Behavioral/Cognitive

Cortical Overlap and Cortical-Hippocampal Interactions Predict Subsequent True and False Memory

Erik A. Wing,[1,2] Benjamin R. Geib,[1] Wei-Chun Wang,[1,3] Zachary Monge,[1] Simon W. Davis,[4] and Roberto Cabeza[1]

[1]Center for Cognitive Neuroscience, Duke University, Durham, North Carolina 27708, [2]Rotman Research Institute, Baycrest, Toronto, Ontario M6A 2E1, Canada, [3]Department of Psychology, University of California Berkeley, Berkeley, California 94720, and [4]Department of Neurology, Duke University School of Medicine, Durham, North Carolina 27708

The declarative memory system allows us to accurately recognize a countless number of items and events, particularly those strengthened by repeated exposure. However, increased familiarity due to repetition can also lead to false recognition of related but new items, particularly when mechanisms supporting fine-grain mnemonic discrimination fail. The hippocampus is thought to be particularly important in separating overlapping cortical inputs during encoding so that similar experiences can be differentiated. In the current study of male and female human subjects, we examine how neural pattern similarity between repeated exemplars of a given concept (e.g., apple) influences true and false memory for target or lure images. Consistent with past work, we found that subsequent true recognition was related to pattern similarity between concept exemplars and the entire encoding set (global encoding similarity), particularly in ventral visual stream. In addition, memory for an individual target exemplar (a specific apple) could be predicted solely by the degree of pattern overlap between the other exemplars (different apple pictures) of that concept (concept-specific encoding similarity). Critically, subsequent false memory for lures was mitigated when high concept-specific similarity in cortical areas was accompanied by differentiated hippocampal representations of the corresponding exemplars. Furthermore, both true and false memory entailed the reinstatement of concept-related information at varying levels of specificity. These results link both true and false memory to a measure of concept strength expressed in the overlap of cortical representations, and importantly, illustrate how the hippocampus serves to separate concurrent cortical overlap in the service of detailed memory.

Key words: encoding; false memory; fMRI; hippocampus; recognition; representational similarity analysis

Significance Statement

In some instances, the same processes that help promote memory for a general idea or concept can also hinder more detailed memory judgments, which may involve differentiating between closely related items. The current study shows that increased overlap in cortical representations for conceptually-related pictures is associated with increased recognition of repeated concept pictures. Whether similar lure items were falsely remembered as old further depended on the hippocampus, where the presence of more distinct representations protected against later false memory. This work suggests that the differentiability of brain patterns during perception is related to the differentiability of items in memory, but that fine-grain discrimination depends on the interaction between cortex and hippocampus.

Introduction

Memories can vary greatly in their level of specificity, and factors that help reinforce memory for an overall concept do not always promote later discrimination of details. For example, passing through a hall of landscape paintings may strengthen one's memory for that general category of artwork but create difficulty in later recognizing a specific painting. Formal models of memory stress that recognition judgments are not made in isolation but instead depend on the relationships between items in an overall encoding set (e.g., other works of art in a museum). Higher correspondence among encoding items is thought to produce stronger familiarity signals during retrieval, leading to a positive relationship between "global similarity" and accurate recognition (Gillund and Shiffrin, 1984; Hintzman, 1988; Nosofsky, 1991) that might also produce false memory for similar but new items (Nosofsky, 1988).

Past neuroimaging work has linked overlap in evoked cortical representations to recognition success both for measures capturing the similarity between a stimulus and all other encoding set

Received July 19, 2019; revised Jan. 8, 2020; accepted Jan. 10, 2020.

Author contributions: E.A.W., B.R.G., W.-C.W., Z.M., S.W.D., and R.C. designed research; E.A.W. performed research; E.A.W., B.R.G., W.-C.W., Z.M., S.W.D., and R.C. analyzed data; E.A.W. wrote the paper.

This work was supported by the National Institute of Aging Grant AG 019731.

The authors declare no competing financial interests.

Correspondence should be addressed to Erik A. Wing at ewing@research.baycrest.org.

https://doi.org/10.1523/JNEUROSCI.1766-19.2020

items (Davis et al., 2014), and between similar (LaRocque et al., 2013; van den Honert et al., 2016) or identical stimulus repetitions (i.e., self-similarity; Xue et al., 2010, 2013; Ward et al., 2013). The idea that pattern similarity provides a neural index of item strength or consistency beneficial for subsequent memory broadly accords with prior work connecting memory to stimulus-based similarity measures like semantic (e.g., word associative strength) or physical (e.g., overlap in visual form) relatedness (for review, see Clark and Gronlund, 1996). Assuming that brain-related measures partly capture shared features between encoding items, relationships between the most closely-related stimuli (e.g., exemplars of the same concept) may be even more important for subsequent memory than general global encoding similarity.

When new items closely resemble past experience, increased encoding similarity can also produce false memory (Roediger et al., 2001; Konkle et al., 2010; Pidgeon and Morcom, 2014, 2016). The hippocampus is thought to separate overlapping cortical inputs during encoding (Norman and O'Reilly, 2003; Norman, 2010; Yassa et al., 2010), thereby protecting against later false recognition of related items. Thus, increased encoding-related representational overlap might increase susceptibility to subsequent false alarms while also reflecting item strength or stability that is beneficial for target recognition. In fact, increased similarity between temporal pole representations during word processing has been shown to predict false alarms to critical lures (Chadwick et al., 2016; Zhu et al., 2019), whereas other studies have focused on corresponding retrieval-related representations (Ye et al., 2016; Lee et al., 2018; Bowman et al., 2019). Along with univariate activity differences in posterior cortex (Gonsalves et al., 2004; Garoff et al., 2005; Abe et al., 2013; Kurkela and Dennis, 2016), several studies have described encoding activity associated with true and false memory formation in the medial temporal lobe (MTL) (Okado and Stark, 2005; Kim and Cabeza, 2007). A major unanswered question concerns how false memory relates to hippocampal processes that operate directly on overlapping stimulus-evoked representations in neocortex.

In the present study, we addressed how the overlap and separation of cortical representations influences true and false memory. During encoding (Fig. 1*A*), subjects viewed different *exemplars* (e.g., a specific dog picture) for a set of general *concepts* (e.g., the concept "dog"). A recognition memory test 1 d later (Fig. 1*B*) included new exemplars of encoded concepts (*lures*), as well as direct repeats (*targets*), and novel concept pictures.

We examined pattern similarity specifically between related exemplars of a concept (concept-specific encoding similarity) but also across the entire encoding set (global encoding similarity). We hypothesized that in posterior cortex, the former might indicate concept-level activation that should increase both true and false memory. To probe the mnemonic consequences of undifferentiated exemplar representations, we explored how the hippocampus might attenuate cortical overlap to enable accurate discrimination of similar lures. Together, these analyses address theoretical accounts of mnemonic strength and separation by examining cortical representations across region and memory phase.

Materials and Methods

Participants

Thirty participants completed both sessions of the experiment in accordance with a protocol approved by the Duke University Institutional Review Board. All participants were right handed and had normal or corrected-to-normal vision and no disclosed history of neurological or psychiatric disorder. Because of a technical error, behavioral data from one participant was not saved, leaving 29 in the final sample ($n = 29$, 21 female; mean age = 24.97 years).

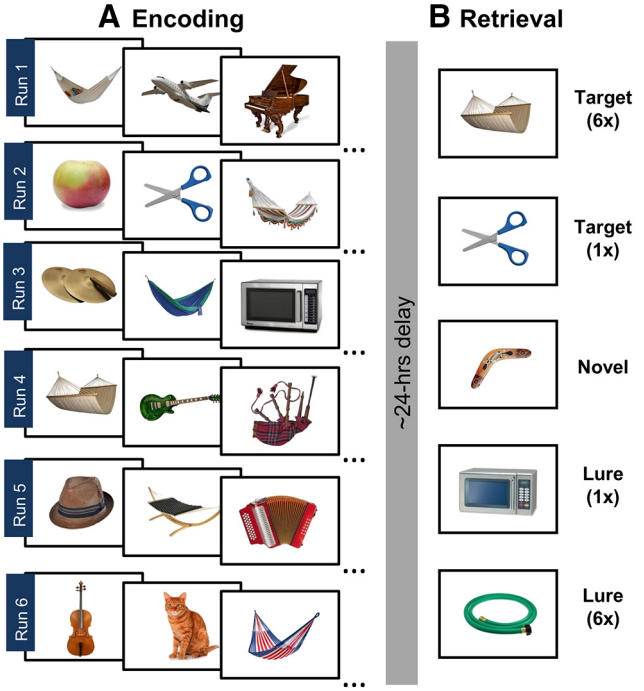

Figure 1. Schematic of experimental paradigm. ***A***, During a category detection encoding task, 72 object concepts were shown 6 times each, with different exemplar versions in each run. Another 72 object concepts were shown only once. ***B***, Retrieval (after 24 h) involved old/new recognition judgments (4-point decision, with confidence) to items including repeated targets and additional lures (unseen exemplars) of encoding concepts, as well as novel objects.

Stimuli

Object picture stimuli were photographs of common, nameable objects appearing on a white background. The images were collected from the internet, resized to 300 × 300 pixels, and centered on a white background to form images whose final dimensions were 600 × 600 pixels. Pictures were drawn from a number of living and nonliving categories (e.g., animals, tools, clothing, furniture, vehicles, musical instruments), which were selected based on past work examining representational structure (Kriegeskorte et al., 2008) and because the concepts they included had many readily available exemplar images. The musical instruments category, which served as the target class for the category detection encoding task (described in the following section), contained 72 unique instrument pictures. The remaining object pictures encompassed 144 concepts (e.g., dog, hammer, boots, chair). The complete set comprised 7 different exemplar versions for each of these 144 concepts, for a total of 1008 object pictures. Two counterbalance lists were formed such that one-half of the concepts in a given list served as non-repeated concepts (only 1 exemplar version was shown at encoding), and the other half served as the exemplar repeat concepts (6 exemplar versions of each concept appeared at encoding).

Experimental design

The experiment took place over the course of two scanning sessions, separated by 24 h (± 1 h). The first session consisted of six similarly structured runs of covert object encoding, with a final run of face encoding, not discussed here. In the object encoding runs, pictures appeared for 1 s, with a jittered fixation intertrial interval (ITI; M = 3 s, range = 2–5 s). Subjects were told to make a categorization decision as quickly and as accurately as possible, pressing with their index finger when pictures of musical instruments appeared (12 per run, 72 total), and with their middle finger for objects of any other category. At focus in the present report, each run contained exemplar pictures (72 per run, 432

total) from a set of 72 concepts that were repeated across each of the six runs, for a total of six exemplar pictures per concept. In addition to the musical instrument target category and these concept repeat exemplars, each run also contained 12 object exemplars for concepts that were not repeated (12 per run, 72 total).

Within runs, the order of trial presentation was random, with the constraint that neither musical instruments nor non-repeated concept exemplars appear on more than two consecutive trials. In addition to counterbalancing which group of 72 concepts served as the repeated or non-repeated set, the assignment of exemplar version to run number was also counterbalanced across participants.

The second session contained two scanned memory tests as well as an unscanned post-test. The first five runs of retrieval contained a word concept recognition test (data not shown). Four runs of object picture recognition followed the word memory test runs. Picture recognition contained targets (identical repeats of pictures shown at encoding), lures (an unseen exemplar version of an encoding concept), and novel object pictures. Excluding musical instruments, the recognition section contained one target and one lure image for each concept from the encoding phase. Lure images for repeated concepts (18 per retrieval run, 72 total) corresponded to the seventh unseen exemplar of each concept, also counterbalanced across participants such that different exemplars served as the concept lure for different participants. Only one of the six encoding exemplars from each repeated concept was shown as a target during the scanned retrieval test. These targets (18 per retrieval run, 72 total) were drawn evenly from each of the six encoding runs. In addition, retrieval runs contained the same number of targets and lures for non-repeated encoding concepts. For all concepts, the order of target or lure appearance was controlled such that for one-half of the concepts, the target version of an object was seen before the lure version, with the reverse order for the other half. The order of trial presentation within run was random, with the constraint that the target and lure of a given concept not appear in adjacent retrieval runs. Picture recognition also included novel objects (8 per retrieval run, 32 total), which were images of nameable items whose concept had not appeared previously anywhere in the study.

Picture recognition trials were shown for 2 s, with a jittered fixation ITI (M = 3 s, range = 2–5 s). Subjects were instructed to make a 4-point recognition memory decision ranging from "definitely old" to "definitely new," according to whether they thought the picture was either a repeat from the prior session (old) or was being seen for the first time (new). Subjects responded with the first four fingers of their right hand, and finger-response mappings were counterbalanced across subjects. Detailed instructions for this task emphasized that pictures were only to be considered old if they were identical to pictures seen at encoding, and not merely if they depicted a previously-shown concept. Responses to targets therefore ranged from high confidence hits to high confidence misses, whereas responses to lures spanned high confidence false alarms to high confidence correct rejections.

After object recognition, participants completed a final two runs of face recognition, which concluded the task-based runs. In a behavioral session directly following the scan session, the remaining five encoding exemplars for each repeated concept that were not presented during scanned picture recognition were shown in a post-scan behavioral memory test. This test (data not shown) was similar in structure to the scanned recognition runs, but contained only targets of repeated concepts, and had a fixed ITI of 1 s.

fMRI data acquisition and processing

MRI data collection occurred on a 3T GE scanner at the Duke University Brain Imaging and Analysis Center. The first session consisted of seven consecutive functional runs. On each run, following a brief calibration scan, functional images were acquired using a SENSE spiral-in sequence (TR = 2000 ms, TE = 30 ms, FOV = 24 cm, 34 oblique slices with voxel dimensions of 3.75 × 3.75 × 3.8 mm, interleaved acquisition). All stimuli were projected onto a screen at the back of the scanner bore, and responses were recorded using a four-button fiber-optic response box. Scanner noise was reduced with earplugs, and head motion was minimized with foam pads. During the second session, both functional and structural data were acquired. Eleven task-based functional runs at the beginning of the session had parameters identical to those used during encoding. Following these runs, a high-resolution anatomical image (96 axial slices parallel to the AC–PC plane with voxel dimensions of 0.9 × 0.9 × 1.9 mm) was collected. A final diffusion-weighted structural scan was collected at the end of the second scanning session.

Preprocessing and data analysis were performed using SPM12 (Wellcome Department of Cognitive Neurology, London, UK). After discarding the initial five volumes of each run to allow for scanner stabilization, images were corrected for slice time acquisition and motion. In addition, the ART toolbox was used to generate regressors corresponding to outlier volumes (default outlier cutoffs; https://www.nitrc.org/projects/artifact_detect/). Functional images were then coregistered to their respective anatomical images. Anatomical images were segmented to produce tissue maps for gray matter, white matter, and cerebrospinal fluid, along with parameters used to normalize both anatomical and functional images to Montreal Neurological Institute space. Finally, to account for confounding sources of physiological noise, normalized functional images were de-noised using the DRIFTER toolbox (Särkkä et al., 2012) and remained unsmoothed.

Statistical analysis

Behavioral data. Behavioral data were analyzed for both the category detection task in Session 1 and for the object recognition memory task in Session 2. Overall accuracy of category detection (proportion of successfully categorized trials within musical instruments and non-instrument items, separately) was calculated for each subject to assess general task engagement during encoding. To test for facilitation in object categorization due to the repetition of concepts, the mean reaction time of repeated and non-repeated concept trials was calculated separately within each run. The linear decrease in reaction times across runs was then compared between the conditions within the context of a 2 (repetition condition) × 6 (encoding run) repeated-measures ANOVA. Behavioral memory performance was examined by computing hit rates and false alarm rates (proportion of hits and false alarms, collapsing high-confidence and low-confidence responses) for the repeated and non-repeated conditions separately, as well as false alarm rate for the novel condition. Corrected recognition (the confidence-collapsed hit rate for target items minus the corresponding confidence-collapsed false alarm rate for lure items) was compared between the repetition conditions with paired samples two-tailed *t* tests, which were also run to ensure that the rate of lure false alarms (average false alarm rate across repeated and non-repeated conditions) was higher than the false alarm rate for novel items.

fMRI data: first-level models and ROIs. Statistical analysis of fMRI data was performed in SPM12 using the general linear model (GLM). Representational similarity analyses were run on unsmoothed trialwise β estimates from a single-trial model constructed with the least-squares single method (Mumford et al., 2012). Accordingly, a separate GLM was run for each trial at encoding and retrieval. A high-pass filter of 128 s and grand mean scaling were applied to the data, and serial autocorrelations in the time series were accounted for using the autoregressive model. Events were modeled using a canonical hemodynamic response, which was applied to a delta function locked to the onset of each trial. Models included one regressor for the trial of interest as well as one regressor for all other trials of interest, plus corresponding regressors for the temporal derivative. In addition, each model contained regressors for six motion parameters generated during motion correction, and also regressors generated by the ART toolbox to flag spikes in motion activity. Finally, the single-trial β volumes for each trial of interest were converted to *t* values by dividing each voxel's β estimate by its standard error.

Representational similarity analyses were run within an anatomical mask of the hippocampus (HC; region of interest (ROI) from the subcortical Harvard-Oxford Atlas, thresholded at 0.25) as well as four cortical regions (see Fig. 3*D*). For examining effects in ventral visual processing stream, we used Brodmann area (BA) ROIs (cerebellar voxels removed) corresponding to early visual cortex (EVC; conjoined BA 17/18) as well as late visual cortex (LVC; BA 37). This division accords with past research that has described differential responses for true and false

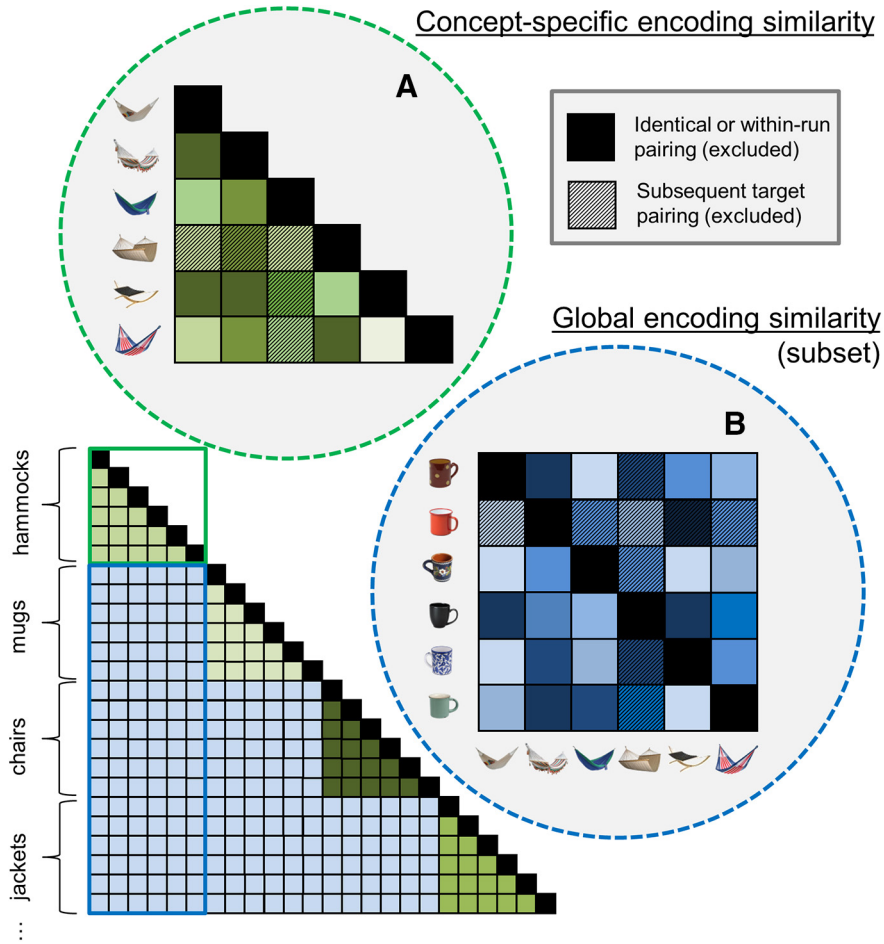

Figure 2. Overview of encoding-based representational similarity analysis for true memory. ***A***, Concept-specific encoding similarity; example of pairwise pattern correlations contributing to concept-specific encoding similarity for a single concept, with hatched cells denoting excluded pairings involving subsequent target exemplars. ***B***, Global encoding similarity; subset of between-concept pairings contributing to global similarity for a given concept (hammocks). For the concept hammock, global similarity includes the depicted hammocks-mugs similarity values along with the corresponding between-concept pairings for hammocks-chairs, hammocks-jackets, etc. Within-run pairings and those involving subsequent targets were excluded.

memory in visual cortex as a function of sensitivity to lower-level versus more integrative features (Slotnick and Schacter, 2004; Kim and Cabeza, 2007; Dennis et al., 2012; Bowman et al., 2019). Recent investigations have sought to explicitly connect memory veracity to differential stages of perceptual processing (Karanian and Slotnick, 2018), making the division of early and late visual cortex particularly relevant (see Discussion). Although ventral visual stream is often the focus of work on stimulus representations, we also examined effects in two parietal regions: angular gyrus (ANG; BA 39) and dorsal parietal cortex (DPC; BA 7). In addition to broadly supporting memory processes (Spaniol et al., 2009; Davis et al., 2018), recent work suggests that these parietal regions also exhibit representational aspects of conceptual and perceptual stimuli (Devereux et al., 2013; Fairhall and Caramazza, 2013; Kuhl and Chun, 2014; Lee and Kuhl, 2016). For each of these five regions, analyses were run within separate left and right hemisphere ROIs.

fMRI data: pattern similarity measures. For each ROI, voxelwise univariate activity patterns were extracted from single-trial β images of concept repeat encoding trials and corresponding target and lure trials at retrieval. Trialwise patterns were then correlated with one another to produce encoding similarity measures (Fisher transformed Pearson's r) and encoding-retrieval similarity measures for each concept.

For true memory analyses, concept-specific encoding similarity measures were constructed to assess whether memory for a given target item could be predicted solely by the relationships between its concept-matched exemplars, which might reflect a measure of activation strength for the given concept. Critically, this measure captured the averaged similarity values between all pairs of exemplars of a given concept, excluding pairs involving the subsequently-presented target (Fig. 2*A*, average of non-hatched cells). In this way, the target measure was composed entirely from data in runs other than those containing the given concept's subsequent target exemplar (because each exemplar was in a separate run). In contrast to typical subsequent memory measures that involve the subsequently presented stimulus, this concept-by-concept metric minimized the contribution of stimulus processing for the individual target item and instead sought to determine whether memory was related to concept-level activation. Corresponding global encoding similarity values reflected the relationship between exemplars of different concepts. For a given concept, global encoding similarity encompassed the average of all pairwise combinations between that concept's exemplars (again excluding the target) and all (nontarget) exemplars of every other concept (Fig. 2*B*, non-hatched cells). Concept-specific pairings were excluded from this global encoding similarity measure. Concept-wise difference measures were created by subtracting global from concept-specific encoding similarity.

Finally, for each concept we calculated three encoding-retrieval similarity measures that involved pattern similarity between target trials at recognition and encoding trials. First, an identical encoding-retrieval similarity measure simply captured the correlation between the retrieval target item and the initial appearance of that same exemplar at encoding. Second, for concept-specific encoding-retrieval similarity, all five pairwise combinations between the recognition target item and the (non-identical) encoding exemplars of the same concept were averaged. Third, global encoding-retrieval similarity averaged the pairwise similarity values between the target and all repeated concept encoding exemplars belonging to different concepts. Three subtraction measures (identical > concept-specific, identical > global, concept-specific > global) were then generated.

Similarity measures for false memory analyses were computed in a similar fashion, with several differences. Concept-specific encoding similarity was calculated for each concept by taking the mean similarity value across all pairwise combinations (15 in total) of the six exemplars for a given concept. Global encoding similarity was calculated by averaging the pairwise encoding similarity values between all exemplars of a given concept and the exemplars from every other concept (excluding pairings from trials in the same encoding run). For each concept, a difference measure was calculated by subtracting the global encoding similarity value from the concept-specific encoding similarity value. Encoding-retrieval similarity values were also calculated using a method similar to that described for target trials, except that concept-specific encoding-retrieval similarity was composed of all six pairings between the lure and corresponding encoding exemplars. Because lures were by definition not previously encoded, no identical encoding-retrieval similarity level existed for false memory analyses.

Using these concept-wise similarity measures, separate analyses were run to assess effects related to true memory (dependent variable = 4-point memory response for target trials) and false memory (dependent variable = 4-point memory response for lure trials). For both encoding similarity and encoding-retrieval similarity, separate regressions tested

the effect of the different levels of similarity (identical, where applicable, as well as concept-specific and global) before differences between these levels were examined. In all regression analyses, predictor values were standardized, and significantly positive β estimates signified a positive relationship between the similarity measure and subjective mnemonic oldness (i.e., similarity predicting true memory for regressions using target items, or predicting false memory for regressions using lure items).

For random effects testing, subject-wise β estimates from each regression were entered into a separate 2 × 5 repeated-measures ANOVA with factors of hemisphere (left/right) and ROI (EVC/LVC/DPC/ANG/HC). Along with main effects and interactions, effects in individual ROIs were examined by submitting β estimates (which captured the relationship between brain similarity and memory outcome) to two-tailed t tests, given that individual ROIs could show meaningful memory-related differences even in the absence of significant F tests. Multiple comparisons were addressed through Bonferroni correction based on the number of ROIs tested (10 ROIs, yielding a corrected threshold of 0.005).

fMRI data: hippocampal interaction analysis. For both target and lure responses, we also tested for the possibility that subsequent memory might be the product of concurrent changes in concept-specific encoding similarity across different ROIs (i.e., a between ROI interaction in concept-specific encoding similarity). These analyses focused on cross ROI interactions involving the hippocampus, given the role of this region in orthogonalizing overlapping inputs from cortical regions. Additional regressions were therefore run where the predictor-of-interest was the interaction of concept-specific encoding similarity between a pair of ROIs, rather than within a single ROI (for an example with lure memory, see Fig. 4*A*). Interaction terms were formed by first z-scoring the series of concept-specific encoding similarity values for each of the two ROIs in the pair (hippocampal ROI and cortical ROI) and computing a concept-wise product (left/right hippocampal ROIs were paired with bilateral cortical ROIs to reduce the number of pairwise comparisons to 8, which yielded a Bonferroni-corrected threshold of 0.0063). In addition to this interaction term, each regression also included the constituent standardized encoding similarity values for the separate ROIs forming the interaction. For the false memory regressions that yielded a significant result, the specific nature of the interaction was visualized by dividing trials into high and low similarity bins (median split on concept-specific encoding similarity measure) for both the cortical and hippocampal ROIs and plotting the average lure memory score for each set of trials (see Fig. 4*B*). To ensure that this interaction reflected the relationship between concurrent cortical and hippocampal patterns within exemplars of a given concept only, we also reran the interaction analysis using a global encoding similarity measure for the hippocampus.

Results

Behavioral results

Accuracy during the category detection encoding task was near ceiling, as expected. Participants correctly categorized objects as non-instruments at a high rate (M = 0.984, SEM = 0.04), and also successfully detected instruments (M = 0.870, SEM = 0.016). Collapsing across runs, reaction times for correct repeated concept trials were slightly but significantly faster than concepts that were not repeated (repeated concept trials: M = 580.30 ms, SEM = 12.20 ms; non-repeated concept trials, M = 570.1, SEM = 12.43; paired $t_{(28)} = 2.468$, $p = 0.002$, $d = 0.644$). This difference reflected the decrease in reaction times across runs for the repeated concept condition, suggesting cross-form priming effects (Koutstaal et al., 2001; Simons et al., 2003). The linear change in reaction times across runs was tested with a 2 × 6 repeated measure ANOVA (condition: repeated/non-repeated by encoding run: 1–6). The test on this linear trend showed a significant interaction between conditions ($F_{(1,28)} = 7.531$, $p = 0.01$), confirming that facilitation across runs was greater in the repeated concept condition, where exemplars corresponded to previously-shown concepts.

Behavioral performance during the scanned object recognition memory test showed an increased false alarm rate (collapsing high and low confidence responses) for repeated concept lures (M = 0.415, SEM = 0.032) compared with non-repeated concept lures (M = 0.265, SEM = 0.022). When considered alongside the corresponding confidence-collapsed hit rates (repeated condition: M = 0.532, SEM = 0.033; non-repeated condition: M = 0.431, SEM = 0.033), corrected recognition (hits − false alarms) was slightly higher (paired $t_{(28)} = -2.065$, $p = 0.049$, d = −0.383) in the non-repeated (M = 0.166, SEM = 0.022) versus repeated condition (M = 0.117, SEM = 0.014). The combined false alarm rate for lure items across both conditions (M = 0.340, SEM = 0.026) was substantially higher (paired $t_{(28)} = 11.376$, $p < 0.001$, $d = 2.112$) than that for novel items (M = 0.093, SEM = 0.019), confirming reliable false memory for lures. A final within-subjects analysis focusing on picture recognition trials for concept repeat items found a small association between successful word recognition of a concept and the likelihood that its corresponding picture was judged old (group mean correlation between word recognition and picture target hit rate: $r = 0.143$, SEM = 0.032; between word recognition and picture lure false alarm rate: $r = 0.127$, SEM = 0.34) with no difference in this relationship between target and lure items (paired t test on subjectwise correlations: $t_{(28)} = 0.492$, $p = 0.62$, $d = 0.091$).

Pattern similarity related to true memory

Encoding similarity: concept-specific and global measures

Although global encoding similarity may contribute strongly to recognition of items from heterogeneous encoding sets, the repetition of concept exemplars in the present design allowed for a measure of conceptual activation strength specific to the exemplars of a given concept. Unlike most subsequent memory analyses, which by definition examine the encoding signatures of subsequently-repeated items, we sought to determine whether memory for a specific target exemplar could be predicted solely on the basis of similarity between other exemplars of the same concept. Therefore, for the purposes of predicting target memory, the concept-specific encoding similarity measure excluded pairwise comparisons involving the subsequent target. We first ran a laterality (left/right) × region (5 regions) ANOVA on β estimates from a regression where concept-specific encoding similarity predicted later true memory. A main effect of region emerged ($F_{(4,112)} = 3.202$, $p = 0.016$, partial $\eta^2 = 0.103$; no significant main effect of laterality or interaction: p values > 0.1), with follow-up tests revealing two regions showing significant effects: left LVC ($t_{(28)} = 4.362$, $p < 0.001$, $d = 0.810$) and left ANG ($t_{(28)} = 3.196$, $p = 0.004$, $d = 0.593$; Fig. 3*A*). Effects of a similar direction were also present in right EVC and LVC but were not significant at a corrected threshold (Table 1). These results show that in both late visual regions and angular gyrus, higher pattern overlap between exemplars of a given concept predicted later memory for the held out subsequent target.

We next tested how this concept-specific encoding similarity measure compared with a global encoding similarity measure that also excluded all items subsequently serving as recognition targets. An ANOVA with parameter estimates capturing global encoding similarity showed no main effects or interactions (all p values > 0.1), although significant effects were again observed in left LVC ($t_{(28)} = 5.055$, $p < 0.001$, $d = 0.939$) and left ANG ($t_{(28)} = 3.518$, $p < 0.001$, $d = 0.653$). Notably however, the concept-specific > global encoding similarity difference measure suggested that concept-specific encoding similarity may have been a greater predictor of memory than global encoding similarity. A

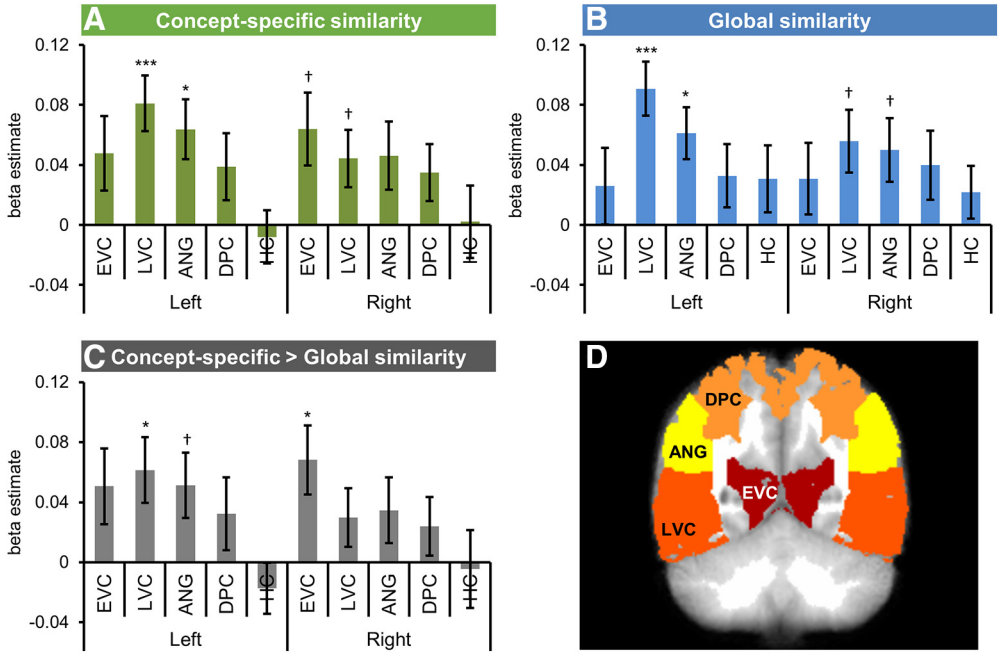

Figure 3. Encoding similarity predicting subsequent target memory. Beta estimates from regression analyses with (***A***) concept-specific encoding similarity, (***B***) global encoding similarity, or (***C***) the difference between these measures predicting subsequent memory. †$p < 0.05$, *$p < 0.01$, ***$p < 0.001$. ***D***, Analyses were run in a subset of ROIs covering posterior cortical regions and in hippocampal ROIs (not depicted), separately within each hemisphere (left/right).

Table 1. Encoding similarity measures predicting subsequent true memory

Encoding similarity ROI	Beta estimate		*t*		*p*	
	Left	Right	Left	Right	Left	Right
Concept-specific						
EVC	0.048	0.064	1.921	2.634	0.065	0.014
LVC	0.081	0.044	4.362	2.317	<0.001	0.028
ANG	0.064	0.046	3.194	2.027	0.003	0.052
DPC	0.039	0.035	1.735	1.831	0.094	0.078
HC	−0.008	0.002	−0.448	0.091	0.657	0.928
Global						
EVC	0.026	0.031	1.010	1.287	0.321	0.209
LVC	0.091	0.056	5.054	2.658	<0.001	0.013
ANG	0.061	0.050	3.518	2.350	0.002	0.026
DPC	0.033	0.040	1.560	1.729	0.13	0.095
HC	0.031	0.022	1.369	1.239	0.182	0.226
Concept-specific > Global						
EVC	0.051	0.068	2.010	2.956	0.054	0.006
LVC	0.061	0.030	2.806	1.531	0.009	0.137
ANG	0.051	0.035	2.363	1.585	0.025	0.124
DPC	0.032	0.024	1.334	1.223	0.193	0.232
HC	−0.017	−0.004	−1.023	−0.172	0.315	0.865

Positive estimates signify positive relationship between similarity measure and subsequent oldness ratings to targets.

significant main effect of region ($F_{(4,112)} = 3.542$, $p < 0.001$, partial $\eta^2 = 0.112$; no significant main effect of laterality or interaction: p values > 0.1) appeared from the ANOVA on the concept-specific versus global difference measure. *Post hoc* tests in right EVC ($t_{(28)} = 2.956$, $p = 0.006$, $d = 0.549$) and left LVC ($t_{(28)} = 2.806$, $p = 0.009$, $d = 0.521$; Fig. 3*C*) identified these regions as showing the strongest memory-related difference between concept-specific and global encoding similarity, although effects fell short of the corrected threshold. These findings indicate that, along with the contribution of broader global similarity, the relationships between concept-matched exemplars may be particularly important for determining later recognition of a given target item.

Analyses focused on hippocampal-cortical interactions in the concept-specific encoding similarity (discussed in more detail in false memory results section) did not show any significant cross-region interactions (all p values > 0.05).

Encoding-retrieval similarity: identical, concept-specific, and global measures

Although our focus in the current study was on the relationship between exemplars at encoding, we also analyzed target memory with respect to encoding-retrieval similarity. ANOVAs on all three levels of encoding-retrieval similarity returned significant main effects of region (identical: $F_{(4,112)} = 3.487$, $p < 0.010$, partial $\eta^2 = 0.111$; concept-specific: $F_{(4,112)} = 3.494$, $p = 0.010$, partial $\eta^2 = 0.111$; global: $F_{(4,112)} = 7.531$, $p < 0.001$, partial $\eta^2 = 0.212$) but no main effects of laterality or laterality by region interactions (all p values > 0.05). Follow-up tests (Table 2, top half) showed that although regions including EVC, LVC, and DPC showed identical encoding-retrieval memory effects, within-concept effects were limited to LVC. Global encoding-retrieval similarity also increased with target memory in LVC, as well as in DPC.

We next compared these levels of encoding-retrieval similarity against each other (Table 2, bottom half). Every ROI showing significant memory-related effects for identical encoding-retrieval similarity remained significant when global encoding-retrieval similarity was subtracted (identical > global encoding-retrieval similarity; main effect of region: $F_{(4,112)} = 2.954$, $p = 0.023$, partial $\eta^2 = 0.095$; no main effect of laterality or laterality by region interaction: p values > 0.1). This result replicates past findings from scene recognition showing that memory-related reinstatement effects are particularly strong in regions of visual cortex (Ritchey et al., 2013).

An even closer comparison was examined by subtracting concept-specific encoding-retrieval similarity from identical similarity (identical > concept-specific encoding-retrieval similarity). This measure isolated memory effects linked to the reca-

Table 2. Encoding-retrieval similarity measures predicting true memory

	Beta estimate		*t*		*p*	
ROI	Left	Right	Left	Right	Left	Right
Encoding-retrieval similarity (single measures)						
Identical						
EVC	0.093	0.065	4.214	2.885	<0.001	0.007
LVC	0.106	0.108	4.382	4.732	<0.001	<0.001
ANG	0.047	0.057	1.865	2.660	0.073	0.013
DPC	0.078	0.085	4.001	4.580	<0.001	<0.001
HC	0.049	0.023	2.424	1.054	0.022	0.301
Concept-specific						
EVC	0.044	0.057	1.940	2.124	0.062	0.043
LVC	0.063	0.078	3.410	5.158	0.002	<0.001
ANG	0.024	0.055	1.028	2.732	0.313	0.011
DPC	0.025	0.036	1.326	1.760	0.196	0.089
HC	0.013	−0.035	0.532	−1.562	0.599	0.130
Global						
EVC	0.010	0.032	0.422	1.150	0.676	0.260
LVC	0.103	0.087	4.658	3.912	<0.001	0.001
ANG	0.051	0.061	2.263	2.728	0.032	0.011
DPC	0.092	0.091	4.888	4.819	<0.001	<0.001
HC	0.004	0.001	0.236	0.033	0.815	0.974
Encoding-retrieval similarity (comparison of measures)						
Identical > Concept-specific						
EVC	0.069	0.037	3.185	1.515	0.004	0.141
LVC	0.072	0.064	2.742	2.736	0.011	0.011
ANG	0.033	0.029	1.264	1.373	0.217	0.181
DPC	0.061	0.067	2.794	3.351	0.009	0.002
HC	0.037	0.037	1.524	1.821	0.139	0.079
Identical > Global						
EVC	0.096	0.065	4.665	3.010	<0.001	0.005
LVC	0.094	0.098	3.886	4.231	0.001	<0.001
ANG	0.041	0.052	1.691	2.373	0.102	0.025
DPC	0.069	0.076	3.495	4.071	0.002	<0.001
HC	0.050	0.022	2.447	1.050	0.021	0.303
Concept-specific > Global						
EVC	0.051	0.060	2.237	2.301	0.033	0.029
LVC	0.039	0.061	2.155	4.007	0.040	<0.001
ANG	0.011	0.044	0.461	2.174	0.649	0.038
DPC	0.004	0.014	0.185	0.681	0.855	0.501
HC	0.015	−0.038	0.640	−1.720	0.527	0.096

Positive estimates signify positive relationship between similarity measure and oldness ratings to targets.

Table 3. Encoding similarity and encoding-retrieval similarity predicting false memory

	Beta estimate		*t*		*p*	
ROI	Left	Right	Left	Right	Left	Right
Encoding similarity						
Concept-specific						
EVC	−0.020	−0.002	−0.952	−0.119	0.349	0.906
LVC	0.014	−0.012	0.739	−0.598	0.466	0.555
ANG	0.013	−0.019	0.587	−1.265	0.562	0.216
DPC	−0.014	−0.019	−0.637	−0.714	0.529	0.481
HC	0.026	−0.001	1.290	−0.070	0.208	0.945
Global						
EVC	−0.016	0.013	−0.745	0.756	0.462	0.456
LVC	0.039	0.022	1.987	0.951	0.057	0.350
ANG	0.020	−0.006	1.235	−0.244	0.227	0.809
DPC	0.012	0.002	0.606	0.076	0.550	0.940
HC	0.033	0.008	1.738	0.324	0.093	0.749
Concept-specific > Global						
EVC	−0.015	−0.009	−0.731	−0.423	0.471	0.676
LVC	0.001	−0.026	0.063	−1.242	0.950	0.224
ANG	0.008	−0.019	0.342	−1.342	0.735	0.190
DPC	−0.020	−0.027	−0.947	−0.995	0.352	0.328
HC	0.022	−0.003	1.035	−0.133	0.309	0.895
Encoding-retrieval similarity						
Concept-specific						
EVC	0.016	−0.010	0.867	−0.780	0.393	0.442
LVC	0.019	0.056	0.895	2.409	0.379	0.023
ANG	0.009	0.031	0.504	1.473	0.618	0.152
DPC	0.024	0.001	1.173	0.045	0.251	0.965
HC	0.038	0.048	1.688	2.970	0.103	0.006
Global						
EVC	−0.021	−0.029	−0.969	−1.537	0.341	0.135
LVC	0.016	0.030	0.662	1.350	0.514	0.188
ANG	0.006	0.032	0.195	1.055	0.847	0.301
DPC	0.034	−0.003	1.181	−0.090	0.248	0.929
HC	0.036	0.034	1.595	1.697	0.122	0.101
Concept-specific > Global						
EVC	0.034	0.004	1.672	0.282	0.106	0.780
LVC	0.019	0.056	0.845	2.313	0.405	0.028
ANG	0.011	0.029	0.617	1.505	0.542	0.144
DPC	0.017	0.002	0.885	0.106	0.384	0.916
HC	0.029	0.043	1.302	2.649	0.204	0.013

Positive estimates signify positive relationship between similarity measure and oldness ratings to lures.

pitulation of the specific target item rather than other exemplars of that same concept. Notably, although *F* tests on the ANOVA using this measure were not significant (p values > 0.1), in left EVC ($t_{(28)} = 3.186$, $p = 0.004$, $d = 0.592$) and right DPC ($t_{(28)} = 3.531$, $p = 0.002$, $d = 0.623$), successful recognition was accompanied by reinstatement of neural patterns specific to the target exemplar rather than the other exemplars of that concept. Bilateral LVC also showed effects in this direction, but did not survive correction for multiple comparisons (Table 2, Identical > Concept-specific).

The final subtraction measure (concept-specific > global encoding-retrieval similarity) produced a significant main effect of region ($F_{(4,112)} = 3.032$, $p = 0.020$, partial $\eta^2 = 0.098$) and region by laterality interaction ($F_{(4,112)} = 2.827$, $p = 0.028$, partial $\eta^2 = 0.092$), but no main effect of laterality ($p > 0.1$). This pattern of effects was driven by right LVC, where reinstatement of patterns corresponding to a target's concept-matched exemplars tracked memory even after global measures had been subtracted. Thus, although it appears that successful recognition is related to reinstatement of patterns at all three levels, memory effects at the highest level of specificity are seen in early visual regions and DPC, whereas effects in later visual cortex may reflect information common to other exemplars of a given target concept.

Pattern similarity related to false memory

Encoding similarity: concept-specific and global measures

To test whether higher similarity across exemplars of a given concept led to later difficulty in distinguishing lure items at test (i.e., increased the chance of subsequent false alarms), we first computed the mean concept-specific encoding similarity for each ROI and used these measures to predict lure memory. Beta estimates from regressions where this similarity value predicted subsequent false memory were analyzed in the context of a repeated-measures ANOVA, with factors of laterality (left/right) and ROI (5 regions). Results from this ANOVA showed neither a significant laterality by region interaction nor significant main effects (all p values > 0.05), nor were memory-related effects significant in any single region (all p values > 0.05; Table 3, top). Corresponding tests from a second ANOVA run with regression values testing the relationship of global encoding similarity and memory were also not significant (all p values > 0.05).

Because encoding similarity might further depend on the function of the hippocampus, we tested whether subsequent lure memory related to the interplay between concept-specific cortical similarity and the overlap or separation of corresponding hippocampal patterns (i.e., an interaction between cortical and hippocampal concept-specific encoding similarity). We com-

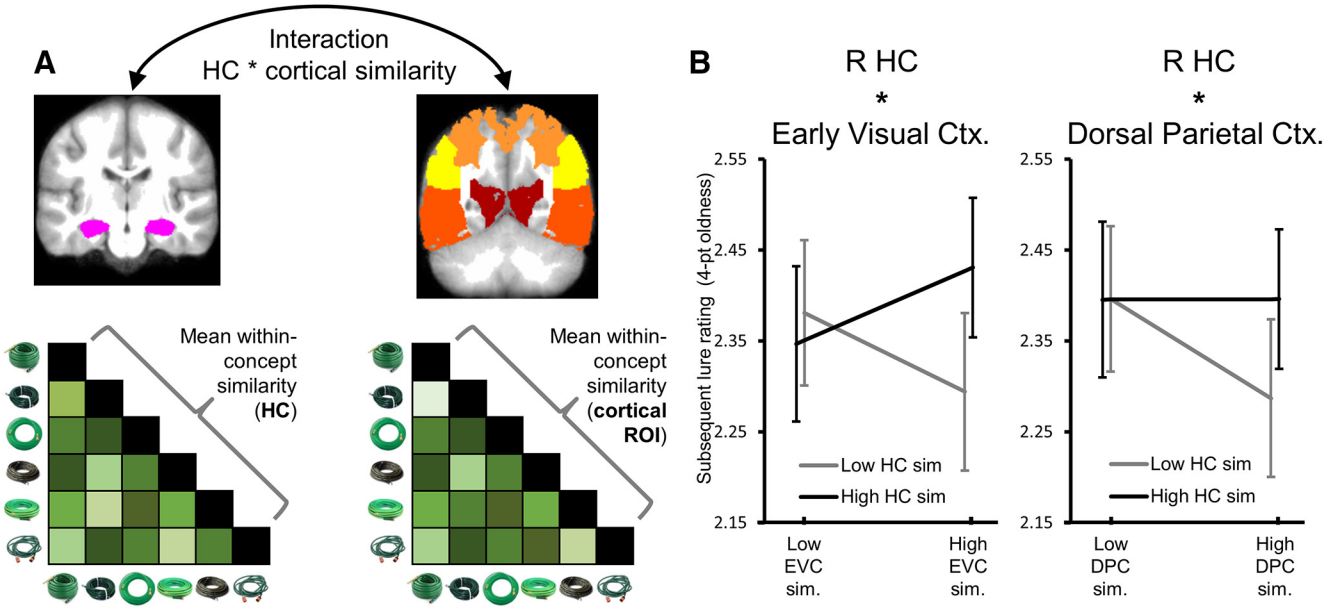

Figure 4. Cross-region encoding similarity interactions and relationship to subsequent lure memory. ***A***, Concept-specific encoding similarity was calculated by averaging the pairwise pattern correlations between each of the six exemplars of a given concept. The cross ROI interaction of these concept-by-concept measures was computed for ROI pairs involving the hippocampus and other cortical regions. These interaction values were used to predict subsequent lure memory. ***B***, Illustration of interaction in EVC and DPC, with average lure oldness rating (i.e., higher for increasing false alarms) plotted for concepts with high/low concept-specific encoding similarity in cortex (error bars reflect SE). In EVC, high concept-specific representational overlap is associated with subsequent false alarms for concepts where hippocampal similarity is also high. False alarms are reduced when high similarity in both EVC and DPC is accompanied by lower similarity between corresponding hippocampal representations.

puted cross ROI interaction terms by multiplying concept-wise similarity measures in the hippocampus (separately for left and right hemisphere ROIs) with those in bilateral cortical ROIs (Fig. 4*A*; Materials and Methods). Parameter estimates for these cross-ROI interaction terms were entered into a 2 × 4 ANOVA with factors of hippocampal laterality and cortical region. A significant main effect of hippocampal laterality was present ($F_{(3,28)} = 4.938$, $p = 0.035$, partial $\eta^2 = 0.150$), with no significant main effect of region or region by laterality interaction (p values > 0.1). *Post hoc* tests revealed a significant cross-ROI interaction effect for two pairs involving right hippocampus: right HC to DPC ($t_{(28)} = 3.602$, $p = 0.003$, $d = 0.669$) and right HC to EVC ($t_{(28)} = 3.027$, $p = 0.005$, $d = 0.562$). These significant interactions indicate that subsequent false memory is not influenced by cortical similarity alone but depends on the structure of corresponding exemplar representations in hippocampus. In both regions, high cortical overlap accompanied by differentiated hippocampal representations (i.e., low concept-specific encoding similarity) is associated with reduced subsequent false alarms. As illustrated in Figure 4*B*, the subsequent false alarm rate is the highest for concepts that have high similarity in EVC while also having undifferentiated hippocampal representations. High cortical similarity accompanied by more dissimilar hippocampal patterns was instead associated with subsequent correct rejections, a pattern also observed in DPC. A corresponding interaction analysis for true memory (i.e., predicting target memory) returned no significant interactions (all p values > 0.05).

As a final test to ensure that the observed false memory interaction effects reflected the influence of elevated hippocampal similarity specifically for the individual concept exemplars whose cortical similarity was also high, we repeated the interaction analysis using the global encoding similarity measure for hippocampus. Here, interaction terms were created by taking the product of concept-specific encoding similarity for a given bilateral cortical ROI, as before, but global encoding similarity for the hippocampus (i.e., similarity between a given concept and all other encoding concepts). Notably, interaction effects computed with this less-specific measure of hippocampal encoding similarity did not also predict subsequent false memory (p values > 0.05), indicating that effective separation of overlapping cortical inputs involves the formation of dissimilar hippocampal representations that correspond to concurrent cortical patterns.

In sum, it appeared that encoding concept similarity within any given ROI was not sufficient, in and of itself, to meaningfully predict later lure memory. Instead, increased concept similarity in visual and parietal ROIs was associated with the subsequent success or failure of lure discrimination only through associated concept-level patterns in the hippocampus. Although influential models of memory propose that the hippocampus operates on distributed cortical inputs (Treves and Rolls, 1994; McClelland et al., 1995; Norman and O'Reilly, 2003), this is, to our knowledge, the first fMRI evidence linking behavioral episodic memory outcomes to an interaction between concurrent representations in the hippocampus and the cortex. Furthermore, from a methodological standpoint, these results illustrate the potential of a representational analog to functional connectivity analysis, which instead of univariate activity examines covariation in representational structure corresponding to sets of trials across an experiment.

Encoding-retrieval similarity: concept-specific and global measures

Additional analyses probed whether false memory was associated with increased pattern similarity between lures at retrieval and various encoding items (Table 3, bottom). An ANOVA for within-concept encoding-retrieval similarity (mean similarity between a lure and corresponding encoding exemplars of the same concept) returned a significant laterality by region interaction ($F_{(4,112)} = 2.484$, $p = 0.048$, partial $\eta^2 = 0.082$; no significant

main effects: p values > 0.1). *Post hoc* tests identified trend-level effects in right LVC ($t_{(28)} = 2.409, p = 0.023, d = 0.447$) and right HC ($t_{(28)} = 2.970$, $p = 0.006$, $d = 0.551$). A test of global encoding-retrieval similarity values yielded no significant main effects or interaction (all p values > 0.1). In a direct comparison between the two levels of similarity (concept-specific $>$ global encoding-retrieval similarity; all F tests: $p > 0.05$), neither right LVC nor right HC survived correction for multiple comparisons (right LVC: $t_{(28)} = 2.313, p = 0.028, d = 0.430$; right HC: $t_{(28)} = 2.649$, $p = 0.013$, $d = 0.492$). Although these results present an equivocal picture, they suggest that, especially in hippocampus, false alarms to lures may be accompanied by reinstatement of encoding patterns associated with concept-matched exemplars.

Discussion

The current study explored how the representational structure of highly similar encoding exemplars related to true and false recognition. We examined conceptual activation strength at encoding by measuring the representational overlap among exemplars of the same concept (concept-specific encoding similarity) and compared this to a set-wide between-concept measure (global encoding similarity). Concept-specific encoding similarity strongly tracked true memory, and subsequent recognition of a target object could be predicted by the degree of pattern overlap between its concept-matched exemplars in angular gyrus and late visual cortex. Increased concept-specific encoding similarity in dorsal parietal and early visual processing regions also related to subsequent false memory, but only through an interaction with corresponding hippocampal patterns. Subsequent false alarms were reduced for concepts where high cortical overlap was accompanied by differentiated hippocampal patterns. This novel finding suggests that the fine-grain discrimination necessary to resist false recognition relates to the success or failure of the hippocampus in disambiguating attendant cortical representations. We discuss the implications of these findings, and their relationships to past work in the following sections.

Pattern similarity related to true memory

Recognition memory is known to depend on the relationship between elements in an encoding set, and this may be particularly true when strong concept representations are generated by the repetition of visosemantic information across items. In support of this idea, we found that greater cortical similarity between nontarget exemplars of a given concept (i.e., concept-specific encoding similarity) was associated with later recognition of that concept's target image, particularly in left angular gyrus and late visual cortex (Fig. 3*B*).

The present results offer some insight into related work on self-similarity, which explores overlap between multiple presentations of an identical stimulus at encoding. This work has found that subsequent memory is associated with higher self-similarity in both frontoparietal (Xue et al., 2010, 2013) and occipitotemporal (Xue et al., 2010; Ward et al., 2013; Hasinski and Sederberg, 2016) regions. The widespread neuroanatomical distribution of these effects suggests that it is not merely the consistency of early perceptual processing across identical presentations that benefits memory. However, the degree to which strengthening of a concept-level representation might influence later memory is hard to quantify given that both stimulus form and meaning are identical in self-similarity analyses. Our data suggest that self-similarity effects may capture a broader conceptual-level strengthening of repeated stimuli in addition to stability of perceptual processing. While the current findings emphasize the beneficial mnemonic consequences of strengthened representations at the level of individual concepts, an important caveat is that distinct exemplars of a given concept also typically overlap in visual form, with perceptually-related brain and stimulus similarity likely contributing to the observed memory effects along with conceptual dimensions.

Although both true memory and false memory were associated with higher concept-specific encoding similarity, only in the latter was this relationship contingent upon concurrent representational overlap in the hippocampus. The absence of hippocampal involvement for true memory is somewhat surprising given various encoding-related findings that the hippocampus supports subsequent cortical reactivation (Gordon et al., 2014; Wing et al., 2015) and even exemplar-level discrimination (van den Honert et al., 2016). In one particularly relevant study where global encoding similarity reflected the relationship between a given word and all other encoding-set words, higher hippocampal pattern overlap positively predicted later memory (Davis et al., 2014). However, better subsequent memory has also been tied to lower hippocampal similarity (i.e., distinctiveness) among items of the same overall category (e.g., faces, scenes, objects; LaRocque et al., 2013). One possibility is that when stimulus sets are conceptually heterogeneous and more uniformly distributed (e.g., lists of unrelated words), overlap benefits memory through general inter-item familiarity, whereas the presence of stimuli that cluster into discrete categories may instead spontaneously promote hippocampal separation processes. However, this influence may have been diminished in the present study if participants coded object exemplars as part of an overarching (non-instrument) category, given past findings that task orientation biases representations in hippocampus and neocortex (Çukur et al., 2013; Aly and Turk-Browne, 2016).

Successful recognition was also related to increased overlap between recognition targets and corresponding encoding items. These results (Table 2) broadly mirror past findings of encoding-retrieval similarity (Staresina et al., 2012; Ritchey et al., 2013; Wing et al., 2015; Xiao et al., 2017), and further show that in early visual and dorsal parietal regions, true memory entails recapitulation of patterns specific enough to differentiate exemplars of the same concept. One potential implication of this finding is that when encoding targets are directly repeated at retrieval, the precision of the cortical match may be sufficient to drive accurate memory, even if related exemplars were not well differentiated at encoding. In cases when memory judgments instead operate on lure items (discussed in the next section), the absence of perceptual repetition may make accurate memory judgments (i.e., correct rejections) more dependent on the ability of the hippocampus to separate overlapping cortical inputs.

Pattern similarity related to false memory

Concept-specific encoding similarity in several cortical regions was found to promote later target recognition, and we predicted that the same type of overlap might also reflect undifferentiated exemplar representations, leading to subsequent false memory. A recurrent finding in the neuroimaging literature is that activity associated with false memory is often found in later visual processing regions, with early visual regions instead tracking true memory (Schacter and Slotnick, 2004; Stark et al., 2010; Dennis et al., 2012). This pattern of findings may reflect the increased influence of top-down processing in higher-level sensory regions, and generally accords with false memory findings in the semantic domain; e.g., that representations in anterior temporal regions during word encoding relate to later false memories (Chadwick et

al., 2016; Zhu et al., 2019). However, similarity between low-level visual stimulus properties has been found to influence subsequent false alarms in short-term memory tests (Kahana et al., 2007), and recent neuroimaging work has now implicated early visual regions in false memory judgments (Karanian and Slotnick, 2017, 2018). Interestingly, no region, visual or otherwise, showed a direct relationship between cortical similarity and false memory in the present data. Instead, the influence of cortical overlap between a concept's exemplars depended on concurrent patterns in the hippocampus, and reinstatement of concept-specific patterns in hippocampus also tracked false alarms.

The present cross-region interactions at encoding (Fig. 4) show that hippocampal pattern differentiation promotes accurate lure discrimination specifically when cortical similarity is high. High cortical overlap in this context may reflect the same type of strong, concept-level activation thought to underpin the beneficial link between encoding similarity and target memory, where effects were broad and not hippocampally-contingent. It is possible that in both cases, concept strengthening occurs through the incidental retrieval of previously-seen concept exemplars during encoding. Past work has linked pattern similarity across repetition and explicit retrieval practice to later memory outcomes (Kuhl and Chun, 2014; Karlsson Wirebring et al., 2015; van den Honert et al., 2016). Notably, a recent study examined reactivation in parietal cortex during associative retrieval and found that reinstatement of item-level representations was linked to successful lure discrimination on a subsequent test, whereas more general category reinstatement predicted false alarms (Lee et al., 2018). Although Lee et al. (2018) found that cortical specificity at retrieval predicts lure discrimination, the current data highlight the further influence of the hippocampus in determining how cortical overlap influences fine-grain memory.

The results also showed tentative evidence that higher encoding-retrieval similarity between lures and concept-matched encoding exemplars tracked false memory (Table 3, concept-specific encoding-retrieval similarity). Although future confirmatory work will be necessary, the present cross-phase hippocampal finding presents a parallel with the encoding-related interaction and suggests that the lack of differentiated hippocampal patterns at encoding may cause a generalized trace to be recovered when participants encounter lure items. Several previous reports have implicated the hippocampus in formation of false memories (Okado and Stark, 2005; Pidgeon and Morcom, 2016) or compared its role in true and false memory retrieval (Cabeza et al., 2001; Dennis et al., 2012; Gutchess and Schacter, 2012). The present findings suggest that false memories may arise from a failure to engage specific mechanisms (McClelland et al., 1995; Norman, 2010) that, when operating effectively, allow for the encoding of individuated representations necessary for accurate memory.

Conclusion

Past theoretical and empirical work has shown that the similarity between items in memory has important implications for subsequent recognition. Although prior research has primarily explored true memory, in the present study we demonstrated that a more focused measure of concept-specific encoding similarity relates not only to true memory, but also to false memory during fine-grain memory discriminations. For predicting false memory, increased concept-specific cortical overlap also depended on the nature of corresponding hippocampal representations: when high cortical similarity was accompanied by differentiated hippocampal patterns, a given concept's lure was more likely to be successfully discriminated. In sum, these results demonstrate how the hippocampus can resolve, or fail to resolve, overlap in cortical representations, which can lead to both true and false memory for concept-related information.

References

Abe N, Fujii T, Suzuki M, Ueno A, Shigemune Y, Mugikura S, Takahashi S, Mori E (2013) Encoding- and retrieval-related brain activity underlying false recognition. Neurosci Res 76:240–250.

Aly M, Turk-Browne NB (2016) Attention stabilizes representations in the human hippocampus. Cereb Cortex 26:783–796.

Bowman CR, Chamberlain JD, Dennis NA (2019) Sensory representations supporting memory specificity: age effects on behavioral and neural discriminability. J Neurosci 39:2265–2275.

Cabeza R, Rao SM, Wagner AD, Mayer AR, Schacter DL (2001) Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. Proc Natl Acad Sci U S A 98:4805–4810.

Chadwick MJ, Anjum RS, Kumaran D, Schacter DL, Spiers HJ, Hassabis D (2016) Semantic representations in the temporal pole predict false memories. Proc Natl Acad Sci U S A 113:10180–10185.

Clark SE, Gronlund SD (1996) Global matching models of recognition memory: how the models match the data. Psychon Bull Rev 3:37–60.

Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16:763–770.

Davis SW, Wing EA, Cabeza R (2018) Contributions of the ventral parietal cortex to declarative memory. Handb Clin Neurol 151:525–553.

Davis T, Xue G, Love BC, Preston AR, Poldrack RA (2014) Global neural pattern similarity as a common basis for categorization and recognition memory. J Neurosci 34:7472–7484.

Dennis NA, Bowman CR, Vandekar SN (2012) True and phantom recollection: an fMRI investigation of similar and distinct neural correlates and connectivity. Neuroimage 59:2982–2993.

Devereux BJ, Clarke A, Marouchos A, Tyler LK (2013) Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. J Neurosci 33:18906–18916.

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552–10558.

Garoff RJ, Slotnick SD, Schacter DL (2005) The neural origins of specific and general memory: the role of the fusiform cortex. Neuropsychologia 43:847–859.

Gillund G, Shiffrin RM (1984) A retrieval model for both recognition and recall. Psychol Rev 91:1–67.

Gonsalves B, Reber PJ, Gitelman DR, Parrish TB, Mesulam MM, Paller KA (2004) Neural evidence that vivid imagining can lead to false remembering. Psychol Sci 15:655–660.

Gordon AM, Rissman J, Kiani R, Wagner AD (2014) Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. Cereb Cortex 24:3350–3364.

Gutchess AH, Schacter DL (2012) The neural correlates of gist-based true and false recognition. Neuroimage 59:3418–3426.

Hasinski AE, Sederberg PB (2016) Trial-level information for individual faces in the fusiform face area depends on subsequent memory. Neuroimage 124:526–535.

Hintzman DL (1988) Judgments of frequency and recognition memory in a multiple-trace memory model. Psychol Rev 95:528–551.

Kahana MJ, Zhou F, Geller AS, Sekuler R (2007) Lure similarity affects visual episodic recognition: detailed tests of a noisy exemplar model. Mem Cognit 35:1222–1232.

Karanian JM, Slotnick SD (2017) False memories for shape activate the lateral occipital complex. Learn Mem 24:552–556.

Karanian JM, Slotnick SD (2018) Confident false memories for spatial location are mediated by V1. Cogn Neurosci 9:139–150.

Karlsson Wirebring L, Wiklund-Hörnqvist C, Eriksson J, Andersson M, Jonsson B, Nyberg L (2015) Lesser neural pattern similarity across repeated tests is associated with better long-term memory retention. J Neurosci 35:9595–9602.

Kim H, Cabeza R (2007) Differential contributions of prefrontal, medial temporal, and sensory-perceptual regions to true and false memory formation. Cereb Cortex 17:2143–2150.

Konkle T, Brady TF, Alvarez GA, Oliva A (2010) Conceptual distinctiveness

supports detailed visual long-term memory for real-world objects. J Exp Psychol Gen 139:558–578.
Koutstaal W, Wagner AD, Rotte M, Maril A, Buckner RL, Schacter DL (2001) Perceptual specificity in visual object priming: functional magnetic resonance imaging evidence for a laterality difference in fusiform cortex. Neuropsychologia 39:184–199.
Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.
Kuhl BA, Chun MM (2014) Successful remembering elicits event-specific activity patterns in lateral parietal cortex. J Neurosci 34:8051–8060.
Kurkela KA, Dennis NA (2016) Event-related fMRI studies of false memory: an activation likelihood estimation meta-analysis. Neuropsychologia 81:149–167.
LaRocque KF, Smith ME, Carr VA, Witthoft N, Grill-Spector K, Wagner AD (2013) Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. J Neurosci 33:5466–5474.
Lee H, Kuhl BA (2016) Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. J Neurosci 36:6069–6082.
Lee H, Samide R, Richter FR, Kuhl BA (2018) Decomposing parietal memory reactivation to predict consequences of remembering. Cereb Cortex 29:3305–3318.
McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev 102:419–457.
Mumford JA, Turner BO, Ashby FG, Poldrack RA (2012) Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59:2636–2643.
Norman KA (2010) How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus 20:1217–1227.
Norman KA, O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol Rev 110:611–646.
Nosofsky RM (1988) Exemplar-based accounts of relations between classification, recognition, and typicality. J Exp Psychol Learn Mem Cogn 14:700–708.
Nosofsky RM (1991) Tests of an exemplar model for relating perceptual classification and recognition memory. J Exp Psychol Hum Percept Perform 17:3–27.
Okado Y, Stark CE (2005) Neural activity during encoding predicts false memories created by misinformation. Learn Mem 12:3–11.
Pidgeon LM, Morcom AM (2014) Age-related increases in false recognition: the role of perceptual and conceptual similarity. Front Aging Neurosci 6:283.
Pidgeon LM, Morcom AM (2016) Cortical pattern separation and item-specific memory encoding. Neuropsychologia 85:256–271.
Ritchey M, Wing EA, LaBar KS, Cabeza R (2013) Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. Cereb Cortex 23:2818–2828.
Roediger HL, Watson JM, McDermott KB, Gallo DA (2001) Factors that determine false recall: a multiple regression analysis. Psychon Bull Rev 8:385–407.
Särkkä S, Solin A, Nummenmaa A, Vehtari A, Auranen T, Vanni S, Lin FH. (2012) Neuroimage dynamic retrospective filtering of physiological noise in BOLD fMRI: DRIFTER. Neuroimage 60:1517–1527.
Schacter DL, Slotnick SD (2004) The cognitive neuroscience of memory distortion. Neuron 44:149–160.
Simons JS, Koutstaal W, Prince S, Wagner AD, Schacter DL (2003) Neural mechanisms of visual object priming: evidence for perceptual and semantic distinctions in fusiform cortex. Neuroimage 19:613–626.
Slotnick SD, Schacter DL (2004) A sensory signature that distinguishes true from false memories. Nat Neurosci 7:664–672.
Spaniol J, Davidson PS, Kim AS, Han H, Moscovitch M, Grady CL (2009) Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. Neuropsychologia 47:1765–1779.
Staresina BP, Henson RN, Kriegeskorte N, Alink A (2012) Episodic reinstatement in the medial temporal lobe. J Neurosci 32:18150–18156.
Stark CE, Okado Y, Loftus EF (2010) Imaging the reconstruction of true and false memories using sensory reactivation and the misinformation paradigms. Learn Mem 17:485–488.
Treves A, Rolls ET (1994) Computational analysis of the role of the hippocampus in memory. Hippocampus 4:374–391.
van den Honert RN, McCarthy G, Johnson MK (2016) Reactivation during encoding supports the later discrimination of similar episodic memories. Hippocampus 26:1168–1178.
Ward EJ, Chun MM, Kuhl BA (2013) Repetition suppression and multivoxel pattern similarity differentially track implicit and explicit visual memory. J Neurosci 33:14749–14757.
Wing EA, Ritchey M, Cabeza R (2015) Reinstatement of individual past events revealed by the similarity of distributed activation patterns during encoding and retrieval. J Cogn Neurosci 27:679–691.
Xiao X, Dong Q, Gao J, Men W, Poldrack RA, Xue G (2017) Transformed neural pattern reinstatement during episodic memory retrieval. J Neurosci 37:2986–2998.
Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA (2010) Greater neural pattern similarity across repetitions is associated with better memory. Science 330:97–101.
Xue G, Dong Q, Chen C, Lu ZL, Mumford JA, Poldrack RA (2013) Complementary role of frontoparietal activity and cortical pattern similarity in successful episodic memory encoding. Cereb Cortex 23:1562–1571.
Yassa MA, Stark SM, Bakker A, Albert MS, Gallagher M, Stark CE (2010) High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic mild cognitive impairment. Neuroimage 51:1242–1252.
Ye Z, Zhu B, Zhuang L, Lu Z, Chen C, Xue G (2016) Neural global pattern similarity underlies true and false memories. J Neurosci 36:6792–6802.
Zhu B, Chen C, Shao X, Liu W, Ye Z, Zhuang L, Zheng L, Loftus EF, Xue G (2019) Multiple interactive memory representations underlie the induction of false memory. Proc Natl Acad Sci U S A 116:3466–3475.