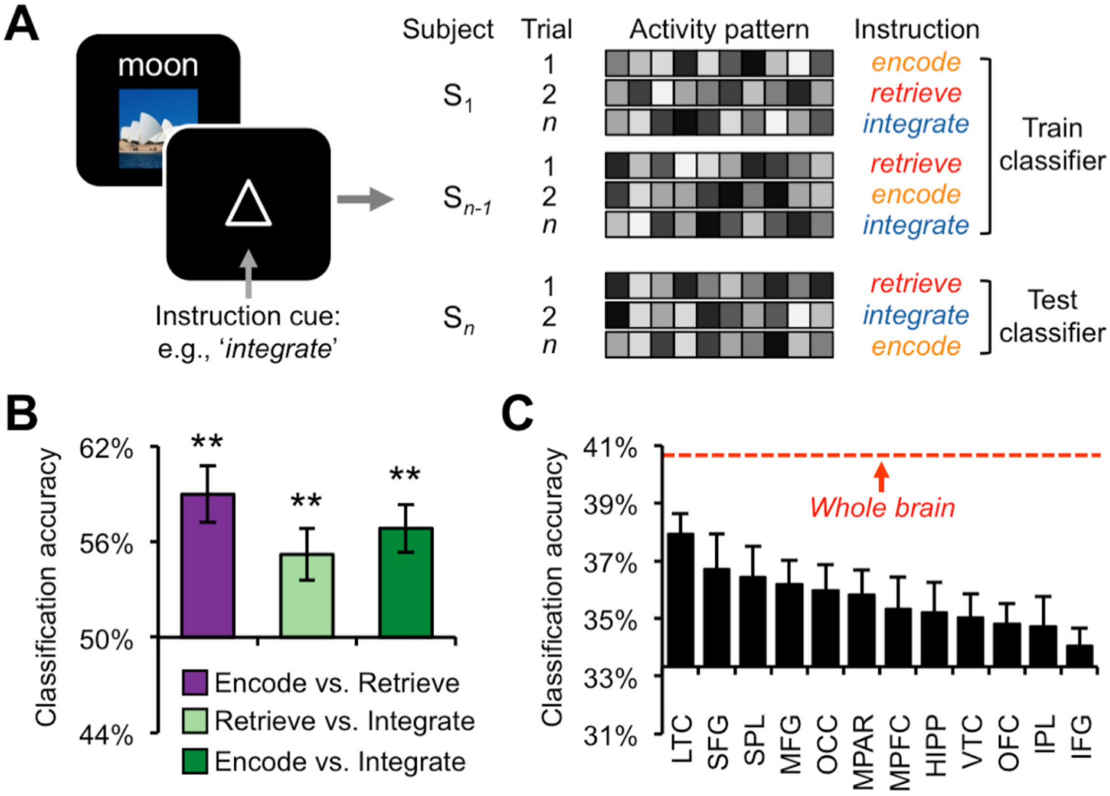

Figure 2.

Decoding mnemonic processing states. (**A**) State decoding was performed using leave-one-subject-out cross-validation, in which classifiers were iteratively trained to decode the instruction received on each trial (retrieve vs. encode vs. integrate) using data from 20/21 subjects and then tested on each trial for the held-out subject. (**B**) Pairwise classification accuracy (for each pair of processing states) for the whole brain mask. (**C**) Three-way classification accuracy in sub-region masks. Dashed red line = performance of whole brain classifier. Error bars correspond to standard error of the mean. Notes: ** p < .005, one-tailed t-test; IFG = inferior frontal gyrus; MFG = middle frontal gyrus; SFG = superior frontal gyrus; MPFC = medial prefrontal cortex; OFC = orbitofrontal cortex; LTC = lateral temporal cortex; VTC = ventral temporal cortex; HIPP = hippocampus; IPL = inferior parietal lobule; SPL = superior parietal lobule; MPAR = medial parietal cortex; OCC = occipital cortex.

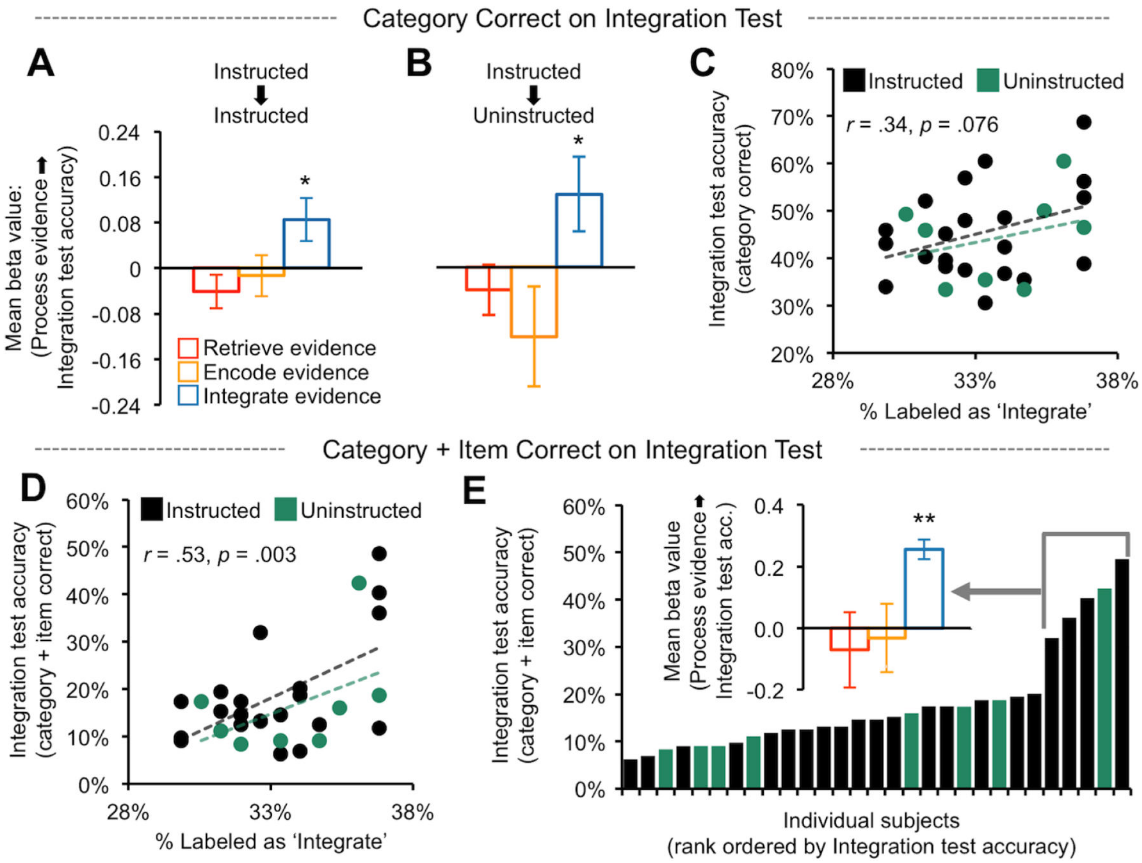

Figure 3.

Predicting integration. (**A**) Classifier evidence for each processing state (i.e., retrieve evidence, encode evidence, integrate evidence) was used to predict behavioral performance on the post-scan integration test. Separate logistic regression analyses were performed for each subject and each instruction condition (to control for effects of instruction). Each bar represents the mean beta values from the regression analyses. Performance on the integration test was selectively predicted by classifier-derived evidence for an integration state during new learning. (**B**) A classifier was trained on data from the full sample of 'instructed' subjects and applied to data from a separate set of 'uninstructed' subjects. Trial-level classifier evidence from the uninstructed subjects was then used to predict performance on the integration test (as in **A**). Again, evidence for an integration state during new learning predicted performance on the integration test. (**C**) The across-subject correlation between percentage of new learning trials labeled by the classifier as 'integrate' and mean category-level accuracy on the subsequent integration test was marginally significant [data are collapsed across instructed (black) and uninstructed (green) samples, but separate trend lines are shown for each group for comparison]. (**D**) Same as (**C**) except that integration test performance (*y*-axis) reflects mean category + item accuracy. (**E**) Rank ordered category + item accuracy for individual subjects [combining across instructed (black) and uninstructed (green) samples]. Among the small sub-group of subjects with accuracy above 30% (*n* = 5), there was a robust trial-level relationship between classifier-derived evidence for an integration state during new learning and category + item level accuracy on the subsequent integration test. Notes: * *p* < .05, ** *p* < .005; a two-tailed t-test was used for (**A**), but a one-tailed t-test was used for (**B**) given that the analysis was a replication with a clear directional prediction.

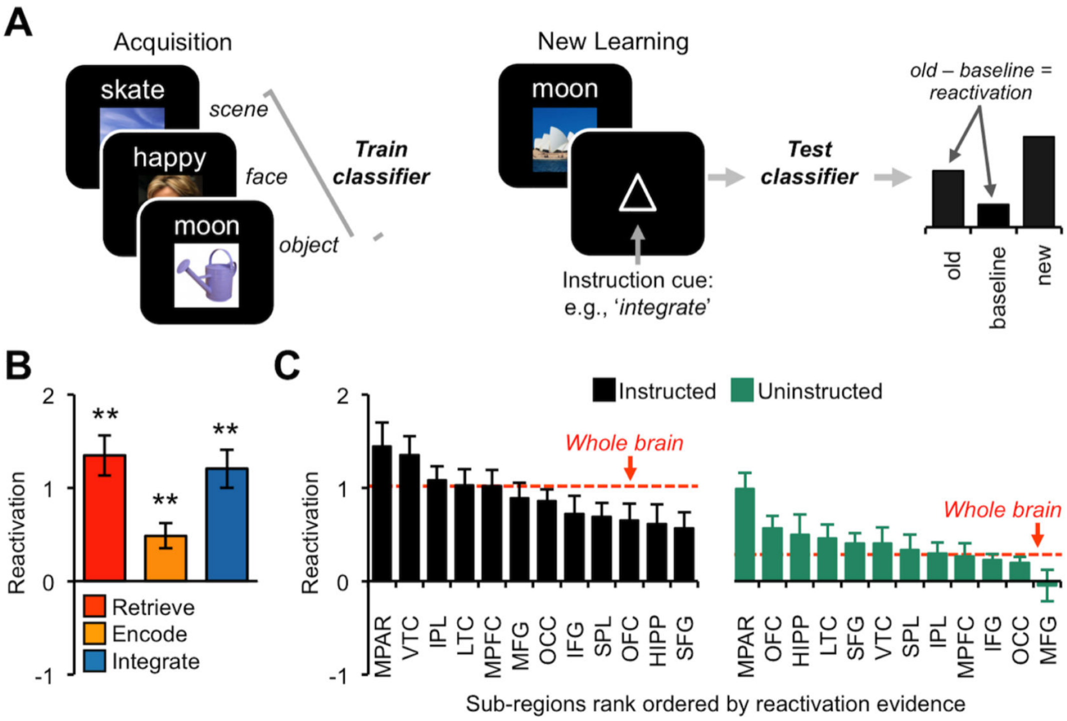

Figure 4.

Decoding reactivation. (**A**) Pattern classifiers were trained to discriminate visual category information (face vs. scene vs. object) using data from the acquisition phase. The classifiers were then tested on each trial in the new learning phase to measure the strength of evidence for the category of the old picture (as well as for the baseline and new picture categories); evidence for the baseline category was subtracted from evidence for the old category to obtain a measure of reactivation. (**B**) Reactivation in the whole brain mask as a function of instruction condition. (**C**) Reactivation in twelve sub-regions of the whole brain mask, separately for the instructed and uninstructed subjects. Error bars correspond to standard error of the mean. Notes: ** $p < .005$, one-tailed t-test; IFG = inferior frontal gyrus; MFG = middle frontal gyrus; SFG = superior frontal gyrus; MPFC = medial prefrontal cortex; OFC = orbitofrontal cortex; LTC = lateral temporal cortex; VTC = ventral temporal cortex; HIPP = hippocampus; IPL = inferior parietal lobule; SPL = superior parietal lobule; MPAR = medial parietal cortex; OCC = occipital cortex.

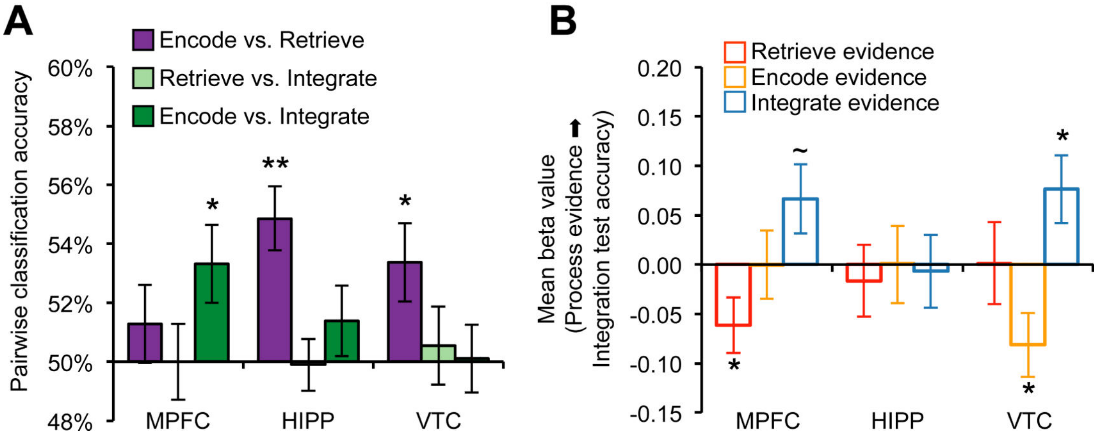

Figure 5.

Process decoding in sub-regions of interest. (**A**) Pairwise classification accuracy for each pair of instruction conditions across sub-regions. (**B**) Trial-by-trial fluctuations in classifier evidence for each processing state were used to predict category-level behavioral performance on the post-scan integration test using logistic regression analyses (as in **Figure 3A**) for each of the sub-regions. Each bar represents the mean beta values from separate regressions for each form of classifier evidence (retrieve, encode, integrate) and each sub-region. Notes: ** *p* < .01; * *p* < .05; ~ *p* < .1. One-tailed t-tests were used for (**A**) given that classifier accuracy was compared to chance, but two-tailed tests were used in (B). Performance on the integration test was positively predicted by classifier-derived integrate evidence in MPFC (marginally significant) and VTC. Integration test performance was negatively predicted by retrieve evidence in MFPC and encode evidence in VTC.

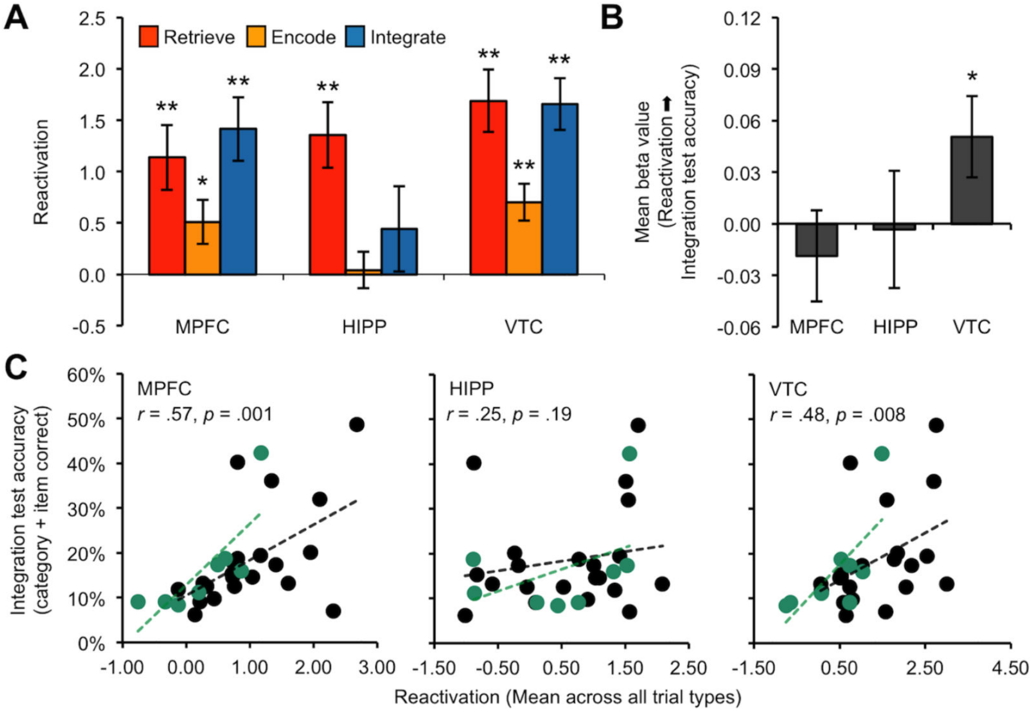

Figure 6.

Reactivation in sub-regions of interest. (**A**) Pattern classifiers were trained to discriminate visual category information (face vs. scene vs. object) using data from the acquisition phase and were then tested on each trial in the new learning phase. Classifier evidence for the baseline category (i.e., the category to which neither the old nor new picture belonged) was subtracted from evidence for the old category to obtain a measure of reactivation. Across the sub-regions, reactivation was greater for integrate and retrieve trials than encode trials. HIPP was characterized by relatively weaker reactivation on integrate than retrieve trials, which contrasted with MPFC. (**B**) Trial-by-trial fluctuations in reactivation strength during new learning were related to category-level accuracy on the subsequent integration test using subject-specific logistic regression analyses (data from instructed and uninstructed subjects were combined). Individual bars reflect mean beta values from these regression analyses, separately for each sub-region. Reactivation in VTC positively predicted subsequent performance on the integration test. (**C**) Individual differences in mean reactivation during the new learning phase were correlated with category + item accuracy on the subsequent integration test, separately for each sub-region. Significant across-subject correlations were observed in MPFC and VTC. [Notes: ** $p < .01$; * $p < .05$; ~ $p < .1$. One-tailed t-tests were used for (**A**) given that reactivation was compared to baseline, but two-tailed tests were used in (B). the correlations combined the instructed (black) and uninstructed (green) samples, but separate trend lines are shown for each group for comparison].

Table 1

Direct association test performance

		Old pairs		New pairs	
		Mean	***(SD)***	**Mean**	***(SD***
Instructed	Retrieve	64.9%	*(21.5%)*	24.3%	*(14.4%)*
	Encode	63.1%	*(22.7%)*	38.4%	*(17.5%)*
	Integrate	64.3%	*(23.4%)*	34.1%	*(17.4%)*
	All trials	64.1%	*(22.1%)*	32.3%	*(14.3%)*
Uninstructed	All trials	49.5%	*(18.3%)*	37.9%	*(18.0%)*

Table 2

Integration test performance

		Category		**Item**		**Category & Item**	
		mean	***(SD)***	**mean**	***(SD)***	**mean**	***(SD)***
Instructed	Retrieve	42.8%	*(11.5%)*	25.6%	*(12.2%)*	13.6%	*(10.9%)*
	Encode	46.3%	*(12.0%)*	32.9%	*(15.3%)*	19.9%	*(12.1%)*
	Integrate	46.8%	*(12.3%)*	37.3%	*(16.8%)*	22.1%	*(14.8%)*
	All trials	45.3%	*(9.7%)*	31.9%	*(13.2%)*	18.6%	*(11.3%)*
Uninstructed	All trials	44.3%	*(9.6%)*	31.3%	*(11.0%)*	16.5%	*(11.2%)*