Computational models suggest that human memory judgments exhibit interference due to the use of distributed representations

Derek J. Huffman*
Department of Psychology, Colby College, Waterville, ME 04901
*Corresponding author:
Email: derek.huffman@colby.edu
5568 Mayflower Hill Drive, Waterville, ME 04901

Ruijia Guan
Colby College, Waterville, ME 04901
University of Michigan, Ann Arbor, MI 48109

Keywords: episodic memory, pattern separation, distributed representations, hippocampus, dentate gyrus, distributed memory models, global matching models

1 Abstract

Episodic memory is a core function that allows us to remember the events of our lives. Given that many events in our life contain overlapping elements (e.g., similar people and places), it is critical to understand how well we can remember the specific events of our lives vs. how susceptible we are to interference between similar memories. Decades of research have aimed to uncover the mechanisms of memory using a combination of behavioral tasks, neuroscientific measures, and computational modeling. Several prominent theories converged on the notion that the brain employs distinct neural mechanisms to support different aspects of episodic memory. Many theories suggest that the hippocampus plays a greater role in processes such as recollection, associative memory, and memory for specific details, while the neocortex plays a more prominent role in domain-specific memory (e.g., perirhinal cortex and object-based memory). Specifically, many studies have supported the role of specific subregions of the hippocampus in the computational mechanism of pattern separation, thus potentially allowing us to store and retrieve unique memories for similar experiences of our lives. However, a critical unaddressed question remains: What is the extent to which we make memory-based decisions using the purported pattern-separated representations in the hippocampus vs. other neural mechanisms that employ more distributed representations? We propose that studying human memory performance on tasks with targets and similar lures provides a critical testbed for comparing the competing predictions the role of hippocampal pattern separation vs. more distributed representations in supporting human episodic memory. We generated predictions from competing computational models of performance on memory tests with targets and similar lures and then we tested these predictions in a large sample of human participants ($N = 145$). We found that the comparison between simulated neural responses in an object-processing region of the brain (area IT) and human memory performance exhibited a linear relationship, thus better supporting the predictions of distributed memory models than models of hippocampal pattern separation. Likewise, we observed strong effects of test format on performance as well as clear and consistent relationships between test formats, and these results were also better accounted for by the distributed memory models than the proposed pattern-separated representations of the hippocampus. Altogether, our results provide an important challenge to prominent theories of human memory and provide an important alternative mechanism for explaining human memory performance; therefore, we discuss the implications of our results for reinterpreting previous behavioral, neuroscientific, and computational modeling research and we propose avenues for future research.

2 Introduction

Episodic memory is a core function that supports and enriches our daily lives. The importance of memory can perhaps be best appreciated by observing patients with Alzheimer's disease, who experience a devastating loss of self. Therefore, decades of interdisciplinary research have focused on studying the neural basis of memory using a variety of techniques, including direct measures of the brain (e.g., lesion studies, neurophysiology) and computational modeling. An interesting set of competing ideas emerged from these lines of research, with one group of prominent theories suggesting that the hippocampus and medial temporal lobe cortex play highly specialized roles in episodic memory (e.g., recall vs. familiarity in dual-process models; [1, 2, 3]) and other theories suggesting that we primarily make memory judgments using a combined or unified memory signal (e.g., [4, 5]), which could be implemented by distributed representations (e.g., a rich literature from mathematical psychology; e.g., [6, 7, 8, 9]). More generally, a lively debate emerged in the neurosciences about the degree to which memory and perception rely on local representations

(e.g., "grandmother cells", localization of function; cf. [10, 11, 12]) vs. more distributed representations (e.g., [13]). The Complementary Learning Systems (CLS) model attempts to unite and integrate these competing frameworks by theorizing that the brain evolved separate neural systems to solve the competing goals of avoiding interference, which is supported by local, pattern-separated representations in the hippocampus, while also allowing for generalization, which is supported by distributed representations in the neocortex (note: the CLS framework is still a dual process model; [14, 15, 16, 17]).

Previous studies have modeled the effects of hippocampal damage on memory performance using the CLS framework, which employs separate neural models of the hippocampus and the neocortex (e.g., medial temporal lobe cortex, which is comprised of areas including the entorhinal cortex, perirhinal cortex, and parahippocampal cortex; [16, 17]). The CLS model predicts that the hippocampus uses sparse, pattern-separated representations, which allows it to avoid interference, even between memories for similar experiences. In contrast, based on a rich literature in processes such as semantic memory (e.g., [18, 19]), the CLS model predicts that the neocortex exhibits distributed representations, such that similar experiences cause similar patterns of neural activity, thus allowing organisms to generalize or make predictions about similar experiences. Recognition memory paradigms generally consist of two types of test formats: 1) old/new: a participant sees a single stimulus at a time and decides whether it is old or new (sometimes these formats additionally include a "similar" response), 2) forced-choice: a participant sees two or more images and is asked to select the one that they have seen in the context of the experiment (see Figure 1). The CLS model predicts that hippocampal damage should specifically impair an organism's ability to discriminate between previously viewed stimuli (i.e., targets) and similar lures but they should exhibit typical performance in discriminating between targets and novel foils during the old/new test format. Specifically, the CLS model predicts that the pattern separation mechanisms of the hippocampus allow it to avoid interference between targets and similar lures whereas the distributed representations in the neocortex exhibit interference between targets and similar lures but can support performance for targets and novel foils (i.e., given the low level of similarity between targets and novel foils). Patients with damage to the hippocampus [20, 21, 22, 23], and the dentate gyrus in particular [24], have been shown to exhibit impaired performance on old/new and old/similar/new tests of targets vs. similar lures but intact performance for targets vs. novel foils, thus supporting the predictions of the CLS model.

Several neuroscientific studies have supported the prediction that the hippocampus uses pattern separation to reduce the similarity between similar experiences. Foundational studies discovered that the rodent hippocampus contains neurons that respond when the animal occupies a particular location in the environment, termed "place cells" (e.g., [25, 26]). More recent studies measured the similarity of patterns of activity of hippocampal place cells following changes in the similarity of the spatial environment (e.g., linearly warping from a square to a circular environment; [27]). The theory of hippocampal pattern separation predicts it should exhibit a nonlinear transfer function between the similarity of the inputs (e.g., from high-level cortical areas or the physical/featural similarity) such that it exhibits less similar patterns of activity than its input regions. These studies have generally found that the hippocampus, and the dentate gyrus in particular, exhibits a nonlinear mapping between environmental similarity and place-cell similarity, thus suggesting that the hippocampus is involved in pattern separation (e.g., [27, 28]; for review see [29]).

Studies in humans have further supported the CLS model's prediction that the hippocampus plays a role in pattern separation. For example, in healthy young adults, functional magnetic resonance imaging (fMRI) has consistently shown the involvement of the hippocampus, and the dentate gyrus

and CA3 subregions in particular, in distinguishing between targets (i.e., exact repetitions) and similar lures (i.e., items that are similar to a previously viewed item; e.g., [30, 31, 32, 33, 34, 35]). Moreover, these studies have provided evidence for a nonlinear transfer function between DG/CA3 responses and target-lure similarity (e.g., [31, 32, 36]). Similar to the findings with patients with damage to the hippocampus, healthy aging has been linked to specific disruptions within the hippocampal circuit, including changes in activity in the DG/CA3 subregion of the hippocampus (e.g., hyperactivity: [37, 38, 36]), disruptions in the integrity of the perforant path (i.e., the input from entorhinal cortex to the hippocampus, as measured by diffusion imaging; [39, 40]), and disruptions within the connectivity of a broader hippocampal network (e.g., the integrity of the fornix, a key interface between the hippocampus and neocortex, as measured by diffusion imaging; [41]). Importantly, these hippocampal-network-level disruptions correlated with performance in discriminating between targets and similar lures (on an old/similar/new test format) but did not significantly correlate with performance in discriminating between targets and novel foils, thus further supporting the CLS model's prediction of a single dissociation in performance for these different types of memory judgments. Moreover, Bakker et al. [42, 43] found that giving a drug to decrease neural excitability, levetiracetam, to patients with amnestic mild cognitive impairment reduced the hyperactivity that was seen in the DG/CA3 subregion in a placebo group of patients (for other studies that have shown hyperactivity in patients with amnestic mild cognitive impairment see: [44, 45]), and the treatment group performed significantly better than the placebo group in discriminating between targets and similar lures (on the old/similar/new test format). Altogether, these findings are consistent with pattern separation in the human hippocampus (for reviews see [32, 46]).

A critical unaddressed question remains: What is the degree to which we make memory-based decisions based on the purported pattern-separated representations within the dentate gyrus of the hippocampus vs. the more distributed representations of other hippocampal subregions and the neocortex? While the studies reviewed above suggest that the hippocampus is necessary for typical performance in distinguishing between targets and similar lures and further suggest that the transfer function between targets and similar lures is nonlinear within the dentate gyrus and CA3, other neuroscientific, behavioral, and computational modeling research has suggested that other mechanisms might be used to support memory judgments. For example, in the aforementioned fMRI studies, target-lure similarity was generated via the mnemonic similarity (i.e., memory performance=the proportion of times participants incorrectly responded "old" to similar lures). Therefore, the nonlinear mapping between mnemonic similarity and DG/CA3 activity actually suggests a dissociation between memory performance and neural signals. In contrast, activity in CA1 tended to track linearly with mnemonic similarity (e.g., [31, 36]; for review see: [32]), which we think suggests that CA1 exhibits a stronger relationship with behavior. Similarly, immediate early gene studies and single-cell electrophysiology have revealed that the transfer function between the similarity of patterns of active neurons in CA1 and environmental similarity appears to be linear (for review see: [47]). In fact, evidence from computational models (e.g., [48, 49]) and fMRI (e.g., [50, 51, 52]) has suggested that the hippocampus, and the CA1 subregion in particular, plays a critical role in novelty detection, perhaps through a match-mismatch mechanism. For example, Duncan et al. [52] used fMRI to show that the human CA1 responded in a linear manner to the degree of changes in an environment. Therefore, these results suggest that the hippocampus might calculate the match vs. mismatch between a past and current experience. Importantly, these kinds of signals would be better instantiated in distributed, overlapping patterns of activity rather than via pattern separation. We will next turn our discussion to distributed memory models and how they have been applied to account for performance on recognition memory tasks.

Different computational models make competing predictions about the nature of the relationship between the similarity of experiences and the resultant memory judgments. The CLS model makes a strong prediction that the hippocampus is generally not useful for computing the global similarity between a probe item and the contents of memory, whereas the neocortex would play a role in calculating the global match [17], which is generally consistent with dual-process models (e.g., [1, 2, 3]). Specifically, pattern separation results in a low level of between-item similarity of patterns of activity in the hippocampus if we use lists of items with little overlap between the various items in the stimulus set (i.e., aside from the target-lure similarity, the similarity among the original targets within the list is low), thus the hippocampus largely makes memory judgments based on the similarity between a probe item and one specific target item (i.e., a selective match rather than a global match). However, a class of competing computational models, which we refer to as distributed memory models, suggest that we primarily rely on distributed representations and the global match between a probe item and the contents of our memory to make memory judgments (i.e., similar to the CLS neocortical model; e.g., [6, 7, 8]). For example, Davis et al. [9] combined computational modeling and fMRI and they found that recognition memory judgments were correlated with the global similarity between a probe item and the responses to the original target items during encoding within several regions of the medial temporal lobe, including the medial temporal lobe cortex and the hippocampus. Combined with the match-mismatch findings above, these results challenge the notion that the hippocampus cannot compute the global match (but see: [53]) and for the hypothesis that pattern separation in the dentate gyrus would result in a direct behavioral readout on memory tasks. Moreover, given the strong relationship between global match and performance [9], these results provide more direct evidence that memory judgments might be supported by a global matching mechanism (note: they did not specifically examine similar lures, which is a primary aim of our work here).

Performance on forced-choice recognition memory tasks has often been found to be better supported by predictions of distributed memory models than by the pattern separation mechanisms of the dentate gyrus. Specifically, there are three basic types of forced-choice test formats: A-X (a target and an unrelated, novel foil; traditional forced-choice recognition), A-A' (a target and a corresponding similar lure; e.g., a picture of an apple and a similar apple), and A-B' (a target and a noncorresponding similar lure; e.g., a picture of an apple and a basketball that is similar to one that was studied in the encoding phase; note, we use the nomenclature of Tulving [54]; see Figure 1). The CLS framework predicts that damage to the hippocampus will cause a test-format-specific impairment in discriminating between targets and similar lures within forced-choice tests: patients should be impaired on the A-B' test format but they should exhibit typical performance on the A-A' test format as well as the A-X test format (see simulation 4 from [16, 17]). In contrast to the findings for the old/new test formats, the support for the predictions of selective impairments on the forced-choice formats has been equivocal. A case study of a single patient suggested that damage to the hippocampus specifically impaired performance in discriminating between targets and similar lures in the old/new test format with intact performance on the A-A' test format [20] (for similar findings in patients with amnestic mild cognitive impairment see: [55]). However, other studies with a larger sample of patients suggest that hippocampal damage impairs performance on both the old/new and the A-A' test formats [21, 22]. Moreover, Jeneson et al. [22] found that patients with selective hippocampal damage exhibited a similar degree of impairment on the A-A' test format as the A-B' test format. Specifically, they did not observe a significant patient-group by test-format interaction and they observed that the patients were significantly impaired on both test formats (as well as the old/new test format). Therefore, these latter results provide evidence that is discordant with the prediction of the standard version of the CLS hippocampal model, and we turn

next to discussing performance on the forced-choice test formats in healthy young adults.

The support for the CLS model's prediction of similar performance on the A-A' and A-B' test formats in healthy young adults (i.e., given their intact hippocampal circuit; see see Figure 2a; cf. [56]) has also been equivocal (but note that [16, 17] discussed the idea that human memory would likely better be modeled by combining the output of the hippococampal and medial temporal lobe cortex model, which we will return to in the Discussion). Migo et al. [56] found that healthy young adults exhibit similar performance on the A-A' and the A-B' forced-choice test formats (i.e., the Recognition condition from Experiment 1 in [56]). However, other, larger studies have shown clear effects of test format (e.g., Experiment 1 and 2 from [57]; [8, 54, 58, 59]). Specifically, the general finding in these studies is that performance on the A-A' test format is significantly better than performance on the A-B' test format, and this finding is consistent within healthy young adults (Experiment 1 and 2 from [57]; [8, 54, 58, 59]), healthy middle-aged adults [22], healthy older adults (Experiment 1 from [57]), and patients with damage to the hippocampus [22]. Altogether, the evidence for the hippocampal model's prediction of a null effect of a difference between the A-A' and the A-B' forced-choice test format (cf. [56]) is tenuous. We suggest that more evidence points toward a significant effect of test format because several published studies have shown a significant A-A' > A-B' test format effect and it is difficult to interpret null effects in the other papers (especially without a Bayesian analysis). Importantly, distributed memory models readily account for the A-A' > A-B' test format effect (e.g., [8, 57]; also see the neocortical model from Simulation 4 in [16]).

We aimed to test contradictory predictions from distributed memory models and the CLS hippocampal model about performance on memory tasks with targets, similar lures, and novel foils. First, we tested for the A-X > A-A' > A-B' test format effect, which should be significant if the distributed memory models are supported but null if the CLS hippocampal model is supported. We further aimed to test the relationship between performance on the forced-choice and old/new test formats to test specific predictions of distributed memory models. Second, distributed memory models and the CLS hippocampal model also make contradictory predictions about the relationship between memory performance and the similarity between targets and similar lures. Specifically, as we show here, distributed memory models predict that the degree of similarity will cause a relatively strong, continuous, and monotonic relationship between performance in discriminating between targets and similar lures. In contrast, the CLS hippocampal model predicts a much more modest effect of target/lure similarity, which exhibits a relatively flat or null effect of stimulus similarity until the target/lure similarity is extremely high, given its use of pattern separation in the dentate gyrus and CA3 (see Figure 2b and [16]). We aimed to quantify a measure of physical or semantic similarity between our target and lure items using a biologically inspired artificial neural network to simulate patterns of activity in the inferior temporal cortex (area IT; [60]), which is a high-level visual region involved in object recognition and it is heavily interconnected with the perirhinal cortex [61]. Thus, we argue that the artificial neural network allows us to empirically determine the transfer function of performance on the task vs. an objective measure of the target/lure similarity. We examined the effects of target/lure similarity on performance across various test formats, including the old/new, A-A', and A-B' test formats. Finally, we aimed to further generate and test predictions from distributed memory models and neural networks to test how they compare with human performance. To foreshadow our findings and conclusions, we provide evidence that performance across several test formats that ask participants to distinguish between targets, similar lures, and novel foils is more consistent with the distributed memory models (vs. hippocampal pattern separation), thus providing important insight into neural and computational theories of human memory and the behavioral tasks that are used to study these systems.

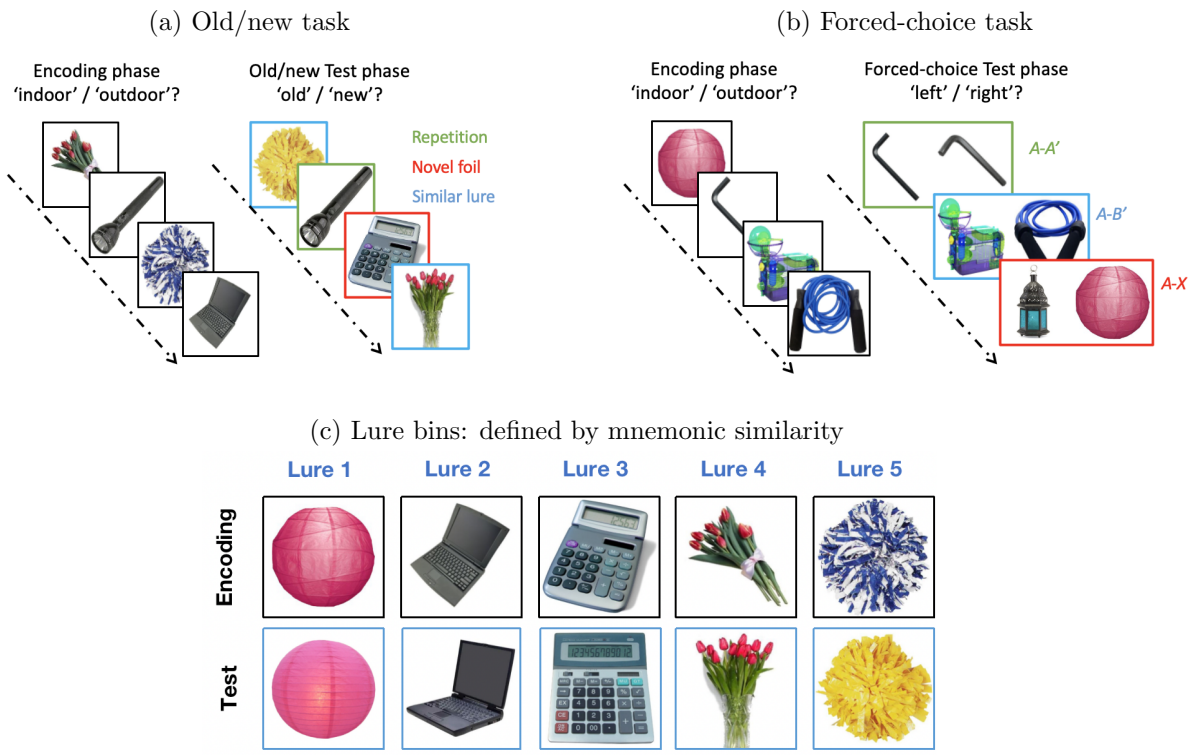


Figure 1: We used two versions of the task and each version contained similar lures that varied as a function of mnemonic similarity. (a) The old/new task was similar the standard versions of the mnemonic similarity task: participants indicated whether each item was "old" or "new". Participants were instructed to respond "old" only to repetitions and "new" to similar lures and novel foils. (b) We also used a forced-choice version of the mnemonic similarity task with three test formats: A-X (a target and a novel foil), A-A' (a target and a corresponding similar lure), and A-B' (a target and a noncorresponding similar lure). (c) In both the old/new and forced-choice tasks, the similar lures varied as a function of mnemonic similarity, which was previously defined as the proportion of times that participants respond "old" to the similar lure [46, 62, 63]. The lure bins varied from the most similar, Lure bin 1, to the most dissimilar, Lure bin 5.

3 Results

3.1 Quantification of target/lure similarity with an artificial neural network

We first aimed to test the nature of the transfer function between performance and target-lure similarity. Specifically, as we discussed in the Introduction, the CLS version of the hippocampal model suggests that the transfer function between performance and the target-lure similarity is highly nonlinear [16, 17]. Specifically, this theory suggests that the hippocampus will largely avoid interference until the target-lure similarity is very high (i.e., above 87.5% overlap; see Figure 2b). In contrast, the distributed memory models predict that the target-lure similarity will more strongly and continuously affect performance (see below; also see the cortical model from [16]). Therefore, we used an artificial neural network to simulate responses in area IT to all of the images in 6 stimulus sets of the Mnemonic Similarity Task [46, 62, 63], which each contained 192 target-lure pairs (i.e., a total of 1152 images).

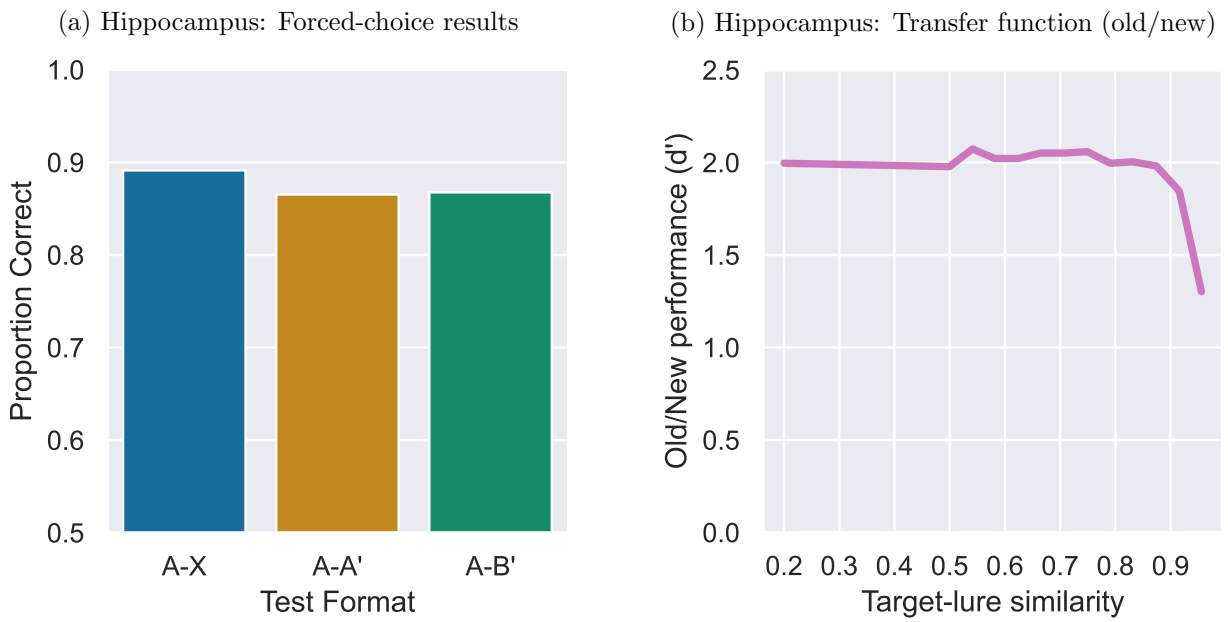


Figure 2: (a) Norman and O'Reilly [16] found that the simulated behavioral performance of the CLS hippocampal model was not significantly different on the A-A' vs. A-B' forced-choice test formats (replotted from the 0.71 overlap condition in Figure 12 of Norman and O'Reilly [16]), and there were cases in which performance on the A-B' test format was actually better than the A-A' test format. (b) Norman and O'Reilly [16] found that the simulated behavioral performance of the CLS hippocampal model on the old/new test format was highly resilient to target-lure similarity, such that there was virtually no difference in performance for discriminating between targets vs. novel foils (target-lure similarity of 0.2) or targets vs. similar lures with a moderate to high level of overlap (i.e., similarity from 0.5 to 0.875). In fact, the hippocampal model did not exhibit interference between targets vs. similar lures until extremely high levels of similarity (i.e., similarity > 0.875; replotted from Figure 9 of Norman and O'Reilly [16]).

As we discuss in the Methods, we extracted patterns of activity from the 3rd convolutional layer of IT from CORnet-S (i.e., the penultimate layer; [60, 64]) and calculated the pattern similarity (Pearson correlation) between the responses to the target and the lure images. We compared these pattern-similarity values to behavioral performance, which we defined here based on the proportion of times that participants made an "old" response to the lures (1-p('old')), as in past research [46, 62, 63]. We observed a strong linear relationship between performance on the task and the simulated pattern similarity (Pearson's $r = -0.41$, $p = 1.71 \times 10^{-48}$; see Figure 3). Note that the magnitude of our observed effect compares favorably to other approaches linking neural network models with human behavior on perceptual tasks [60, 64]. Moreover, we averaged the pattern similarity and performance within each lure bin to reduce noise (i.e., because the simulations are not a perfect recapitulation of actual neural responses), and we again observed a strong linear relationship between these variables (Pearson's $r = -0.99$). Therefore, these results suggest that the transfer function between mnemonic discrimination performance and target-lure similarity exhibits a strong, continuous, and perhaps linear relationship. Moreover, while it is difficult to interpret the pattern similarity values directly, the magnitude of these values span a broad range and performance is not solely impaired by only the most similar lures (e.g., pattern similarity greater than 0.75). Altogether, these results support the predictions of the distributed

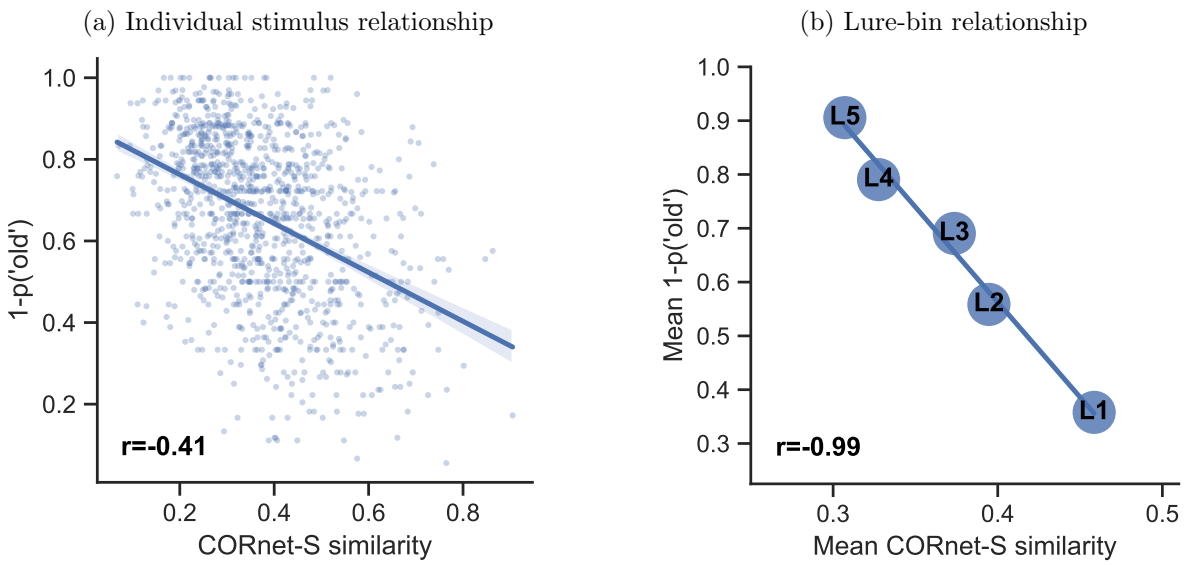


Figure 3: We observed a significant correlation between the simulated target-lure pattern similarity in the IT of CORnet-S and mnemonic discrimination performance in the data that were used to define the original lure bins [46, 62, 63]. (a) We observed a significant correlation between CORnet-S similarity and 1-p('old') for individual stimulus pairs. Here, the correlation between the model and behavior compares favorably to other studies that have compared pattern similarity in artificial neural networks and behavioral performance (e.g., on perceptual tasks). (b) We also averaged the pattern similarity and the memory performance values (1-p('old')) within each lure bin to reduce noise, and we again found a striking relationship between pattern similarity and performance on the task. The points indicate the lure bin of each data point (see Figure 1). Altogether, these data support the predictions of the distributed memory models (see Figure 4b, Figure 5b, Figure 6b, Figure 7b) but are at odds with the predictions of the CLS hippocampal model (see Figure 2b).

memory models (see below) and contradict the predictions from the hippocampal model [16, 17] (see Figure 2b).

3.2 Comparison of predictions from the computational models

We next aimed to generate predictions from several computational models regarding performance on memory tasks with targets, similar lures, and novel foils (e.g., the Mnemonic Similarity Task; [65, 62, 63, 46]; see Figure 1). Here, we specifically compared the predictions from the CLS model of the hippocampus [16, 17] with several models, including MINERVA 2 (a multiple-trace global-matching model; [66, 8]), theory of distributed and associative memory (TODAM; a unitrace global-matching model; [6, 7]), and two implementations of a Hopfield/autoassociative neural network (similar computational architecture to the CA3 subregion of the hippocampus: [67, 68, 69]). Given that all of these models implement distributed representations, we will refer to these collectively as "distributed memory models" (which also stands in contrast to the CLS hippocampal model, which emphasizes sparse, pattern-separated representations). We then tested the predictions from these models with empirical behavioral performance in a large sample of human participants.

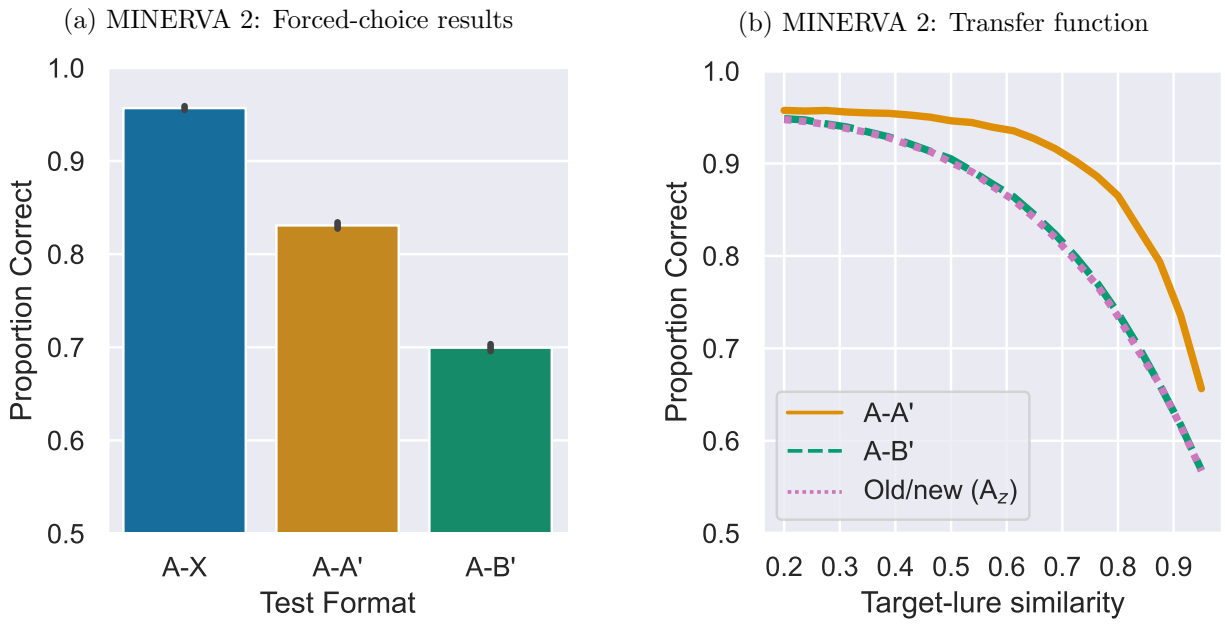


Figure 4: (a) MINERVA 2 predicts a clear test-format effect: A-X > A-A' > A-B' (note the difference from the predictions of the CLS hippocampal model). (b) MINERVA 2 predicts a clear, monotonic relationship between target-lure similarity and performance on the A-A', A-B', and old/new (A_z) test formats (note the difference from the predictions of the CLS hippocampal model, which predicts that performance on the old/new test format will only exhibit interference at very high levels of target-lure similarity). Each of these simulations included 1,000 simulated participants. Error bars indicate the 95% confidence intervals. Please compare these results to the empirical results in Figure 3, Figure 11, and Figure 12.

3.2.1 Predictions of forced-choice performance

As we discussed in the Introduction, the CLS model predicts that the hippocampus will exhibit similar performance on the A-A' and the A-B' forced-choice test formats (see Figure 2a and [16]). In fact, Norman and O'Reilly [16] reported that there are even times in which the hippocampus performs better on the A-B' than the A-A' test format because the hippocampus employs recall-to-reject in their model, thus the A-B' test format gives the hippocampus two chances to recall-to-reject the stimulus, thus any benefit of covariance in the A-A' is offset (or even overcompensated for) by the additional opportunity to recall-to-reject a given stimulus. In contrast, the global matching models, MINERVA 2 (see Figure 4a) and TODAM (see Figure 5a), predict a strong A-A' > A-B' test format effect [8, 57]. Moreover, we found that both the "familiarity" (see Figure 6a) and the "recall" version (see Figure 7a) of the Hopfield/autoassociative neural network predicted a similar A-A' > A-B' test format effect. Therefore, in summary, MINERVA 2, TODAM, and both versions of the Hopfield/autoassociative neural network suggest that there will be a strong A-X > A-A' > A-B' test format effect (also see the neocortical model from Simulation 4 in [16]), which differs from the CLS hippocampal model, which predicts a null A-A' > A-B' test format effect (see [16, 56]). As we report below, we found a clear effect of test format, thus supporting the predictions of the distributed memory models and contradicting the predictions of the CLS hippocampal model.

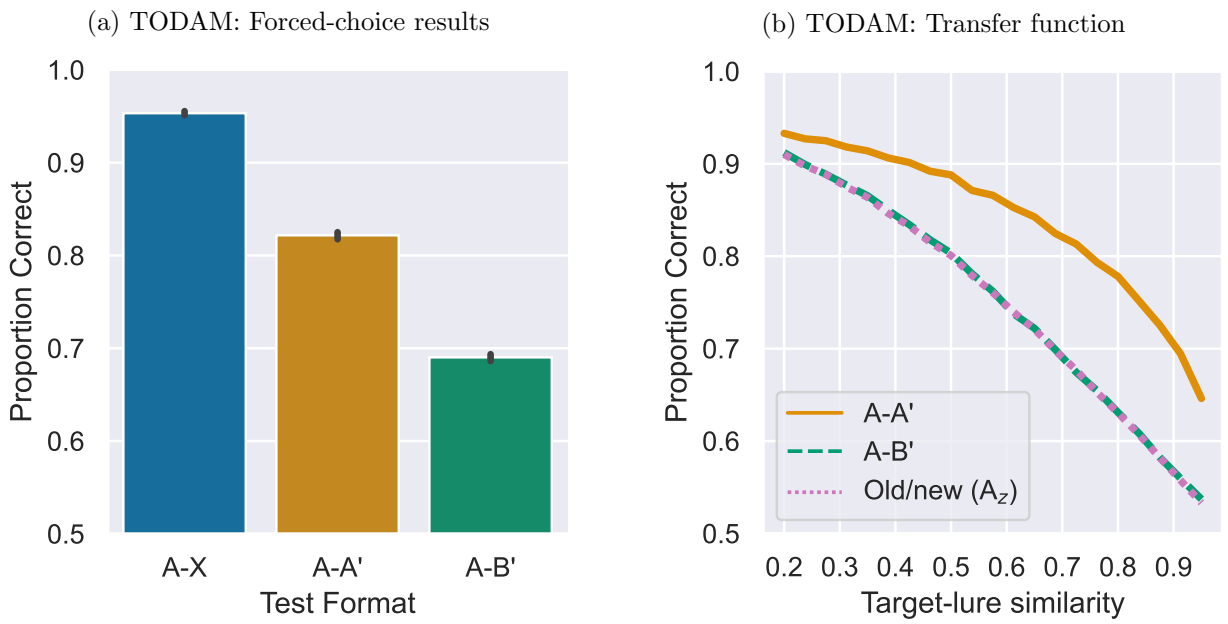


Figure 5: (a) TODAM predicts a clear test-format effect: A-X > A-A' > A-B' (note the difference from the predictions of the CLS hippocampal model). (b) TODAM predicts a clear, monotonic relationship between target-lure similarity and performance on the A-A', A-B', and old/new (A_z) test formats (note the difference from the predictions of the CLS hippocampal model, which predicts that performance on the old/new test format will only exhibit interference at very high levels of target-lure similarity). Each of these simulations included 1,000 simulated participants. Error bars indicate the 95% confidence intervals. Please compare these results to the empirical results in Figure 3, Figure 11, and Figure 12.

3.2.2 Predictions of the transfer function between performance and target-lure similarity

As we discussed in the Introduction, the CLS model predicts that the hippocampus would not exhibit a clear and continuous transfer function between simulated performance and target-lure similarity. Specifically, the transfer function between performance and target-lure similarity is flat or null until only very high levels of similarity (i.e., > 0.875; see Figure 2b and [16]). In contrast, the CLS model of the neocortex exhibits a much more continuous effect of target-lure similarity on simulated performance [16]. In general, all of the distributed memory models predict a relatively continuous transfer function between memory performance and target-lure similarity (MINERVA 2: Figure 4b; TODAM: Figure 5b, the familiarity-based Hopfield autoassociative neural network: Figure 6b; the recall-based Hopfield autoassociative neural network: Figure 7b). Therefore, we aimed to test these competing predictions to determine which candidate models make predictions that are closer to human behavioral performance. As we reported above (see 3.1), we found a continuous (and possibly linear) mapping between performance and simulated target-lure similarity, thus supporting the distributed memory models.

3.2.3 Predictions of the relationship between forced-choice and old/new performance

We next aimed to generate predictions about the relationship between performance on various test formats. Norman and O'Reilly [16] suggested that there are many cases in which hippocampal and

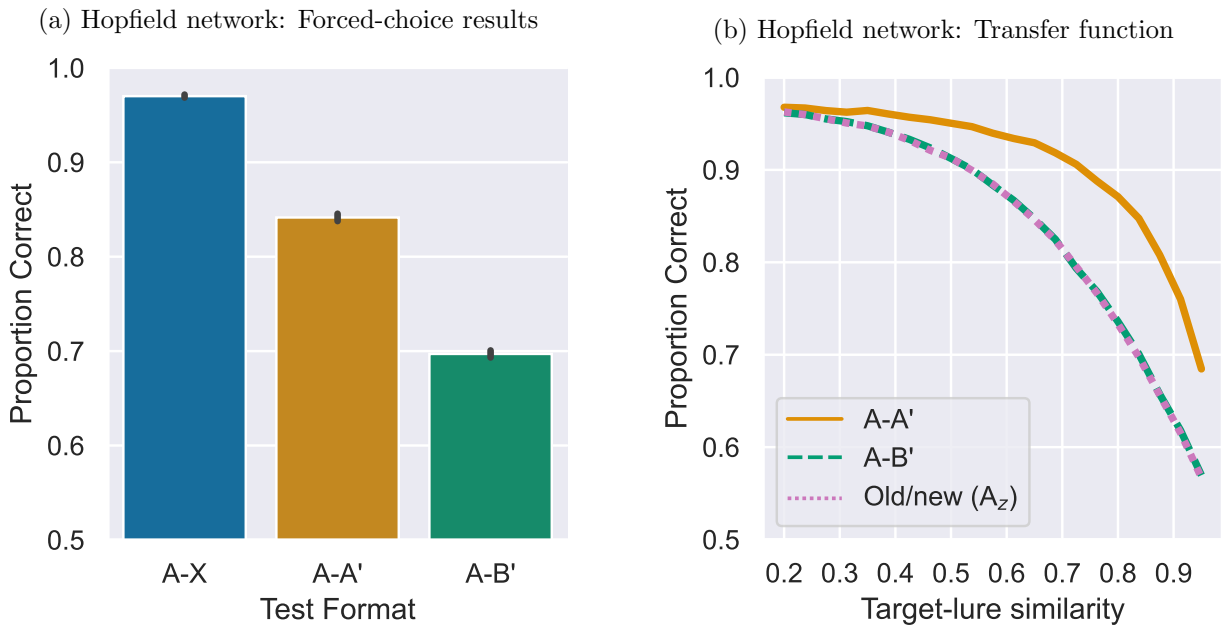
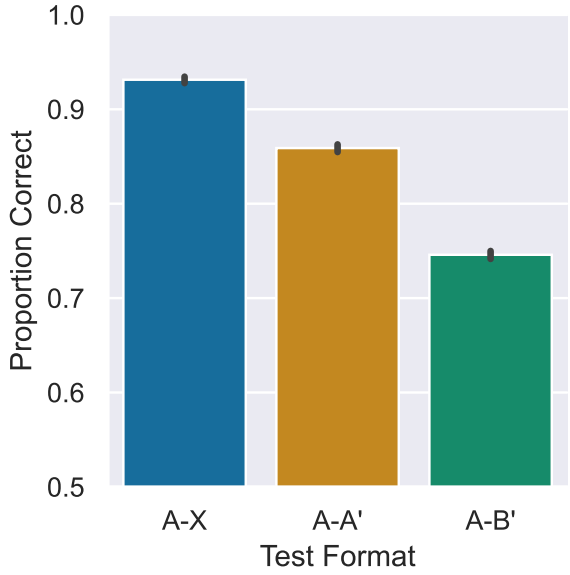


Figure 6: (a) The familiarity-based version of the Hopfield autoassociative neural network predicts a clear test-format effect: A-X > A-A' > A-B' (note the difference from the predictions of the CLS hippocampal model). (b) The familiarity-based version of the Hopfield autoassociative neural network predicts a clear, monotonic relationship between target-lure similarity and performance on the A-A', A-B', and old/new (A_z) test formats (note the difference from the predictions of the CLS hippocampal model, which predicts that performance on the old/new test format will only exhibit interference at very high levels of target-lure similarity). Each of these simulations included 1,000 simulated participants. Error bars indicate the 95% confidence intervals. Please compare these results to the empirical results in Figure 3, Figure 11, and Figure 12.

neocortical performance will not be correlated, which means that performance across test formats will not always be correlated (e.g., A-A' performance might not correlate with A-B' performance). Moreover, as we mentioned above, the CLS hippocampal model predicts that performance will generally be the same across the A-A', A-B', and the old/new with targets and similar lures [16]. In contrast, the distributed memory models make clear predictions about the relationship between performance on the various test formats. For example, we modeled individual differences in performance by changing the learning rate in the models. We found that the models predicted clear differences in performance as a function of learning rate (see Figure 8; note: this figure depicts MINERVA 2, but results were similar in the other distributed memory models). Moreover, when we investigated the relationship between performance on the forced-choice vs. old/new test formats, we found that the distributed memory models predicted that performance on the A-X was roughly equivalent to the old/new test format with targets and novel foils (i.e., the regression between performance on the two formats produced a slope of approximately 1; see Figure 9a), performance on the A-A' test format is predicted to be better than performance on the old/new test format with targets and similar lures (i.e., the regression between performance on the two formats produced a slope that is greater than 1; see Figure 9b), and performance on the A-B' test format was roughly equivalent to performance on the old/new test format with targets and similar lures (i.e., the regression between performance on the two formats produced a slope of approximately

(a) Hopfield network (recall): Forced-choice results



(b) Hopfield network (recall): Transfer function

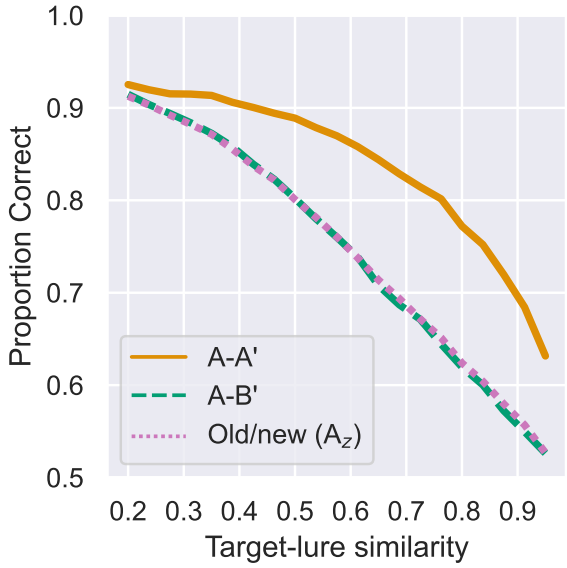


Figure 7: (a) The recall-based version of the Hopfield autoassociative neural network predicts a clear test-format effect: A-X > A-A' > A-B' (note the difference from the predictions of the CLS hippocampal model). (b) The recall-based version of the Hopfield autoassociative neural network predicts a clear, monotonic relationship between target-lure similarity and performance on the A-A', A-B', and old/new (A_z) test formats (note the difference from the predictions of the CLS hippocampal model, which predicts that performance on the old/new test format will only exhibit interference at very high levels of target-lure similarity). Each of these simulations included 1,000 simulated participants. Error bars indicate the 95% confidence intervals. Please compare these results to the empirical results in Figure 3, Figure 11, and Figure 12.

1; see Figure 9c; also see below, including Figure 14). Additionally, we generated predictions of the similarity of performance on all 5 test formats using a random-effects-style simulation and we found that the distributed memory models suggest that performance will be significantly correlated between all of the test formats (see Figure 10). Thus, while there are conditions in which the CLS model predicts correlated performance between tasks [16], if we observe 1) roughly equivalent performance between the A-X and old/new test format with targets and novel foils, 2) better performance on the A-A' test format than both the A-B' test format and old/new test format with targets and similar lures, 3) roughly equivalent performance between the A-B' and the old/new test format with targets and similar lures, 4) strong correlations between test formats, it will provide stronger support for the distributed memory models than the CLS hippocampal model.

3.3 Empirical results

We designed two versions of the Mnemonic Similarity Task [62, 63, 46]: the old/new MST and the forced-choice MST (see Figure 1). In both tasks, participants first completed an incidental encoding phase in which they made indoor/outdoor judgments for each image. During the test phase of the old/new MST, participants saw repeated items (targets), similar items (similar lures), and novel items (novel foils), and we asked them to respond "old" for items that were exactly the same as one they saw during the indoor/outdoor task and to respond "new" to items that were

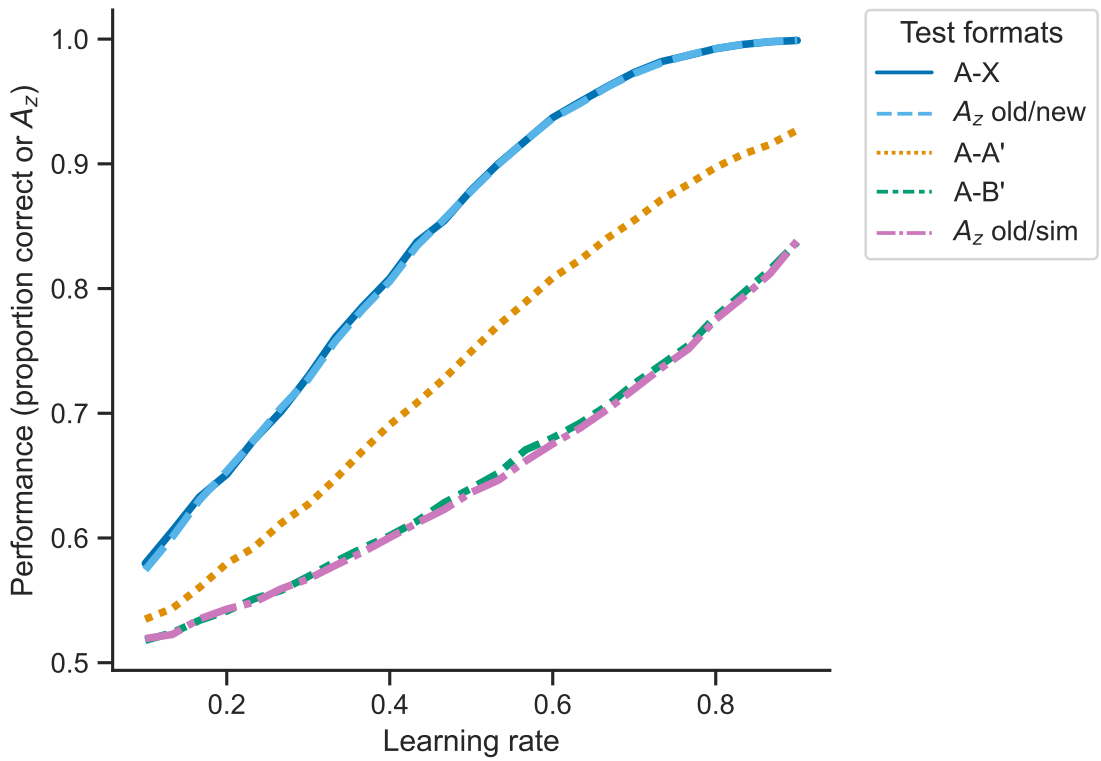


Figure 8: Simulations of individual differences in learning rates reveal performance tuning curves across the 5 test formats. The distributed memory models predict that performance on the A-X and A_z for targets vs. novel foils will be roughly equivalent as will performance on the A-B' and the A_z for targets vs. similar lures, while performance on the A-A' test format will be intermediate. Here, we show the results of MINERVA 2, but the other distributed memory models made very similar predictions.

similar to or totally new from any images they saw during the indoor/outdoor task. We used a within-participants design, in which each participant completed both versions of the MST in a randomized order (N=145). We report the results of each version as well as the comparisons between the tasks below.

3.3.1 Forced-choice data

We first investigated whether there was an effect of test format on performance of the forced-choice version of MST (see Figure 1b) using a generalized linear mixed model with a binomial family. Specifically, we first compared a null model (random slopes and intercepts) with a test-format model that incorporated the mean performance from Experiment 2 in [57] (A-X: logit(0.959) A-A': logit(0.836), A-B': logit(0.756)). A likelihood ratio test revealed that the test-format model accounted for significantly more variance than the intercept-only model ($\chi^2(2, N = 145) = 193.69$, $p < 2.2 \times 10^{-16}$; intercept-only model $BIC = 2,284.2$, test-format model $BIC = 2,096.6$; see Figure 11a). The fixed-effects beta estimates for the test-format model were an intercept of -0.186 and a slope of 1.09. Thus, these results replicate our previous findings and suggest that the magnitude of our observed effects here are similar (i.e., our intercept is close to zero and our slope is close to 1). We next compared our test-format model with the predictions of the CLS hippocampal model

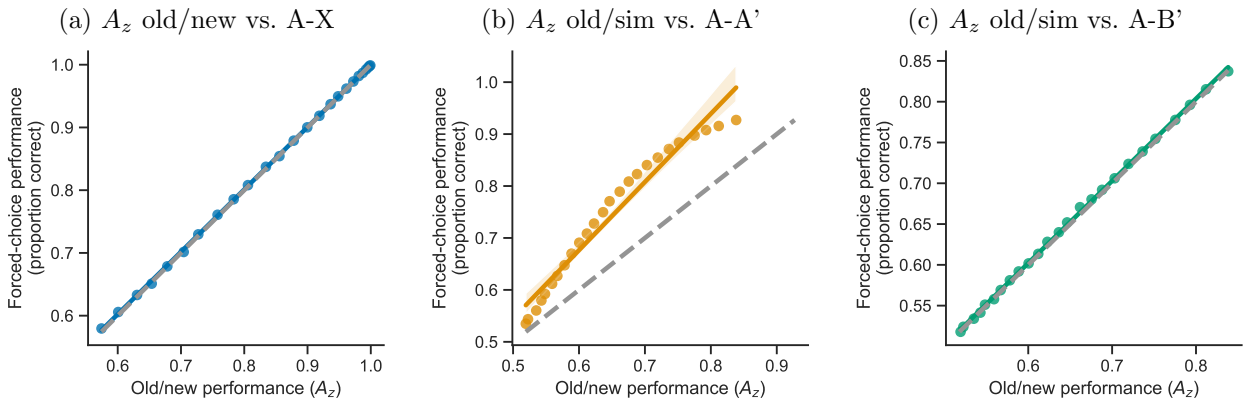


Figure 9: The distributed memory models make clear predictions about the relationship of performance on the forced-choice and old/new test formats. (a) The models predict that performance on the A-X and the old/new test format with targets and novel foils will be approximately equivalent (i.e., a slope of approximately 1). (b) The models predict that performance on the A-A' test format will be significantly better but also strongly related to performance on the old/new test format with targets and similar lures (i.e., a slope that is greater than 1). (c) The models predict that performance on the A-B' test format will be approximately equivalent to performance on the old/new test format with targets and similar lures (i.e., a slope of approximately 1). Please compare these results to Figure 13, Figure 14, and Figure 15.

of equal performance on the A-A' and A-B' test (A-A' = A-B': mean(logit((0.836, 0.756))); see Figure 2a; cf. [16, 56]). We compared the Bayesian information criterion (BIC; lower values indicate model preference) and found evidence that was very much in favor of the distributed memory models vs. the CLS hippocampal model (BIC test-format-model: 2,096.6; BIC CLS hippocampal model: 2,222.1; $BIC_{distributed} - BIC_{hippocampus} = -125.5$; see Methods). Moreover, we used the BIC approximation for calculating Bayes factors [70, 71] (see Equation (14)) and we found that the evidence was in favor of the distributed model vs. the hippocampal model ($BF_{10} \approx 1.74 \times 10^{27}$; where 1 indicated distributed memory models and 0 indicates the CLS hippocampal model), thus suggesting that the data are 1.74×10^{27} times more likely for the predictions of the distributed memory models relative to the predictions of the CLS hippocampal model.

We next performed planned follow-up tests to determine whether performance on the A-X format was better than the A-A' format and whether performance on the A-A' format was better than the A-B' format (i.e., as in [57]). First, we refit the null model and the test-format model while only including data from the A-X and A-A' test formats. We again found that the test-format model accounted for significantly more variance than the intercept-only model ($\chi^2(1, N = 145) = 173.15$, $p < 2.2 \times 10^{-16}$; null model $BIC = 1,516.3$, test-format model $BIC = 1,348.8$; intercept = 0.052, slope = 0.98), thus indicating that performance on the A-X test format is significantly better than performance on the A-A' test format. Second, we refit the intercept-only model and the test-format model while only including data from the A-A' and A-B' test format. We again found that the test-format model accounted for significantly more variance than the intercept-only model ($\chi^2(1, N = 145) = 97.49$, $p < 2.2 \times 10^{-16}$; null model $BIC = 1,660.2$, test-format model $BIC = 1,568.4$; intercept = -0.45, slope = 1.29), thus indicating that performance on the A-A' test format is significantly better than performance on the A-B' test format. Altogether, these results replicate our previous findings [57], and strongly suggest that there is an effect of test format, thus

Figure 10: The distributed memory models predict that individual differences in learning rates will lead to correlated performance between all 5 test formats. The correlation matrix depicts simulated performance in MINERVA 2, but all of the models made similar predictions. Please compare these results to the empirical results in Figure 16.

further bolstering the notion that participants perform better on the A-A' test format than the A-B' test format. Therefore, these results are consistent with the predictions of the distributed memory models (MINERVA 2: Figure 4b, TODAM: Figure 5b, both versions of the Hopfield autoassociative neural network: Figure 6b and Figure 7b; also see the neocortical model from Simulation 4 in [16]) but not with the CLS hippocampal model (see Figure 2a; [16]).

We next investigated the effect of target-lure similarity, determined by lure bin levels, on the performance (see Figure 1c). We focus here on the results of A-A' and A-B' test formats because the A-X test format trials did not manipulate the similarity between the encoding objects and test objects. For this analysis, we determined target-lure similarity using the lure bin levels, which were generated from the empirical mnemonic similarity from previous experiments (i.e., defined by the p("old") to each stimulus, which is a measure of confusability; the lure bins were generated from the old/similar/new test format: [31, 46, 62, 63]). We compared the test-format model with a test-and-lure-bin model and found that the test-format-and-lure-bin model accounted for significantly more variance than the test-format model ($\chi^2(1, N = 145) = 135.24$, $p < 2.2 \times 10^{-16}$; test-format model $BIC = 4,410$, test-format-and-lure-bin model $BIC = 4,282$), thus demonstrating that participants' performance varied as a function of the lure bins (i.e., target-lure similarity; see Figure 11b). We next tested whether the effect of the lure bin varied as a function of test format (e.g., was more pronounced in one test format). Interestingly, a test-format by lure-bin interaction model did not account for significantly more variance than the test-and-lure-bin model ($\chi^2(1, N = 145) = 0.13$, $p = 0.72$; test-format by lure-bin interaction model $BIC = 4,289.2$, note that the BIC for this model is higher than the test-format-and-lure-bin model), thus suggesting

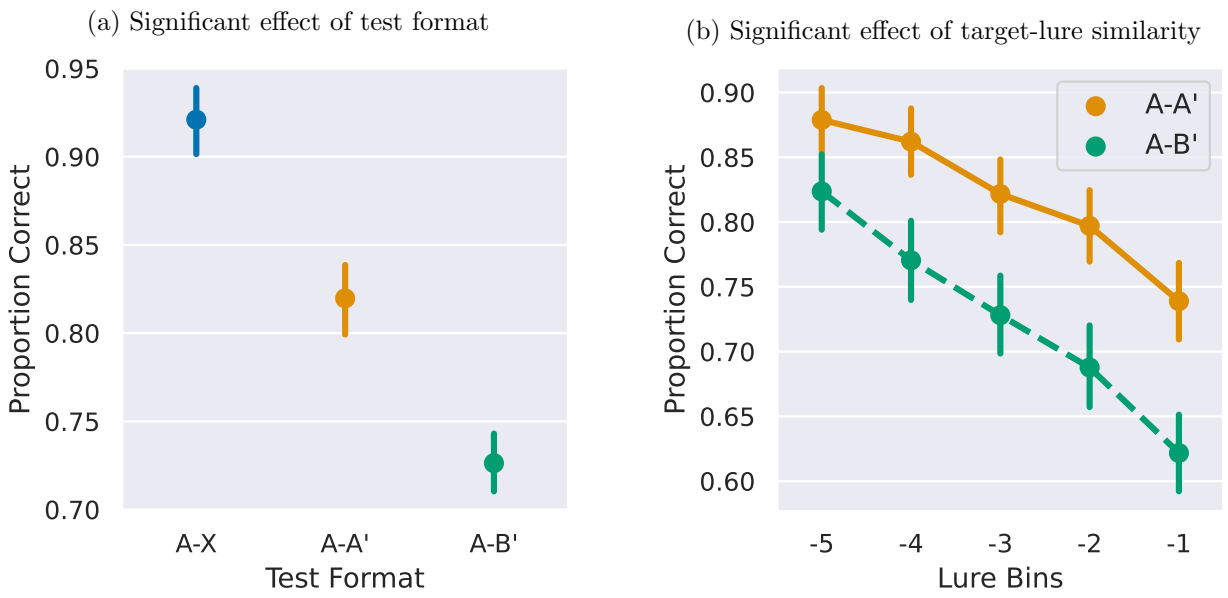


Figure 11: Performance on the forced-choice test formats differed as a function of test format and as a function of target-lure similarity. (a) We observed a significant effect of test format: participants performed best on the A-X format followed by the A-A' format followed by the A-B' format, which is consistent with the distributed memory models (but not the CLS hippocampal model). (b) We observed a significant effect of lure bin (i.e., mnemonic similarity from previous old/similar/new experiments) on performance of the A-A' and A-B' test formats. Moreover, the lure-bin effect was remarkably similar between conditions, although participants performed significantly better on the A-A' than the A-B' format. Note: for visualization, the data are plotted here as the proportion correct across all trials; however, we analyzed our data using generalized linear models with a binomial family. Additionally, for visualization, we multiplied the lure bins by -1 so that the least similar bins are to the left of the plot and the most similar bins are to the right of the plot, thus enabling an easier comparison with the predictions of the computational models. Altogether, these results support the predictions of the distributed memory models (see Figure 4, Figure 5, Figure 6, and Figure 7) but are at odds with the predictions of the CLS hippocampal model (see Figure 2). Points represent the mean and the error bars indicate the 95% confidence intervals.

that the lure bin effect was similar between the A-A' and A-B' test formats ($BF_{01} \approx 36.6$). To further explore the similarity of the lure bin effect on the A-A' and A-B' test formats, we next compared a lure bin model with an intercept-only model within each test format separately. We found that the lure-bin model accounted for significantly more variance than the intercept-only model for both the A-A' test format ($\chi^2(1, N = 145) = 88.71$, $p < 2.2 \times 10^{-16}$; intercept $= 1.70$, lure bin beta $= 0.26$; intercept-only model $BIC = 2,080.7$, lure-bin model $BIC = 1,998.5$) and the A-B' test format ($\chi^2(1, N = 145) = 126.71$, $p < 2.2 \times 10^{-16}$; intercept $= 1.05$, lure bin beta $= 0.26$; intercept-only $BIC = 2,434.9$, lure-bin model $BIC = 2,310.2$). Note that the beta estimates for the lure bin effect were remarkably similar across these two test formats: 0.26 for both the A-A' and the A-B' test formats. Taken together, these results clearly indicate that there is an effect of the lure bin on performance of both the A-A' and the A-B' test format and further suggest that this effect is similar across these two formats.

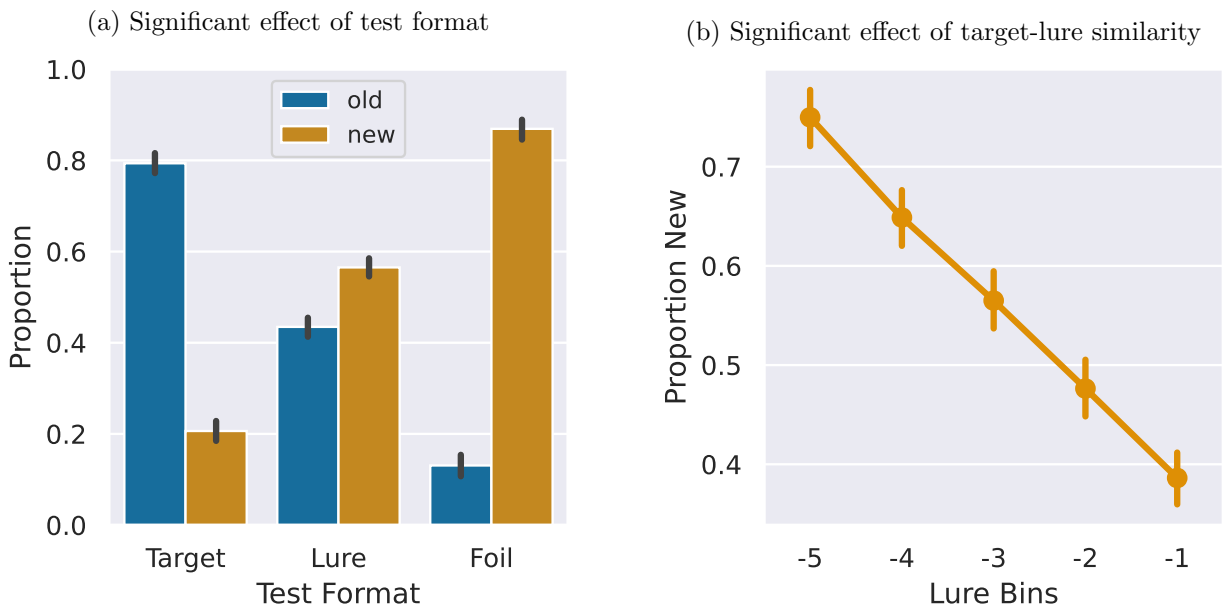


Figure 12: Performance on the old/new format differed as a function of test format and as a function of target-lure similarity. (a) We observed a significant effect of test format: participants made a different number of "old" and "new" responses across the targets, similar lures, and novel foils. (b) We observed a significant effect of lure bin (i.e., mnemonic similarity from previous old/similar/new experiments) on the number of times participants responded "old" to similar lures. Note: for visualization, the data are plotted here as the proportion correct across all trials; however, we analyzed our data using generalized linear models with a binomial family. Additionally, for visualization, we multiplied the lure bins by -1 so that the least similar bins are to the left of the plot and the most similar bins are to the right of the plot, thus enabling an easier comparison with the predictions of the computational models. Altogether, these results support the predictions of the distributed memory models (see Figure 4, Figure 5, Figure 6, and Figure 7) but are at odds with the predictions of the CLS hippocampal model (see Figure 2). Bars and points represent the mean and the error bars indicate the 95% confidence intervals.

3.3.2 Old/new data

We next investigated whether there was an effect of test format on performance on the old/new version of the MST (for the raw proportion of responses on each trial, see Figure 12a). We calculated the area under the ROC curve, A_z (see 5.4.3) and then applied the logit transformation to the resultant values to compare performance between conditions. We found that performance was significantly better than chance for the target/novel foil discrimination (mean= 2.527, $t_{144} = 25.96$, $p < 2.2 \times 10^{-16}$; $BF_{10} \approx 4.54 \times 10^{52}$ [the t was very large, so the function used an approximation]; Cohen's d= 2.16; see the A_z old/new in Figure 15) and for the target/similar lure discrimination (mean= 1.20, $t_{144} = 23.20$, $p < 2.2 \times 10^{-16}$; $BF_{10} \approx 1.08 \times 10^{47}$ [the t was very large, so the function used an approximation], Cohen's d= 1.93; see the A_z old/lure in Figure 13b). We also found that performance was significantly better for the target/novel foil discrimination than the target/similar lure discrimination (mean difference= 1.33, $t_{144} = 20.92$, $p < 2.2 \times 10^{-16}$, $BF_{10} \approx 1.22 \times 10^{42}$ [the t was very large, so the function used an approximation], Cohen's d= 1.737; also see Results: Comparison of the forced-choice and old/new test formats).

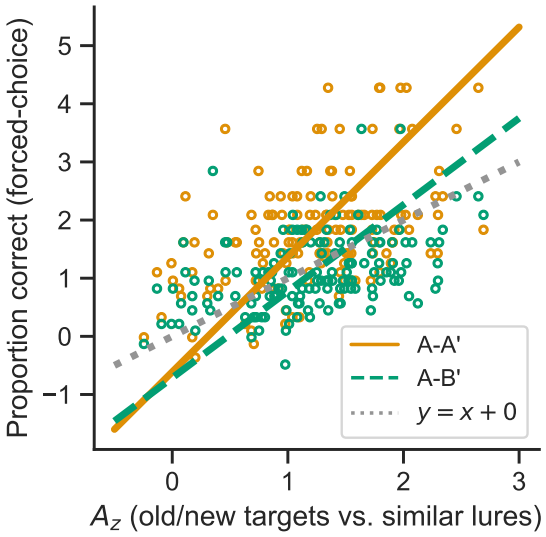
We next tested for our expected effect of a relationship between target-lure similarity and performance on the A-A' and A-B' test formats. Here, we fit a generalized linear model (binomial family) to the proportion of "new" responses to similar lures (note: the correct answer is "new", while the incorrect response is "old" because these are all similar lures; see Figure 1c). We found that the lure-bin model (based on the empirically defined mnemonic similarity; [31, 46, 62, 63]) accounted for significantly more variance than the intercept-only model for the similar lure data ($\chi^2(1, N = 145) = 195.93$, $p < 2.2 \times 10^{-16}$; intercept $= 0.31$, $\beta_{lure_bin} = 0.41$; null model $BIC = 3,353.4$, lure-bin model $BIC = 3,141.2$; see Figure 12b). Therefore, these results replicate and extend previous work that showed that performance varies in a monotonic manner with the differences of lure bins (i.e., mnemonic similarity) on the old/similar/new test format as well as our findings here on the A-A' and A-B' forced-choice test formats. Altogether, our results of the artificial neural network suggest that the transfer function between memory performance and the stimulated target-lure pattern similarity in area IT is approximately linear (see 3.1 and Figure 3), which extends to the A-A', A-B', and old/new test format with targets and similar lures.

3.3.3 Comparison of the forced-choice and old/new test formats

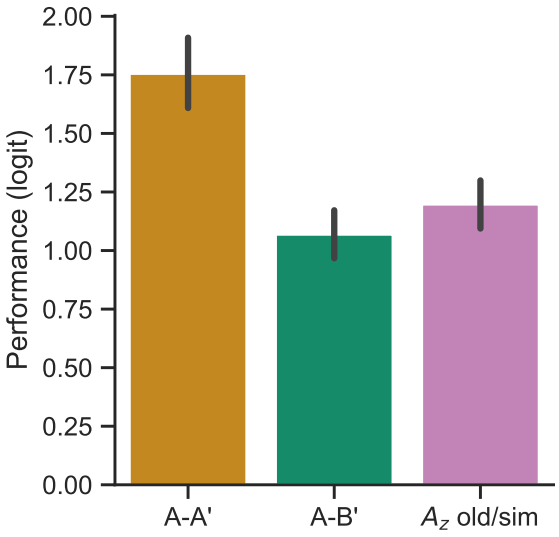
We next aimed to compare performance between the forced-choice and old/new test formats to test the predictions from the distributed memory models vs. the CLS hippocampal model. The distributed memory models predict that performance on the old/new test format with targets and similar lures will be most similar to the A-B' test format. Specifically, in the case of running many simulations and calculating the average performance at the group level (i.e., akin to a fixed-effects analysis), the distributed memory models predict that performance on the old/new test format with targets and similar lures will be equivalent to the A-B' test format, which is in line with previous theoretical and empirical findings for traditional recognition memory paradigms with targets and novel foils [72, 73, 74]. Note that in contrast to the predictions of the distributed memory models, the CLS hippocampal model predicts equivalent performance on the old/new test format with targets and similar lures, the A-A' test format, and the A-B' test format. Moreover, the distributed memory models predict that performance will be correlated across all of the test formats, including the formats with targets vs. novel foils (i.e., driven by individual differences in the learning rate). Thus, we performed several analyses to test the similarity between the forced-choice and old/new test formats.

First, we tested the distributed memory models' prediction that performance on the old/new test format with targets and similar lures would be closest to the A-B' test format. We first calculated the area under the receiver operating characteristic (ROC) curve, which is referred to as A_z (see 5.4.3). We then converted the performance scores with the logit function. We then ran a two-stage regression model to estimate the slope and intercepts of the relationship between the test formats with targets and similar lures. We observed that the estimated slope was relatively close to 1 for the comparison between the A_z for the old/similar lures and the A-B' test format (estimated slope= 1.49, estimated intercept= -0.71), which is in line with the predictions of the distributed memory models. Likewise, the estimated slope for the comparison between the A_z for the old/similar lures and the A-A' test format was (numerically) larger (estimated slope= 1.98, estimated intercept= -0.61), which is also in line with the predictions of the distributed memory models (see Figure 13a). To further assess the similarity between A_z for old/similar lures and the A-B' test formats, we next performed difference tests to compare the formats with targets and similar lures. A paired t-test revealed that performance was significantly better on the A-A' test format than the A_z for targets and similar lures (mean difference= 0.558, $t_{144} = 7.87$, $p = 7.91 \times 10^{-13}$; $BF_{10} =$

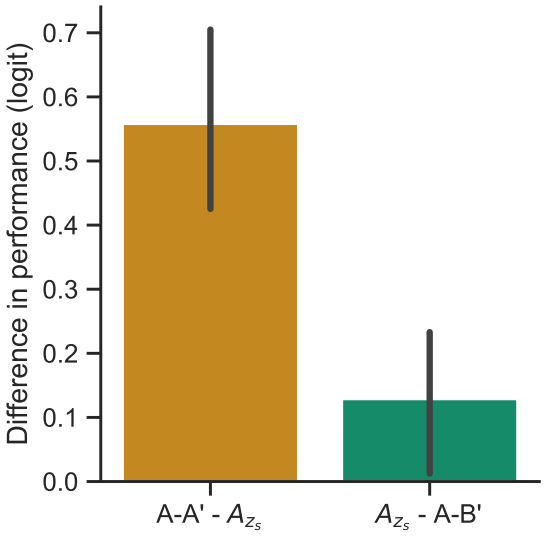
(a) Relationship between test formats



(b) Performance on test formats with similar lures



(c) Difference between old/new and forced-choice



(d) Dendrogram of differences in performance

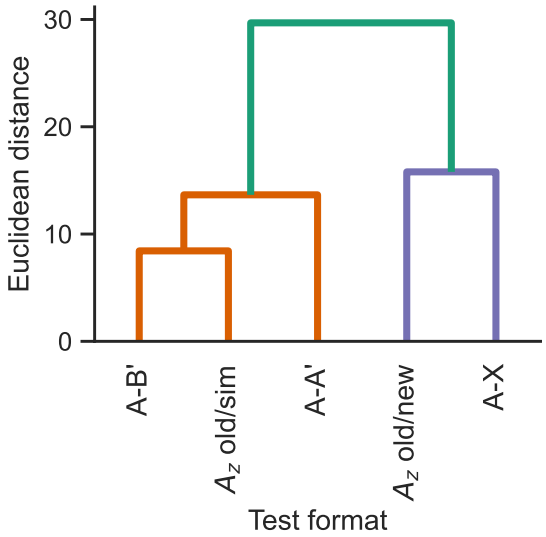


Figure 13: Evidence that performance on the old/new test format with targets and similar lures is most similar to the A-B' test format. (a) The comparison between the old/new test format with targets and similar lures vs. the A-A' and A-B' test formats. Note that the comparison with the A-B' test format produced the closest slope to 1. (b) The logit-transformed performance revealed that (c) performance was most similar between the A-B' test format and the old/new test format with targets and similar lures. (d) Dendrogram analysis also revealed clustering of the test formats. Altogether, these results support the predictions of the distributed memory models (see Figure 8, Figure 9, and Figure 14). Error bars represent the 95% confidence intervals.

9.16×10^9; Cohen's d= 0.653). A paired t-test also revealed that the A_z for targets and similar lures was (numerically) greater than performance on the A-B' test format (mean difference= -0.129,

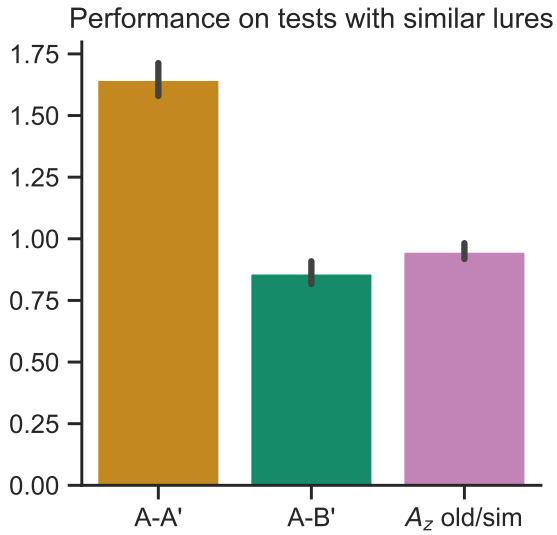
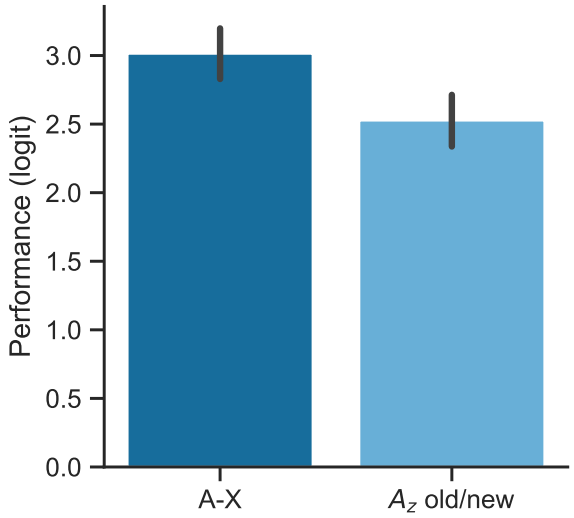


Figure 14: A follow-up, random-effects-style analysis of the distributed memory models revealed the same findings as the empirical data (see Figure 13b). This plot is from MINERVA 2, but the other distributed memory modes made the same predictions. Bars represent the mean and the error bars indicate the 95% confidence intervals.

$t_{144} = -2.24$, $p = 0.026$) but the magnitude of this effect was weak overall ($BF_{10} = 1.04$, Cohen's d= 0.186; see Figure 13b). To further test whether performance of the A_z for targets and similar lures was closer to the A-B' test format than the A-A' test format, we next performed a difference test. We found that the difference between the A-A' vs. A_z for targets and similar lures was significantly greater than the difference between the A_z for targets and similar lures vs. the A-B' test format (mean difference= 0.429, $t_{144} = 3.92$, $p = 0.00014$, $BF_{10} = 119.1$, Cohen's d= 0.325; see Figure 13c). Altogether, these results suggest that performance on the A_z for targets and similar lures is numerically closer to the A-B' test format than the A-A' test format, thus supporting the predictions of the distributed memory models.

We next went back to the computational models to determine whether the empirical effect of slightly enhanced performance on the old/new test format with targets and similar lures vs. the A-B' test format could be explained within the distributed memory models. Briefly, in our original simulations, we combined the data from all of the simulated participants in our calculations of performance between the test formats. Under those conditions, which are akin to a fixed-effects analysis, we found that performance on the old/new test format with targets and similar lures was equivalent to performance on the A-B' test format. Additionally, this is a case with a (very) large number of data points for each condition. Thus, we were interested in modeling the effects at the level of individual participants (i.e., akin to a random-effects analysis). Interestingly, we found that the models predicted a significant advantage for the old/new test format with targets and similar lures relative to the A-B' test format in this analysis. Specifically, the models predicted that performance on the A-A' test format is better than the A_z with targets and similar lures, the A_z with targets and similar lures is better than A-B', and the difference between the A-A' and A_z is great than the difference between A_z and A-B' (for example, see the results from MINERVA 2 in Figure 14; the results were the same in the other distributed memory models). Therefore, while

(a) Performance on tests with novel foils (including all participants)



(b) Performance on tests with novel foils (excluding participants with $\geq 34/35$ correct on A-X)

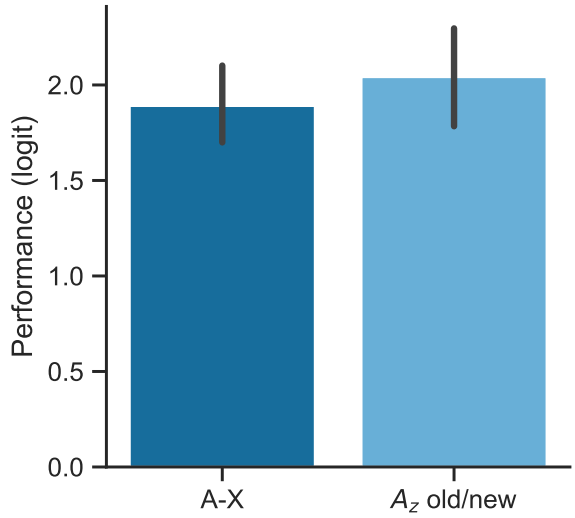


Figure 15: Comparison of performance on the A-X and old/new test format with targets and novel fils indicated that performance was better on the A-X test format, but this might be driven by a ceiling effect in the A-X test format. (a) When we included all participants, we found that A-X performance was significantly better than A_z with targets and novel foils. (b) However, when we tested whether this simply indicated a ceiling effect (here, we limited our analysis to participants with fewer than 34/35 correct on the A-X format), we observed that the A-X $>$ A_z old/new advantage now failed to reach significance, thus suggesting a ceiling effect was driving the original effect. Overall, these results are consistent with the predictions of the distributed memory models (see Figure 8 and Figure 9). Bars represent the mean and the error bars indicate the 95% confidence intervals.

this was not part of our pre-experimental modeling (and thus we temper our interpretations here), we think that these results provide further support for the notion that the distributed memory models at can capture performance on various versions of the Mnemonic Similarity Task.

We next aimed to further examine the relationship between performance on all five test formats. A dendrogram analysis of the Euclidean distance between performance on all 5 test formats revealed that A_z for targets and similar lures was most similar to the A-B' test format (see Figure 13d) and the A-A' test format was in a similar branch of the dendrogram results as the A_z for targets and similar lures as well as the A-B' test format. Likewise, the A_z for targets and novel foils was in cluster with the A-X test format. For completeness, we compared the A_z for targets and novel foils vs. the A-X test format and we found that performance was significantly better on the A-X test format (mean difference= 0.487, $t_{144} = 4.80$, $p = 3.91 \times 10^{-6}$, $BF_{10} = 3,356.1$, Cohen's d= 0.399; see Figure 15a). However, we wondered whether this could be a spurious ceiling effect driving this difference in the A-X test format. Specifically, a large number of participants (81 of 145) exhibited perfect (35/35) or near perfect (34/35) performance on the A-X test format. Therefore, we tested whether there was still a test format effect in participants that were correct on fewer than 34 of the 35 trials on the A-X test format (which resulted in 64 participants in the analysis and 81 excluded participants). Interestingly, the apparent A-X $>$ A_z old/new effect disappeared

Figure 16: We observed significant correlations between all test formats, thus supporting the predictions of the distributed memory models (i.e., compare the results here to Figure 10).

under these conditions (mean difference= -0.151, $t_{63} = 1.22$, $BF_{01} = 3.60$ [i.e., in favor of the null hypothesis], Cohen's d= 0.152; see Figure 15b), thus suggesting that a ceiling effect was driving the difference between these test formats. Altogether, the finding of clustering in the dendrogram analysis matches the predictions of the distributed memory models.

We next aimed to test the prediction from the distributed memory models that performance would be correlated on all of the forced-choice and old/new test formats. Consistent with the predictions of the distributed memory models, we found a significant correlation between all test formats (see Figure 16). Specifically, all of the forced-choice test formats exhibited a significant correlation (A-X vs. A-A': $r = 0.62$, $t_{143} = 9.37$, $p < 2.2 \times 10^{-16}$, $BF_{10} = 3.61 \times 10^{13}$; A-X vs. A-B': $r = 0.46$, $t_{143} = 6.13$, $p < 8.27 \times 10^{-9}$, $BF_{10} = 1.70 \times 10^{6}$; A-A' vs. A-B': $r = 0.53$, $t_{143} = 7.48$, $p < 6.69 \times 10^{-12}$, $BF_{10} = 1.38 \times 10^{9}$). Moreover, all three forced-choice test formats exhibited a significant correlation with the old/new test format with targets and novel foils (A-X vs. A_z old/new: $r = 0.46$, $t_{143} = 6.26$, $p < 4.25 \times 10^{-9}$, $BF_{10} = 3.18 \times 10^{6}$; A-A' vs. A_z old/new: $r = 0.41$, $t_{143} = 5.41$, $p < 2.56 \times 10^{-7}$, $BF_{10} = 6.91 \times 10^{4}$; A-B' vs. A_z old/new: $r = 0.36$, $t_{143} = 4.63$, $p < 7.98 \times 10^{-6}$, $BF_{10} = 2.86 \times 10^{3}$) and the old/new test format with targets and similar lures (A-X vs. A_z old/sim: $r = 0.42$, $t_{143} = 5.51$, $p < 1.60 \times 10^{-7}$, $BF_{10} = 1.07 \times 10^{5}$; A-A' vs. A_z old/new: $r = 0.47$, $t_{143} = 6.43$, $p < 1.77 \times 10^{-9}$, $BF_{10} = 7.24 \times 10^{6}$; A-B' vs. A_z old/new: $r = 0.42$, $t_{143} = 5.48$, $p < 1.86 \times 10^{-7}$, $BF_{10} = 9.32 \times 10^{6}$). Finally, the comparison between the two old/new formats relies upon data that are not independent, but we report the correlation here for completeness ($r = 0.81$). Altogether, our finding of significant correlations between all of the test formats supports the predictions of the distributed memory models.

4 Discussion

We aimed to develop and test the predictions between competing computational models of human performance on memory tests with targets, similar lures, and novel foils. As we discussed above, we believe that these test formats provide a critical testbed for comparing competing models of human memory. Specifically, we compared predictions of the CLS hippocampal model [16, 17]—which emphasizes sparse, pattern-separated representations for avoiding interference between similar memories—vs. distributed memory models—which emphasize distributed representations (see our modeling results above). We focused our empirical analyses on areas in which the two classes of models make competing predictions: 1) the nature of the transfer function between performance and target-lure similarity, 2) whether or not there would be a significant effect of test format on performance of memory tests with targets and similar lures, and 3) whether different memory tasks would produce correlated or dissociable performance via individual differences in performance. Our results clearly favored the predictions of the distributed memory models over the predictions of the CLS hippocampal model: 1) we found an approximately linear transfer function between performance and the simulated target-lure similarity in an artificial neural network of area IT, 2) we observed systematic differences in performance between the forced-choice and old/new test formats, and 3) individual differences analyses suggest that differences in learning rate lead to correlated performance on all 5 test formats, which we discuss in turn below. We also discuss implications of our modeling and empirical results for future research using behavioral, neural, and computational modeling approaches for studying interference resolution in human memory.

4.1 The transfer function between memory performance and the simulated target-lure pattern similarity in area IT appears to be approximately linear

Tasks that include targets and similar lures typically aim to tap into the pattern-separation mechanisms of the hippocampus (e.g., the Mnemonic Similarity Task [46, 62, 63, 65], which had been used in more than 100 behavioral and neuroimaging publications as of 2019 [46]). The CLS hippocampal model and the distributed memory models make competing predictions about the degree to which target-lure similarity will influence memory performance. Specifically, the CLS hippocampal model predicts that humans will only exhibit interference when target-lure similarity is very high or when the similarity between all items in a list is very high (see Figure 2a and [16, 17]). We aimed to generate an objective measure of the target-lure similarity using a biologically inspired artificial neural network, CORnet-S [60, 64], to compare the pattern similarity of simulated activity within area IT, which is heavily interconnected with perirhinal cortex [61]. Specifically, extent models of the medial temporal lobe predict that area IT would project strongly into the perirhinal cortex [61], which would then project to entorhinal cortex and into the hippocampus (e.g., [3, 75, 76, 77, 78]). Models of hippocampal pattern separation posit that the hippocampus, and the dentate gyrus in particular, will decrease the pattern similarity between targets and similar lures relative to its input patterns from the neocortex (e.g., [14, 15, 16, 17, 79]). Therefore, if human participants rely on the pattern-separated representations of the dentate gyrus to make memory judgments (i.e., based on previous neuroscientific findings; e.g., [27, 28, 29, 31, 32, 36]), then we would expect behavior to follow a nonlinear function and only exhibit interference for very high levels of target-lure similarity in its input regions (see Figure 2a and [16, 17]). Conversely, as we reported here, the distributed memory models predict that the degree of target-lure similarity will exhibit a strong, monotonic, and relatively continuous influence on memory performance (MINERVA 2: Figure 4b, TODAM: Figure 5b, the familiarity-based Hopfield autoassociative neural network: Figure 6b, and the recall-based Hopfield autoassociative neural network: Figure 7b; also see the medial temporal lobe cortex

model from [16, 17]). Therefore, testing these competing predictions can shed important light on the underlying mechanisms of human memory.

Our results provide evidence for what appears to be a linear mapping between mnemonic discrimination and simulated target-lure pattern similarity in area IT. Specifically, when we compared the resultant simulated pattern similarity in area IT to the original behavioral data that was used to generate the lure bins of the stimuli in the Mnemonic Similarity Task [46, 62, 63, 65], we observed evidence of an approximately linear transfer function between mnemonic discrimination performance and the simulated target-lure similarity in area IT (see Figure 3). Moreover, in our current empirical data, we replicated the effect of the influence of lure bins on memory performance on the old/new test format with targets and similar lures (see Figure 12b; e.g., [31, 46, 62, 63]). Here, we also extended the finding of an effect of the lure bin on performance of both the A-A' and A-B' test formats (see Figure 11b). The finding that the lure similarity effect was not significantly different between the A-A' and A-B' test formats (and that it was significant in both test formats independently) suggests that target-lure similarity affects performance across all of these test formats. Moreover, combined with the simulated pattern similarity results in area IT, our results further suggest that the transfer function mnemonic discrimination and target-lure similarity is approximately linear in the old/new, A-A', and A-B' test formats.

While our study is the first (to our knowledge) to test the transfer function between mnemonic discrimination performance and target-lure similarity using an artificial neural network, Motley and Kirwan [33] created a parametric version of the Mnemonic Similarity Task in which they modulated the target-lure similarity by parametrically rotating the objects to form the similar lures (rotations: 15°, 25°, 35°, and 55°). Interestingly, they found that behavioral performance (the proportion of times that participants responded "old" as well as "rotated"="similar") varied linearly as a function of the target-lure rotation. Therefore, our results extend these findings from their parametric manipulation of target-lure similarity. Given the widespread use of the Mnemonic Similarity Task and other similar measures, we argue that our results provide an interesting alternative view about the nature of the underlying representations that support performance on the task. Specifically, our results suggest that mnemonic discrimination performance on the task is better captured by the distributed memory models than the pattern separation mechanisms of the dentate gyrus.

As we discussed in the Introduction, previous fMRI research using the Mnemonic Similarity Task (and related paradigms) have generally found a nonlinear mapping between the response profile of the dentate gyrus/CA3 and the mnemonic similarity of target-lure pairs but a linear relationship between the response profile of CA1 and the mnemonic similarity of target-lure pairs (e.g., [31, 36]; for review see: [32]). Therefore, given that mnemonic similarity was operationally defined as mnemonic discrimination (i.e., 1-p('old'); [31, 46, 62, 63]), these findings also reveal a potential disconnect between the dentate gyrus/CA3 and memory performance. In contrast, the linear relationship between CA1 activity and mnemonic similarity suggests that CA1 may play a stronger role in behavioral performance. Moreover, previous computational modeling and empirical research suggests that CA1 might play a role in match-mismatch detection (e.g., via a match-mismatch mechanism; [48, 49]) and previous research has shown a linear transfer function between response profiles of CA1 and environmental/stimulus similarity (e.g., [47, 50, 51, 52]). Therefore, a possible explanation for our findings here is that participants rely on neural signals other than pattern-separated representations in the dentate gyrus to make memory-based judgments on tasks with targets and similar lures. For example, the underlying representations of the distributed models that we use here are similar to the representations of the neocortex (e.g., perirhinal cortex e.g., see the neocortical model from [15, 16, 17]) and CA1. Additionally, given the functional similarity of

the Hopfield autoassociative neural network to area CA3 (cf. [67, 80, 81, 82]), it is possible that CA3 might play a key role in memory performance on these tasks. Moreover, a recall-based version of the Hopfield autoassociative neural network was previously shown to account for performance on recall tasks (i.e., in addition to the recognition memory tasks that we used here (e.g., [69]), thus suggesting it could potentially be a more general memory mechanism. Finally, it is possible that memory-based decisions are driven by the accumulation of evidence throughout the brain (i.e., in a more holistic, global manner; cf. [4, 5, 16, 17]).

While the CLS model predicts that the hippocampus will generally not be useful for computing the global similarity between stimuli (i.e., given its use of pattern separation; [16, 17]), as we discussed in the Introduction, Davis et al. [9] found that the global similarity between targets and foils in the hippocampus and other areas of the medial temporal lobe was related to memory-based decisions in a recognition memory paradigm. Therefore, their results suggest that there are cases in which the hippocampus would be able to compute the global match, which would, in turn, be related to memory-based decisions. Combined with our findings here, we argue that it will be interesting for future research to further investigate the degree to which we make memory decisions based on the purported pattern-separated representations in the dentate gyrus vs. distributed representations in the neocortex (or the hippocampus) vs. combined evidence from all of these sources (cf. [4, 5, 16, 17]).

In summary, the results of the transfer function between mnemonic discrimination performance and target-lure similarity in input regions of the hippocampus appears to be linear, which is better fit by the predictions of the distributed memory models than by the pattern-separation mechanisms of the dentate gyrus. Moreover, we argue that when we reexamine the findings from previous fMRI studies, the evidence actually points toward a closer relationship between mnemonic discrimination performance and activity in areas outside of the dentate gyrus (e.g., CA1, perirhinal cortex). We return to the discussion of other possible reasons that the behavior did not fit the predictions of the standard version of the CLS hippocampal model (see 4.4).

4.2 The test format effect on the forced-choice tests favors the distributed memory models

We compared the performance of several computational models at accounting for human behavioral performance on different test formats of the Mnemonic Similarity Task (the CLS hippocampal model: Figure 2a; MINERVA 2: Figure 4b, TODAM: Figure 5b, the familiarity-based Hopfield autoassociative neural network: Figure 6b, and the recall-based Hopfield autoassociative neural network: Figure 7b; also see the medial temporal lobe cortex model from [16, 17]). One of our key findings is an effect of differences in performance as a function of the test format. For example, for the forced-choice test formats, performance was best on the A-X test format, followed by the A-A' test format, followed by the A-B' test format (see Figure 11a). All of the models that can account for this effect of test format—and in particular the A-A' > A-B' effect—suggest that encoding variability plays a key role in this effect (cf. Simulation 4 from [16]; [8, 57]). Specifically, the models predict that encoding variability leads to an increased number of errors on the A-B' test format because sometimes the encoding of the original B item exceeds the encoding of the original A items. In this situation, the response to the B' item can trigger a stronger global memory signal because it matches more features in memory. For example, when encoding variability is minimized, the difference between the A-A' and A-B' test formats diminishes (cf. Figure 4 from [57] with MINERVA 2). Interestingly, Rollins et al. [58] used eye tracking to test whether performance on the A-B' test format varies as a function of encoding variability, as indexed by the number of fixations. Specifically, they found a significant test-format (A-A' vs. A-B') by subsequent memory accuracy

(correct vs. incorrect) interaction. A follow-up t-test showed that participants made significantly more fixations (a measure of encoding strength) to B items than A items on subsequently incorrect trials of the A-B' forced-choice test format. Therefore, their findings supported a key prediction from the distributed memory models that the impaired performance on the A-B' test format relative to the A-A' test format is influenced by encoding variability (i.e., better encoding of the original B item than the original A item). Future research can further test these predictions using other measures of encoding strength (e.g., subsequent memory effects with fMRI and EEG).

Counter to the prediction of an effect of encoding variability leading to better performance on the A-A' test format relative to the A-B' test format, the standard CLS hippocampal model predicts that performance will be similar on these two test formats (see Figure 2a; [16, 56]) even with simulated encoding variability (see Simulation 4 from [16]). In fact, the CLS model predicts that encoding variability will influence performance in the neocortical model but not the hippocampal model. They suggest that the pattern separation and recall-to-reject mechanisms of the hippocampus would give it two opportunities to reject the items in the A-B' test condition (i.e., once for the A item and once for the B' item), thus any benefit of the covariance of the items in the A-A' test format is offset by the additional ability for recall-to-reject on A-B' test format.

Although our behavioral results here as well as in other studies [8, 22, 54, 57, 58, 59] are discordant with the CLS hippocampal model's prediction of a null effect of a difference between performance on the A-A' vs. A-B' test formats, we suggest that there could be modifications under which the CLS hippocampal model would account for the extant behavioral findings. First, the output of the hippocampal model and the neocortical model could be combined and a separate neural network could be used to model decision making (cf. [16, 17]). In this situation, the combined output could predict an A-A' > A-B' test format because the neocortical model already predicts such an effect (i.e., when there is encoding variability). As a second approach, we predict that different learning rules for implementing encoding variability might lead to the A-A' > A-B' test format effect. Norman and O'Reilly [16] implemented encoding variability by universally scaling the learning-rate parameter on all of the weights. Specifically, for each trial, they multiplied the learning rate by a variable encoding strength (see Simulation 4 from [16]). In effect, this changes the encoding strength of all features uniformly. In contrast, other approaches for implementing encoding variability could be to use either a connection-wise probabilistic encoding rule (e.g., as applied here for the Hopfield/autoassociative neural network; [69]) or a feature-wise probabilistic encoding rule (e.g., as applied here for MINERVA 2 and TODAM). Specifically, the connection-wise probabilistic learning rule would essentially model random variability in terms of how many connections are modified. Likewise, the feature-wise probabilistic learning rule would essentially model random variability in terms of whether or not stimulus features were properly activated or stored during the original encoding phase (e.g., we can imagine that we neither pay attention to nor memorize every feature in any given trial). In either situation, the number of features as well as which specific features are encoded would be variable on each trial. For example, the Hopfield/autoassociative neural network that we used here predicted a test format effect. In effect, the connection-wise or the feature-wise probabilistic learning rule would also allow the encoding variability to interact with the sampling variability (e.g., in some cases encoding more of the inactive units than the active units). Interestingly, encoding variability within Hopfield autoassociative neural networks was also found to be important for associative recall tasks, thus suggesting a possible extension and connection to different behavioral paradigms [69]. Given the structural similarity of the Hopfield/autoassociative neural network to the CA3 subregion of the hippocampus (cf. [67, 80, 81, 82]), we think it is possible that such a change to the hippocampal network will account for the test-format effects that we observed in our empirical data, and we plan to explore this possibility in future work.

4.3 The relationships between performance on the forced-choice and old/new test formats favor the distributed memory models

The distributed memory models predict specific relationships between performance on the forced-choice and old/new test formats: performance on the old/new test format will be similar to performance on the A-X test format, and performance on the old/new test format with targets and similar lures will be significantly worse than performance on the A-A' test format and most similar to performance on the A-B' test format (see Figure 9 and Figure 14). Conversely, the standard CLS hippocampal model predicts that performance will be virtually the same across all of the test formats with targets and similar lures. Our results fit with the predictions of the distributed memory models (see Figure 13 and Figure 15). Moreover, the distributed memory models predict that performance across all 5 test formats (including comparisons between test formats with targets and novel foils as well as targets and similar lures; see Figure 10). In contrast, the CLS model predicts that there are many cases in which performance would not be correlated between test formats [16]. We observed clear evidence of relationships between performance on all of the test formats (see Figure 16), thus supporting the predictions of the distributed memory models. We note that Norman and O'Reilly [16] also found that under conditions of high levels of encoding variability that the combined CLS model did predict there would be correlations between performance on different test formats. Thus, our finding of correlated performance between all 5 test formats is not necessarily discordant with all of the predictions the CLS model, but our results are naturally accounted for by the distributed memory models. It will be interesting for future studies to try to constrain encoding variability to test the strong prediction of the CLS model that the relationship between test formats would disappear in the absence of encoding variability; however, future computational modeling work should also examine how changed in encoding variability would change the predictions of the distributed memory models (cf. [57]). Therefore, we hope that our empirical and modeling results will inform future research that seeks to explore the underlying mechanisms that explain the correlated performance that we observed here. We also think that our results provide an important cautionary tale about making inferences that these test formats (e.g., the comparison between test formats with similar lures compared to traditional recognition memory paradigms with novel foils; the comparison between the forced-choice test formats such as the A-A' and A-B' test formats) tap into fundamentally different cognitive or neural processes.

4.4 Avenues for testing the generality of our empirical and modeling results

While our results clearly favor the distributed memory models, we think there are many other cases in which the CLS hippocampal model could be favored. For example, it is possible that task designs with so many objects could overload the pattern separation capabilities of the hippocampus. For example, participants in these types of paradigms view more than 100 images and have to make fine-grained decisions about them in the matter of a short period of time. Therefore, it is possible that under these circumstances, participants rely on the more distributed representations of the neocortex and that even the hippocampus exhibits a relatively high degree of pattern overlap (cf. [16, 17]). More recent theories suggest that event boundaries and schema knowledge exert a profound influence on human memory (e.g., [83]). For example, the hippocampus appears to respond most prominently at event boundaries, thus suggesting that the hippocampus plays a key role in encoding memories around event boundaries (e.g., [84, 85, 86]). Therefore, it is possible that memory paradigms that do not incorporate event-like structure might be harder for participants to parse, learn, and remember (i.e., because the sequence of images is essentially random and thus lacks event-like structure). It will be interesting for future experiments to combine more naturalistic

approaches with the study of the transfer function between mnemonic discrimination and event similarity. Another intriguing idea is that expertise can help us avoid interference between similar stimuli or memories. For example, people with domain expertise can rapidly detect and report differences between similar species of animals, similar faces, and other similar task stimuli such as Greebles (e.g., [87, 88, 89, 90, 91, 92, 93]). Likewise, in our daily lives, we might be experts in our knowledge of typically recurring experiences (e.g., between events that contain similar, familiar people and places). Thus, it is possible that schema knowledge could aid in our ability to distinguish between similar experiences. Therefore, it will be interesting for future research to examine whether more naturalistic designs, that incorporate event schemas and expertise, confer protection from interference between similar memories as well as the neural mechanisms that support behavioral performance under these conditions.

A related possibility is that the hippocampus might play a stronger role in the pattern separation of memories that are inherently associative, such as spatial or temporal memories. Several prominent theories suggest that the hippocampus plays a key role in binding items in context or in spatial memory more generally (e.g., [3, 75]). For example, previous research has provided evidence of pattern separation between similar spatial environments in rodents [27, 28] and humans [94, 95]. Likewise, future research could further leverage temporal memory paradigms (e.g., [96, 97, 98]) to investigate whether there is a more linear relationship between mnemonic discrimination and hippocampal activity patterns. Therefore, it is possible that the humans would exhibit less interference for distinguishing between similar stimulus-stimulus associations, spatial memories, or temporal sequence memories. Future research could leverage novel advances in virtual reality technology to address these possibilities, especially regarding changes in spatial environments, which could be based on previous neuroscience research [27, 28, 94, 95].

4.5 Conclusion

We generated and tested the predictions of distributed memory models vs. the standard CLS hippocampal model on several test formats that included targets, similar lures, and novel foils. All of our empirical data were better fit by the predictions of the distributed memory models vs. the CLS hippocampal model, thus providing an important challenge for influential theories about the cognitive neuroscience of memory. Our behavioral and computational modeling results provide an important alternative framework and can provide novel avenues for future behavioral, neuroscientific, and computational modeling research.

5 Materials and Methods

5.1 Participants

Participants consisted of 153 undergraduate students at Colby College. The mean age of participants was 19.5 years old, with a standard deviation of 1.17 and a range between 18 and 22. 108 participants identified as female, 42 participants identified as male, 2 participants identified as non-binary, and 1 participant used another term or option not listed. Their participation lasted no longer than 60 minutes, and all participants consented to participation in accordance with the Institutional Review Board at Colby College. Participants were recruited within the Colby Psychology Department, and they received course credit for their participation. We planned to collect data from at least 100 participants given that some of our measures included studying individual differences and correlations between performance on different test formats. To ensure the success of the Receiver operating characteristic (ROC) analysis (see below), we excluded 5 participants

whose answers were not distributed across confidence bins (e.g., if they only used two confidence bins). We also excluded 3 participants whose responses in the encoding phase were biased (i.e., if they selected 10 times more 'outdoor' than 'indoor' or vice versa, or if they did not respond to more than 100 trials) and the correct rate in all test formats of the forced-choice version were less than 0.7. The final analysis included 145 participants.

5.2 Behavioral task design

The Mnemonic Similarity Task (MST) [46, 57, 62, 63, 65] has been developed to study memory performance for targets, similar lures, and novel foils. The MST has been employed with several versions of the test format. The typical versions of the task include an Old/Similar/New test format, in which participants are instructed to respond "old" if they think that they have previously seen the exact item, "similar" if they think that they have previously seen a similar item, and "new" if they think that they did not see this item or an item like it in the encoding phase. Other studies have used an Old/New test format, in which participants are instructed to respond "old" if they believe that they saw the exact item and to respond "new" otherwise (i.e., to similar lures and novel foils; cf. the "veridical" version of the task in Experiment 4 in [63]). Moreover, studies using the Old/New test format have also occasionally included confidence ratings, which allows for analysis with methods from signal detection theory, including the receiver operating characteristic (ROC) curve (cf. [63, 99]). Finally, previous studies have also studied performance on a forced-choice version of the task (e.g., [57, 58]; cf. [22, 54]). Specifically, participants in these experiments were tested under three forced-choice formats: A-X (a target and an unrelated, novel foil; traditional forced-choice recognition), A-A' (a target and a corresponding similar lure; e.g., a picture of an apple and a similar apple), and A-B' (a target and a noncorresponding similar lure; e.g., a picture of an apple and a basketball that is similar to one that was studied in the encoding phase; note, we use the nomenclature of Tulving [54]). Therefore, these various test formats allow us to test how the test format itself influences performance on the task, which we explore here.

Our experiment consisted of two versions of the Mnemonic Similarity Task (forced-choice and old/new with confidence ratings), one automatic 10-minute rest between the two versions, and questionnaires at the end. We ran our experiments online within the `cognition.run` platform. After the welcome and introduction messages, the experiment website randomly assigned participants to either begin with the forced-choice version of MST or the old/new version. In both versions, participants first performed an encoding task in which they indicated whether each presented image depicted more of an indoor or an outdoor item (i.e., the indoor/outdoor task, which is an incidental encoding task; see Figure 1). Each encoding phase consisted of 140 images, which were displayed for 2,000 ms with a 500-ms interstimulus interval.

In the test phase of the forced-choice MST (see Figure 1), participants performed a memory test that contained three test formats: 1) A-X (i.e., a target and a novel foil), 2) A-A' (i.e., a target and a corresponding similar lure), and 3) A-B' (i.e., a target and a noncorresponding lure object from a different pair), as in our previous research [57]. Each test format included 35 trials. We matched the "lure bin" (i.e., similarity level) of the target and lure/foil item in each pair of stimuli to account for any potential effects of encoding difficulty of the stimuli (as in [57]). The test formats were presented in a random order on a trial-by-trial basis. Every trial contained two objects, one on the left side of the screen and one on the right side, and the target was randomly assigned to one side on each trial. Participants were instructed to select the exact object they saw during the encoding phase via button press. The images were displayed either until the participant made a response or for 4 s, after which the image disappeared but the participants had an unlimited

response window.

In the test phase of the old/new MST, participants performed a memory test that contained three types of objects: 1) targets (i.e., exact repetitions of images they saw during the indoor/outdoor task), 2) similar lures (i.e., images that were similar to one they saw during the indoor/outdoor task), 3) novel foils (i.e., new images that were not in a similar image pair to one they saw during the indoor/outdoor task). The task included 70 target, 70 similar lures, and 52 novel foil trials (i.e., the remainder of the number of available stimuli because each MST image set has 192 images). On each test trial, one object was presented in the center of the screen, and participants were instructed to identify whether they saw it during the encoding phase (i.e., "old" object) or not (i.e., "new" object) via button press. We instructed participants to respond "old" if they saw this exact image during the indoor/outdoor task (i.e., they should respond "old" only to targets) and to respond "new" otherwise (i.e., to novel foils and to similar lures). The images were displayed until participants made a response or for 2 s, after which the image disappeared and left an unlimited response window. Following the old/new response, participants indicated their confidence in their decision, with the following options: "very sure", "somewhat sure", "somewhat unsure", "very unsure". As we describe below, we were able to convert the two responses (i.e., old, new) and the four confidence bins to a total of 8 possible responses (i.e., 2×4) for the ROC analysis.

After the MST, participants completed questionnaires about their demographic information, video game and GPS use, autobiographical memory, sleepiness, and COVID-19 information. Specifically, we used the Survey of Autobiographical Memory (SAM; [100]) to assess participants' trait mnemonics and Stanford Sleepiness Scale (SSS; [101]) to assess participants' sleeping performance. We did not analyze the questionnaires here.

5.3 Computational modeling methods

5.3.1 Global matching models: MINERVA 2 and TODAM

We previously found that MINERVA 2 and TODAM can account for many aspects of performance on the forced-choice versions of the Mnemonic Similarity Task [57] (cf. [8]). Specifically, these models suggest that performance on the tasks will vary as a function of test format: A-X > A-A' > A-B', similar to our previous empirical data. Here, we again used these models to replicate the finding of the test format effect. Additionally, we extended these models to make novel predictions, which we tested in our empirical data.

We modeled both MINERVA 2 [8, 66] and TODAM [7] in Python. To model items in MINERVA 2, we drew random integers of -1, 0, and 1 using the randint function from numpy. To generate similar lures, we randomly redrew from the random integers for a specified proportion of the features using the binomial function from numpy to select the features to be changed. We modeled learning by having the model learn each feature with probability of L and to not learn each feature with probability 1 - L. Specifically, we used the binomial function from numpy to select which features would be remembered for each trial. We then multiplied this binomial matrix by the encoding list to simulate the learning phase. Thus, the learning phase generated a matrix with M rows (i.e., memory traces) and N columns (i.e., features). During the testing phase, we calculated the summed similarity of the probe to the memory matrix using the following equations [8, 66]:

$$s_i = \left(\frac{p \cdot T_i}{n_i} \right)^3 \tag{1}$$

where p is the probe item, T_i is one of the items in memory (i.e., one of the rows of the memory

matrix), and n_i is the number of features that are nonzero in either p or T_i (i.e., the part of the equation within the parentheses is a normalized dot product, which calculates the match between the probe and a given memory trace and normalizes by the number of nonzero features). The cubing function makes the match nonlinear, which can be thought of as a form of pattern separation. We then summed the similarity values across all items in the memory matrix using the following equation [8, 66]:

$$g = \sum_{i=1}^{M} s_i \tag{2}$$

The encoding phase of MINERVA 2 is an instantiation of a multiple-trace model. However, the nature of the way that memory retrieval is modeled makes it a global matching model (i.e., because it sums the similarity across all memory traces to calculate the overall memory match for each probe item).

We modeled TODAM using the item-only version of the model [7]; however, TODAM has traditionally focused on modeling associative memory (e.g., [6]). Items in TODAM are represented as vectors, similar to MINERVA 2. Each feature of the item vectors in TODAM is generated as a random draw from a normal distribution with mean 0 and standard deviation $\sqrt{P/N}$, where N is the number of features (i.e., the length of the item vector) and in our case P is equal to 1. Note that setting the standard deviation with P equal to 1 has the desirable effect of making the item vectors have approximately unit length, thus allowing us to calculate the similarity using the dot product (i.e., this has a normalizing effect on the calculation of the dot product; [6]). To generate a similar lure item ($f_{j'}$) for a given target item (f_j), we used the following equation [7]:

$$f_{j'} = \rho f_j + \left(\sqrt{1-\rho^2}\right) g_j \tag{3}$$

where ρ indicates the desired similarity between the target item and the similar lure item and g_j represents an independent random vector. Here, the expected value of the dot product (i.e., similarity) between a target item and the resultant similar lure is ρ. To model the learning phase of the item-only version of TODAM, we calculated the memory vector with the following equation ([7]; pg. 105 from [102]):

$$m_t = \alpha m_{t-1} + B_t f_t \tag{4}$$

where α is a forgetting parameter (i.e., the opposite of a retention parameter, where 0 indicates complete forgetting of previous memories and 1 indicates that the new memory is added to the memory vector from the previous trial without any forgetting), m_{t-1} is the memory vector from the previous trial, B_t is a vector indicating which features should be learned (more information below), and f_t is the item that is presented at time t in the encoding phase of the task. B_t is generated by drawing from the binomial function in numpy with the probability of encoding a feature equal to p (i.e., the learning rate of the model) and not encoded with probability 1 - p. Therefore, p in TODAM is isomorphic to L in MINERVA 2. In our current simulations, we set α to be equal to 1 because we were not modeling the effects of list position effects and this makes the instantiation more similar to MINERVA 2.

During the test phase, we calculated the global match, g, of a probe item to the memory vector using the dot product (using the dot function in numpy):

$$g = p \cdot m \tag{5}$$

where p is the probe item and m is the memory vector. MINERVA 2 uses a cube function to generate a nonlinear mapping between the match and the summed similarity; however, TODAM

uses a linear similarity function. Also, because TODAM assumes that memory is stored in a single memory vector, the calculation of the similarity is naturally a measure of summed similarity.

For both the MINERVA 2 and TODAM, we simulated forced-choice performance and old/new performance using the general methods described below (see 5.3.3).

5.3.2 The Hopfield/autoassociative neural network

We also modeled performance using a Hopfield/autoassociative neural network [68, 69] within Python. Hopfield/autoassociative neural networks are a fully connected, single-layer recurrent neural network that shares functional overlap with the structure and function of the CA3 subregion of the hippocampus (cf. [67, 80, 81]). For example, the majority of connections in CA3 come from the recurrent collaterals (i.e., inputs from other CA3 neurons). For example, Wilson et al. [82] estimated that more than 95% of the inputs to CA3 neurons are recurrent collaterals. Therefore, while this model is still a form of an abstract neural network (i.e., it is not specifically designed to model the anatomical and functional properties of the brain), it bears resemblance to the anatomical, functional, and computational properties of CA3.

Items are also represented as vectors in the Hopfield/autoassociative neural networks. Specifically, we generated items with the features of -1 and 1, with an equal probability of either feature. These values can be thought of as modeling activity levels that are lower, -1, or higher, 1, than the baseline or average activity (pg. 165-166 from [102]). To generate similar lure items, we followed a similar procedure to our approach in MINERVA 2. Specifically, we randomly redrew features (i.e., -1 and 1) for a specified proportion of the features using the binomial function from numpy to select the features to be changed.

To simulate learning, we used a probabilistic Hebbian learning paradigm [69]. The first part of the learning model was the standard autoassociative Hebbian learning rule, which is implemented as the sum of the outer products across all of the trials (cf. [69]):

$$W = \sum_{t=1}^{M} (a_t \otimes a_t) = \sum_{t=1}^{M} a_t a_t^T \tag{6}$$

where W is the weight matrix, at is the activity vector at time t (a column vector), $\otimes$ represents the outer product (here, using the outer function in numpy), and a^T indicates the transpose of a (i.e., a row vector). Note that following [68] and (pg. 166 from [102]) we set the connection between a unit and itself (i.e., the diagonal of the weight matrix) to be equal to 0. We then implemented a probabilistic learning rule by allowing the network to learn each feature (i.e., each connection between simulated neurons/each entry in the weight matrix) with probability γ and to not learn each feature with probability $1 - \gamma$, which we calculated separately on each trial [69]. We calculated γ using draws from a truncated normal distribution, which allows us to incorporate encoding variability into the encoding phase. Specifically, we generated the probabilities from a normal distribution with a mean of 0.4 and a standard deviation of 0.1 using the function numpy.random.normal. Whenever the draw from the random distribution returned a value that was less than 0 or greater than 1, we redrew from the normal distribution (i.e., this is the truncated aspect of the truncated normal distribution; [69]). Note that Rizzuto and Kahana (2001) modeled different versions of the probabilistic encoding rule; e.g., to allow for symmetric vs. asymmetric learning of associations between items in an associative memory task. However, in our case, we implemented a fully symmetric learning rule. Once we calculated γ for a given trial, we then used this value to generate the features to encode via the binomial function in numpy. Specifically, we generated a binomial

vector of values with a probability of success equal to γ, with a length equal to $(N \times (N-1))/2$; i.e., corresponding to the upper triangle of the change-in-weights matrix: Δ. We then placed these features in the upper triangle of the binary change-in-weights matrix (which was initialized as all zeros) and then we made the matrix symmetric with the following equation:

$$x = x + x^T \tag{7}$$

Altogether, the weight changes for the probabilistic Hebbian learning rule for each trial were given by the element-wise multiplication of the change-in-weights matrix with the outer product of the item vectors, which we then summed across all trials, as shown by following equation (cf. [69]):

$$W = \sum_{t=1}^{M} (\Delta_t \odot (a_t \otimes a_t)) \tag{8}$$

where Δ_t represents a binary change-in-weights matrix (on trial t) that implements the probabilistic learning rule described above (values of 1 for weights that will change and 0 otherwise; note that the t subscript indicates that the matrix representing the weights to change is trial unique), $\odot$ represents the Hadamard product (i.e., the element-wise product) of the two resultant matrices, a_t is the activity vector at time t, and $\otimes$ represents the outer product (here, using the outer function in numpy). As we mentioned above, we set the connection between a unit and itself (i.e., the diagonal of the weight matrix) to be equal to 0 ([68] and pg. 166 from [102]).

To model the test phase, we used two approaches, one based on a familiarity-based model (for a similar approach to modeling memory within the perirhinal cortex see Bogacz et al., 1999, 2001) and another based on a recall-based model (for a similar approach to modeling the hippocampus see Norman and O'Reilly, 2003). For the familiarity-based model, we calculated the harmony (H, i.e., the match) of the probe item, p, to the memory matrix, W, using the harmony function (cf. [68, 103]; see pg. 107 from [104]):

$$H = 0.5 p^T W p \tag{9}$$

where p represents the probe vector (a column vector), W represents the weight matrix, and p^T represents the transpose of the probe vector (i.e., a row vector). Specifically, we used the dot function within numpy for the calculation (0.5 * np.dot(np.dot(p, W), p))). Note that [105, 106] applied a similar approach to modeling familiarity-based discriminations in the perirhinal cortex (i.e., using the energy of the network, E, which is simply the negative value of the harmony of the network: $E = -H$).

For the recall-based model, we used Hopfield's (1982) dynamical rule for updating the trace. Specifically, for each iteration, we calculated the output of the network with the following equation ([68, 69] and pg. 170 from [102]):

$$s_{t+1}(i) = sgn\,(W s_t) \tag{10}$$

where sgn is the sign function (here, set to +1 if the activity of the unit is greater than 0, to -1 if the activity of the unit is less than 0, and randomly to +1 or -1 if the activity of the unit is equal to 0), W represents the weight matrix, and s_t represents the trace at time t. The initial value of s at the start of a test trial (i.e., s_0) was one of the probes (i.e., either a target, a similar lure, or a novel foil item). We used the dynamical retrieval rule, in which we updated the activity of a single unit per iteration (i.e., time step: t, denoted as $s_{t+1}(i)$, where the ith unit is updated; [68, 69]). Here, we randomly selected one of the units that did not have the same value as the previous time step. We allowed the network to iterate until it reached a match criterion, θ, to one of the items in the

training lexicon (here, θ = cosine similarity > 0.99) or for a maximum number of iterations ($Imax$ = 800; [69]). After each trial ended (i.e., either exceeding or reaching $Imax$), we then calculated the similarity between the probe and the retrieved pattern (cf. [16]) using the cosine similarity (i.e., values close to 1 indicate stronger match between the probe item and the recalled item, whereas values closer to 0 indicate a weaker match between the probe item and the recalled item).

For both the familiarity-based model and the recall-based model, we simulated forced-choice performance and old/new performance using the general methods described below (see 5.3.3).

5.3.3 Simulations of forced-choice performance

To simulate performance on the forced-choice test formats, we calculated the relevant memory scores for each model. Then, we simulated the decision on the forced-choice formats using the following equation [57] (cf. [8]):

$$Pr\{A > X\} = \frac{1}{M}\sum_{i=1}^{M}[I(g_{A_i} > g_{X_i}) + (0.5 \times I(g_{A_i} = g_{X_i}))] \tag{11}$$

where M represents the list length, $I()$ is the indicator function that returns the value 1 if the statement is true and 0 otherwise, g_{A_i} is the memory score for the target item on trial i, and g_{x_i} is the memory score for the distractor item (i.e., a novel foil or a similar lure) on trial i. Note that the second part of the equation deals with guessing by simulating that the model would get 0.5 of the guess trials correct (i.e., the model would randomly guess when the memory score for the target and distractor items match). We could have also simulated "tie-breaking" by randomly selecting an item on a trial-by-trial basis (cf. [16]); however, the equation above will produce the same results on average.

5.3.4 Simulations of old/new performance

To simulate performance on the old/new test formats, we calculated the relevant memory scores for each model. Then, we simulated the old/new test formats with confidence ratings by calculating the proportion of hits and false alarms at several values of the memory strength—i.e., we modeled the confidence ratings as increasing values of the memory scores. Specifically, here we used 10 confidence bins by linearly spacing from the minimum value of the target memory score (i.e., the lowest value that would induce any hits) to the maximum value of the distractor memory score (i.e., the highest value that would induce any false alarms). We then calculated the simulated hit rates and false alarm rates by calculating the number of responses within each simulated confidence bin: i.e., the number of values that are greater than or equal to the minimum cutoff and less than the cutoff of the next confidence bin. Note that for the weakest memory condition, we also added all of the distractor responses that were less than the minimum cutoff (i.e., the lowest memory score for the target distribution) and for the strongest memory condition, we added all of the target responses that were greater than the maximum cutoff (i.e., the highest memory score for the distractor distribution). Note, we calculated the minimum/maximum confidence bins in this manner (e.g., rather than allowing a nonzero hit rate with a zero false alarm rate) because the binormal model that we used for calculating performance assumes that the hit rate/false alarm rate plot (i.e., the receiving operator characteristic curve) passes through the origin: i.e., (0, 0). We then calculated the cumulative proportions of hits and false alarms based on these simulated confidence bins (from the "most confident old"/"least confident new" to the "least confident old"/"most confident new"). Then, we calculated our measure of performance, the area under the receiver

operating characteristic (ROC) curve (A_z), using the same techniques as we used for our empirical data (see 5.4.3 for more details and rationale of using this measure).

5.3.5 Quantification of target/lure similarity with an artificial neural network

We used an artificial neural network to quantify the similarity between our target and lure items. Briefly, previous research has focused on defining the similarity between targets and lures using a measure of mnemonic performance on the task [46, 62, 63]. Specifically, these studies have combined items into lure bins by quantifying the proportion of times that a participant makes an "old" response to a similar lure item, with a higher proportion of "old" responses indicating higher mnemonic similarity. These approaches are very helpful for quantifying and comparing the mnemonic similarity between stimuli. However, we think that it would also be interesting to study the relationship between a measure of stimulus similarity (i.e., a physical or semantic measure of similarity between targets and lures) to compare the transfer function between this objective measure of stimulus similarity and the degree of memory interference for the similar lures. Accordingly, we extracted features (i.e., modeled neural responses) from the 3rd convolutional layer of IT (i.e., the penultimate layer) from CORnet-S [60] using the THINGSvision toolbox [64]. Briefly, CORnet-S has been shown to provide a good approximation both to neural responses in IT as well as to behavioral data across several studies [60, 64]. Moreover, area IT is shown to provide the main inputs into perirhinal cortex [61], and thus the simulated activity in IT can be seen as an approximation of the input to the medial temporal lobe for object-memory tasks.

THINGSvision automatically resizes images to 256 pixels using bilinear interpolation and then performs a center crop to 224×224 pixels. Therefore, we changed our images to all be a square, while maintaining image proportions, by applying a white border (note all images in the Mnemonic Similarity Task are objects on a white background). Here, we set the white border to be 256 / 224 times the maximum dimension of the image (i.e., to avoid the model cropping part of the image). After extracting the simulated responses from IT, we quantified the similarity between the target-lure pairs by calculating Pearson's correlation coefficient between the model's response to the target item and the lure item (i.e., we calculated the pattern similarity; [107]). We then calculated Pearson's correlation coefficient between the resultant pattern similarity scores and 1-p("old") data from previous reports [46, 62, 63]. Here, we were specifically interested in comparing whether the transfer function between mnemonic discrimination and target/lure similarity is monotonic and relatively continuous (i.e., in line with the distributed memory models; MINERVA 2: Figure 4b, TODAM: Figure 5b, the familiarity-based Hopfield autoassociative neural network: Figure 6b, and the recall-based Hopfield autoassociative neural network: Figure 7b) or whether it is solely driven by only extremely high similarity values (i.e., target-lure similarity values that are greater than 0.875, which is what the CLS hippocampal model predicted; see Figure 2b; [16]).

5.4 Analytic methods

5.4.1 Forced-choice data

To analyze the data for the forced-choice data, we used generalized linear models with the lme4 package in R (version: 1.1-21; [108]). Specifically, we used the glmer function with a binomial family because our data were binomial (i.e., on each trial a participant either selected the correct answer or the incorrect answer). We fit our models with random intercepts for the participants. We assessed significance of the effects using a likelihood ratio test, using the anova function in R with the test set to "LRT". We also report the Bayesian information criterion (BIC), where lower

values indicate better model fits while penalizing for model complexity. To test whether there was a significant effect of test format, as in previous papers [57], we compared a model with the effect of test format to a random intercepts and slopes model:

Random-intercepts-and-slopes model:

$$\begin{aligned} null_model = glmer(Accuracy \sim 1 + (TestFormatPrev|Subject)\,, \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{12}$$

Test-format model:

$$\begin{aligned} full_model = glmer(Accuracy \sim TestFormat + (TestFormatPrev|Subject)\,, \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{13}$$

To test whether the empirical data fit better with the predictions of the distributed memory models (i.e., A-X > A-A' > A-B') than the CLS hippocampal model (i.e., A-X > A-A' = A-B'; a null effect of the A-A' and A-B' test formats), we compared the fit of a model with the test formats equal to the mean performance from Experiment 2 in [57] (A-X: logit(0.959) A-A': logit(0.836), A-B': logit(0.756)) vs. a model with equal performance on the A-A' and A-B' test (A-A' = A-B': mean(logit((0.836, 0.756)))). Specifically, we ran Equation (13) two times, once with the previous performance values and once with the CLS hippocampal model predictions (i.e., the $TestFormat$ values differed for each run of the model). To keep the models consistent in terms of the random effects, in both models the random effects structure included the random effects data from the previous performance values (i.e., termed $TestFormatPrev$ above). We then compared the Bayesian information criteria (BIC) between these two models (i.e., because zero parameters differed), by comparing both the raw BIC values (i.e., lower BIC indicated the preferred model) as well as the approximation of Bayes factors using the following equation [70, 71]:

$$BF_{10} \approx e^{(BIC_0 - BIC_1)/2} \tag{14}$$

To test whether there was a significant effect of lure bins, as in previous papers (e.g., for old/similar/new test format: [62]), we included a parameter for the lure bins in our model. Specifically, we compared the test-format model with a test-format-and-lure-bin model.

Random intercepts and slopes (for shortening equations below):

$$random_part = (TestFormat + LureBin||Subject) \tag{15}$$

We used a random effects model with random slopes and intercepts for the test format and the lure bin within each subject (and uncorrelated between the slopes and intercepts, hence the || in the equation above). We did not include the random effects for the interaction because the model produced a singular fit and failed to converge. Thus, we chose the maximal model that converged [109, 110]; moreover, as we discuss in the Results, the interaction model was not significant anyway, thus the random effects model in no way limits our conclusions here. We then compared several models to determine how well each model fit the data.

Random-intercepts-and-slopes model:

$$\begin{aligned} null_model = glmer(Accuracy \sim 1 + random_part, \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{16}$$

Test-format model (for comparison to Equation (18) with the A-A' and A-B' data):

$$\begin{aligned} test_format_model = glmer(Accuracy \sim TestFormat + random_part, \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{17}$$

Test-format-and-lure-bin model:

$$\begin{aligned} test_lure_model = glmer(Accuracy \sim TestFormat + LureBin + random_part, \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{18}$$

Here, we focused our analysis on the A-A' and A-B' test formats since the A-X test format did not manipulate the similarity to previously viewed items. We also investigated whether the effect of lure bins differed as a function of test format by comparing a test-format by lure-bin interaction model with the test-format-and-lure-bin model.

Test-format by lure-bin interaction model:

$$\begin{aligned} interaction_model = glmer(Accuracy \sim TestFormat * LureBin + random_part, \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{19}$$

Note, for these analyses, we normalized our lure bins (range: [1..5]) so that they center around 0 by subtracting 3 from each lure bin (range: [−2..2]). The lure bins are organized from the most similar (lowest lure bin) to the least similar (highest lure bin; see Figure 1), thus a positive beta estimate from these models would be indicative of a positive monotonic relationship between empirical mnemonic similarity and performance on the task.

We also followed up these analyses by comparing the following lure-bin model with an intercept-only version of the model (i.e., replacing $LureBin$ with 1 in Equation (20)), separately for each condition (i.e., the A-A' and A-B' test format data).

Lure-bin model (for testing within a single test format):

$$\begin{aligned} lure_bin_model = glmer(Outcome \sim LureBin + (1|Subject), \\ family = binomial, data = forced_choice_data, weights = n) \end{aligned} \tag{20}$$

5.4.2 Old/new data

To test whether there was a significant effect of lure bins, as in previous papers (e.g., for old/similar/new test format; [46]), we included a parameter for the lure bins in our model. First, we converted the participants' responses from "old" and "new" to binary responses, 0 if they selected "old" (i.e., the incorrect response) and 1 if they selected "new" (i.e., the correct response). Then we compared a lure-bin model (zero-meaned as in our forced-choice analysis; see above) to an intercept-only model using a binomial model as we did for our forced-choice data. Note that the more maximal model did not produce a singular fit here, thus our models were as follows:

Random-intercepts-and-slopes model for the old/new data:

$$\begin{aligned} null_model = glmer(proportion_new \sim 1 + (LureBin|Subject), \\ family = binomial, data = old_new_data, weights = n) \end{aligned} \tag{21}$$

Lure-bin model for the old/new data:

$$\begin{aligned} full_model = glmer(proportion_new \sim LureBin + (LureBin|Subject), \\ family = binomial, data = old_new_data, weights = n) \end{aligned} \tag{22}$$

5.4.3 Comparison of the forced-choice and old/new test formats

One of the key motivations of the current set of experiments was to test predictions from the global matching models (MINERVA 2, TODAM), the Hopfield/autoassociative neural network, the medial temporal lobe cortex model, and the hippocampal model about the relationships between performance on the old/new and forced-choice test formats. To allow us to compare the results of the old/new test format to the forced-choice test format, we calculated the area under the receiver operating characteristic (ROC) curve based on the old/new with confidence ratings responses for our various conditions. Previous theories and empirical research have suggested that the area under the ROC curve is mathematically equivalent to proportion correct on a two-alternative forced-choice test format [72, 73, 74]. Swets and Pickett [73] and Stanislaw and Todorov [74] argued that the best measure for calculating the area under the ROC curve is to use maximum likelihood estimation to fit the z-transformed ROC curve with a binormal model, a measure referred to as A_z. Given that the procedure to calculate A_z uses a binormal model, it does not assume that the target and distractor distributions have equal variance (i.e., they can have unequal variance).

We developed our own code for the binormal model within R. The binormal model involves estimating the intercept (a), slope (b), and estimates of latent cut points within the decision space based on the confidence ratings (the Z_k's, where the number of cut points is equal to the number of confidence bins - 1). We calculated the initial a (intercept) and b (slope) using the "unbiased" slope method described by Stanislaw and Todorov (1999). Specifically, we first z-transformed the hit rate and false-alarm rates using the qnorm function in R (a function often referred to as Φ^{-1} [the "inverse phi" function] in signal detection theory, which converts probabilities to z scores; [74]). Then, we fit two linear regression models: 1) a linear regression with the z-transformed hit rate as the independent variable and the z-transformed false-alarm rate as the dependent variable, 2) we fit another model with the z-transformed false-alarm rate as the independent variable and the z-transformed hit rate as the dependent variable. Then, we calculated the "unbiased" slope by combining the slopes from these two models [74]:

$$Slope^* = 0.5\left(Slope_1 + \frac{1}{Slope_2}\right) \tag{23}$$

We calculated the initial intercept using the following equation [74]:

$$Intercept^* = \bar{z}_H - (Slope^* \times \bar{z}_F) \tag{24}$$

where $\bar{z}_H$ is the mean of the z-transformed hit rates and $\bar{z}_F$ is the mean of the z-transformed false-alarm rates. We then calculated the initial break points (i.e., Z_k) by passing the true negative rates, the initial x_k values, (for the distractor responses; i.e., 1 - false-alarm rates) to the *qnorm* function in R a function often referred to as Φ^{-1} [the "inverse phi" function] in signal detection theory, which converts probabilities to z scores; [74]). We then used maximum likelihood estimation with a binormal model to optimize the estimates of a, b, and the Z_k values using the following steps. For the distractor distribution (i.e., the novel foils or the similar lures), we calculated the probability that a participant made response j, R_j, given stimulus 1, S_1, using the following equation [111]:

$$P\left(R_j|S_1\right) = F\left(Z_{k=j}\right) - F\left(Z_{k=j-1}\right) \tag{25}$$

Where $Z_k = (x_k - \mu_1)/\sigma_1$ (note: these values are then updated in the optimization function described below; see above for explanation of how we calculated the initial Z_k values), and F is the cumulative normal distribution (a function often referred to as the Φ ["phi"] function in signal

detection theory, which converts z scores into probabilities; [74]), which we calculated with the *pnorm* function in R.

For the target distribution, we calculated the probability that a participant made response j, R_j, given stimulus 2, S_2, using the following equation [111]:

$$P\left(R_j|S_2\right) = F\left(bZ_{k=j} - a\right) - F\left(bZ_{k=j-1} - a\right) \tag{26}$$

where $b = \sigma_1/\sigma_2$, and $a = \left(\mu_2 - \mu_1\right)/\sigma_2$. Then, we maximized the following log-likelihood function [111]:

$$logL = \sum_{i=1}^{2}\sum_{j=1}^{k} r_{ij} log\left(P\left(R_j|S_i\right)\right) \tag{27}$$

where r_{ij} represents the number of responses, R_j's, to stimulus i (i.e., either the distractor [$i = 1$] or the target [$i = 2$]), and we used the *optim* function in R with the method set to BFGS (i.e., the Broyden-Fletcher-Goldfarb-Shanno algorithm; [112, 113, 114, 115]). Note, the optim function minimizes a given input function, thus we passed the negative value of the log-likelihood function above to the optim function, which, in practice, maximizes the log-likelihood function. Maximizing the log-likelihood function simultaneously optimizes the intercept (a), slope (b), and the Z_k values. Note, if participants made zero responses at a given confidence level for both the distractor trials and the target trials, then we dropped that bin from the analysis above (i.e., if a participant only used 7 of the possible 8 bins, then k would equal 6 rather than the typical value of 7 [i.e., n - 1 = 8 - 1 = 7]). Additionally, if any bins resulted in a cumulative hit or false alarm rate of 0, we converted it to 0.5 / trials, or if any bins resulted in a cumulative hit or false alarm rate of 1, we converted it to (trials - 0.5) / trials [74, 116].

We then calculated the area under the ROC curve (Az) using the following equation:

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{28}$$

where a and b are calculated via the binormal model above and Φ is the standard normal cumulative distribution function (calculated with the *pnorm* function in R).

To compare the old/new data to the A-A' and A-B' test formats, we calculated A_z for the comparison between targets and similar lures within each participant. Similarly, to compare the old/new data to the A-X test format, we calculated A_z for the comparison between targets and novel foils within each participant.

We then compared the predictions of the distributed memory models by comparing performance between the various test formats. While the A_z measure is directly interpretable as the predicted proportion correct on a two-alternative forced-choice test format, we cannot use binomial models to analyze the data because these are not proportional data (i.e., we would not get integer counts of performance). Therefore, we instead applied the logit function to transform our data for analysis with other methods, as we describe below.

First, we aimed to test the distributed memory models' prediction that performance on the old/new test format with targets and similar lures would be the most similar to the A-B' test format. Accordingly, we ran two separate regressions between performance on the old/new test format with targets and similar lures vs. the A-A' and A-B' test formats (see Equations (23) and (24)). We then compared which of the two slopes was closer to 1. We next compared the raw performance

values between the test formats and we performed contrasts between the conditions to determine which conditions were the most similar. Additionally, we used a dendrogram analysis in which we calculated the Euclidean distance between performance on the various test formats. In accordance with the distributed memory models, we predicted that these analyses would reveal that the A-B' and the old/new test format with targets vs. similar lures would be the most similar, followed by the A-A' test format. Likewise, we predicted that the A-X and the old/new with targets and novel foils would be similar to each other.

We tested the prediction of the distributed memory models that performance would be correlated across all of the forced-choice and the old/new test formats. Specifically, although previous research suggests that there are dissociations in performance on these test formats, the distributed memory models predict that performance will be similar between all test formats (i.e., as a function of the learning rate of the model/participant). Therefore, positive correlations between the test formats would support the predictions of the distributed memory models while null correlations would support the notion that these test formats recruit distinct cognitive and neural processes.

References

[1] John P. Aggleton and Malcolm W. Brown. Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences*, 22(3):425–444, June 1999.

[2] Andrew P Yonelinas. The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, 46(3):441–517, April 2002.

[3] H. Eichenbaum, A.P. Yonelinas, and C. Ranganath. The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, 30(1):123–152, July 2007.

[4] Larry R. Squire, John T. Wixted, and Robert E. Clark. Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, 8(11):872–883, November 2007.

[5] John T. Wixted. Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, 114(1):152–176, 2007.

[6] Bennet B. Murdock. A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89(6):609–626, 1982.

[7] Bennet B. Murdock. Similarity in a Distributed Memory Model. *Journal of Mathematical Psychology*, 39(3):251–264, September 1995.

[8] Douglas L. Hintzman. Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95(4):528–551, 1988.

[9] T. Davis, G. Xue, B. C. Love, A. R. Preston, and R. A. Poldrack. Global Neural Pattern Similarity as a Common Basis for Categorization and Recognition Memory. *Journal of Neuroscience*, 34(22):7472–7484, May 2014.

[10] R. Quian Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, June 2005.

[11] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11):4302–4311, June 1997.

[12] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, April 1998.

[13] J. V. Haxby. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539):2425–2430, September 2001.

[14] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, July 1995.

[15] Randall C O'Reilly and Kenneth A Norman. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, 6(12):505–510, December 2002.

[16] Kenneth A. Norman and Randall C. O'Reilly. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4):611–646, 2003.

[17] Kenneth A. Norman. How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus*, 20(11):1217–1227, November 2010.

[18] David E. Rumelhart and Peter M. Todd. Learning and connectionist representations. In *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience.*, pages 3–30. The MIT Press, Cambridge, MA, US, 1993.

[19] James L. McClelland and Timothy T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, April 2003.

[20] J.S. Holdstock, A.R. Mayes, N. Roberts, E. Cezayirli, C.L. Isaac, R.C. O'Reilly, and K.A. Norman. Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus*, 12(3):341–351, 2002.

[21] P. J. Bayley, J. T. Wixted, R. O. Hopkins, and L. R. Squire. Yes/No Recognition, Forced-choice Recognition, and the Human Hippocampus. *Journal of Cognitive Neuroscience*, 20(3):505–512, March 2008.

[22] Annette Jeneson, C. Brock Kirwan, Ramona O. Hopkins, John T. Wixted, and Larry R. Squire. Recognition memory and the hippocampus: A test of the hippocampal contribution to recollection and familiarity. *Learning & Memory*, 17(1):63–70, January 2010.

[23] C. Brock Kirwan, Andrew Hartshorn, Shauna M. Stark, Naomi J. Goodrich-Hunsaker, Ramona O. Hopkins, and Craig E.L. Stark. Pattern separation deficits following damage to the hippocampus. *Neuropsychologia*, 50(10):2408–2414, August 2012.

[24] Stevenson Baker, Paula Vieweg, Fuqiang Gao, Asaf Gilboa, Thomas Wolbers, Sandra E. Black, and R. Shayna Rosenbaum. The Human Dentate Gyrus Plays a Necessary Role in Discriminating New Memories. *Current Biology*, 26(19):2629–2634, October 2016.

[25] John O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, November 1971.

[26] John O'Keefe and Lynn Nadel. *The hippocampus as a cognitive map.* Clarendon Press ; Oxford University Press, Oxford : New York, 1978.

[27] J. K. Leutgeb, S. Leutgeb, M.-B. Moser, and E. I. Moser. Pattern Separation in the Dentate Gyrus and CA3 of the Hippocampus. *Science*, 315(5814):961–966, February 2007.

[28] Joshua P. Neunuebel and James J. Knierim. CA3 Retrieves Coherent Representations from Degraded Input: Direct Evidence for CA3 Pattern Completion and Dentate Gyrus Pattern Separation. *Neuron*, 81(2):416–427, January 2014.

[29] S. Leutgeb and J. K. Leutgeb. Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. *Learning & Memory*, 14(11):745–757, November 2007.

[30] Arnold Bakker, C. Brock Kirwan, Michael Miller, and Craig E. L. Stark. Pattern Separation in the Human Hippocampal CA3 and Dentate Gyrus. *Science*, 319(5870):1640–1642, March 2008.

[31] Joyce W. Lacy, Michael A. Yassa, Shauna M. Stark, L. Tugan Muftuler, and Craig E.L. Stark. Distinct pattern separation related transfer functions in human CA3/dentate and CA1 revealed using high-resolution fMRI and variable mnemonic similarity. *Learning & Memory*, 18(1):15–18, January 2011.

[32] Michael A. Yassa and Craig E.L. Stark. Pattern separation in the hippocampus. *Trends in Neurosciences*, 34(10):515–525, October 2011.

[33] S. E. Motley and C. B. Kirwan. A Parametric Investigation of Pattern Separation Processes in the Medial Temporal Lobe. *Journal of Neuroscience*, 32(38):13076–13084, September 2012.

[34] Zachariah M. Reagh and Michael A. Yassa. Object and spatial mnemonic interference differentially engage lateral and medial entorhinal cortex in humans. *Proceedings of the National Academy of Sciences*, 111(40), October 2014.

[35] D. Berron, H. Schutze, A. Maass, A. Cardenas-Blanco, H. J. Kuijf, D. Kumaran, and E. Duzel. Strong Evidence for Pattern Separation in Human Dentate Gyrus. *Journal of Neuroscience*, 36(29):7569–7579, July 2016.

[36] Zachariah M. Reagh, Jessica A. Noche, Nicholas J. Tustison, Derek Delisle, Elizabeth A. Murray, and Michael A. Yassa. Functional Imbalance of Anterolateral Entorhinal Cortex and Hippocampal Dentate/CA3 Underlies Age-Related Object Pattern Separation Deficits. *Neuron*, 97(5):1187–1198.e4, March 2018.

[37] Michael A. Yassa, Joyce W. Lacy, Shauna M. Stark, Marilyn S. Albert, Michela Gallagher, and Craig E.L. Stark. Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus*, pages n/a–n/a, 2010.

[38] Michael A. Yassa, Aaron T. Mattfeld, Shauna M. Stark, and Craig E. L. Stark. Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proceedings of the National Academy of Sciences*, 108(21):8873–8878, May 2011.

[39] Michael A. Yassa, L. Tugan Muftuler, and Craig E. L. Stark. Ultrahigh-resolution microstructural diffusion tensor imaging reveals perforant path degradation in aged humans in vivo. *Proceedings of the National Academy of Sciences*, 107(28):12687–12691, July 2010.

[40] Ilana J. Bennett and Craig E.L. Stark. Mnemonic discrimination relates to perforant path integrity: An ultra-high resolution diffusion tensor imaging study. *Neurobiology of Learning and Memory*, 129:107–112, March 2016.

[41] Ilana J. Bennett, Derek J. Huffman, and Craig E.L. Stark. Limbic Tract Integrity Contributes to Pattern Separation Performance Across the Lifespan. *Cerebral Cortex*, 25(9):2988–2999, September 2015.

[42] Arnold Bakker, Gregory L. Krauss, Marilyn S. Albert, Caroline L. Speck, Lauren R. Jones, Craig E. Stark, Michael A. Yassa, Susan S. Bassett, Amy L. Shelton, and Michela Gallagher. Reduction of Hippocampal Hyperactivity Improves Cognition in Amnestic Mild Cognitive Impairment. *Neuron*, 74(3):467–474, May 2012.

[43] Arnold Bakker, Marilyn S. Albert, Gregory Krauss, Caroline L. Speck, and Michela Gallagher. Response of the medial temporal lobe network in amnestic mild cognitive impairment to therapeutic intervention assessed by fMRI and memory task performance. *NeuroImage: Clinical*, 7:688–698, 2015.

[44] Michael A. Yassa, Shauna M. Stark, Arnold Bakker, Marilyn S. Albert, Michela Gallagher, and Craig E.L. Stark. High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic Mild Cognitive Impairment. *NeuroImage*, 51(3):1242–1252, July 2010.

[45] Tammy T. Tran, Caroline L. Speck, Aparna Pisupati, Michela Gallagher, and Arnold Bakker. Increased hippocampal activation in ApoE-4 carriers and non-carriers with amnestic mild cognitive impairment. *NeuroImage: Clinical*, 13:237–245, 2017.

[46] Shauna M. Stark, C. Brock Kirwan, and Craig E.L. Stark. Mnemonic Similarity Task: A Tool for Assessing Hippocampal Integrity. *Trends in Cognitive Sciences*, 23(11):938–951, November 2019.

[47] John F. Guzowski, James J. Knierim, and Edvard I. Moser. Ensemble Dynamics of Hippocampal Regions CA3 and CA1. *Neuron*, 44(4):581–584, November 2004.

[48] Michael E. Hasselmo, Bradley P. Wyble, and Gene V. Wallenstein. Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus*, 6(6):693–708, 1996.

[49] M. Meeter, J.M.J. Murre, and L.M. Talamini. Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, 14(6):722–741, 2004.

[50] Dharshan Kumaran and Eleanor A Maguire. An Unexpected Sequence of Events: Mismatch Detection in the Human Hippocampus. *PLoS Biology*, 4(12):e424, November 2006.

[51] D. Kumaran and E. A. Maguire. Match Mismatch Processes Underlie Human Hippocampal Responses to Associative Novelty. *Journal of Neuroscience*, 27(32):8517–8524, August 2007.

[52] Katherine Duncan, Nicholas Ketz, Souheil J. Inati, and Lila Davachi. Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. *Hippocampus*, 22(3):389–398, March 2012.

[53] K. F. LaRocque, M. E. Smith, V. A. Carr, N. Witthoft, K. Grill-Spector, and A. D. Wagner. Global Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. *Journal of Neuroscience*, 33(13):5466–5474, March 2013.

[54] Endel Tulving. Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, 20(5):479–496, October 1981.

[55] Carmen E. Westerberg, Ken A. Paller, Sandra Weintraub, M.-Marsel Mesulam, Juliet S. Holdstock, Andrew R. Mayes, and Paul J. Reber. When memory does not fail: Familiarity-based recognition in mild cognitive impairment and Alzheimer's disease. *Neuropsychology*, 20(2):193–205, March 2006.

[56] Ellen Migo, Daniela Montaldi, Kenneth A. Norman, Joel Quamme, and Andrew Mayes. The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, 62(6):1198–1215, June 2009.

[57] Derek J. Huffman and Craig E. L. Stark. Age-related impairment on a forced-choice version of the Mnemonic Similarity Task. *Behavioral Neuroscience*, 131(1):55–67, February 2017.

[58] Leslie Rollins, Alexis Khuu, and Nafeesa Lodi. Encoding variability accounts for false recognition of noncorresponding lures on the forced-choice Mnemonic Similarity Task. *Learning & Memory*, 26(8):280–283, August 2019.

[59] Yana Fandakova, Elliott G Johnson, and Simona Ghetti. Distinct neural mechanisms underlie subjective and objective recollection and guide memory-based decision making. *eLife*, 10:e62520, March 2021.

[60] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs, October 2019. arXiv:1909.06161 [cs, eess, q-bio].

[61] Wendy L. Suzuki and David G. Amaral. Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Journal of Comparative Neurology*, 350(4):497–533, December 1994.

[62] Shauna M. Stark, Michael A. Yassa, Joyce W. Lacy, and Craig E.L. Stark. A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia*, 51(12):2442–2449, October 2013.

[63] Shauna M. Stark, Rebecca Stevenson, Claudia Wu, Samantha Rutledge, and Craig E. L. Stark. Stability of age-related deficits in the mnemonic similarity task across task variations. *Behavioral Neuroscience*, 129(3):257–268, June 2015.

[64] Lukas Muttenthaler and Martin N. Hebart. THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. *Frontiers in Neuroinformatics*, 15:679838, September 2021.

[65] C. Brock Kirwan and Craig E.L. Stark. Overcoming interference: An fMRI investigation of pattern separation in the medial temporal lobe. *Learning & Memory*, 14(9):625–633, September 2007.

[66] Douglas L. Hintzman. MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, &amp Computers*, 16(2):96–101, March 1984.

[67] E. T. Rolls. An attractor network in the hippocampus: Theory and neurophysiology. *Learning & Memory*, 14(11):714–731, November 2007.

[68] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982.

[69] Daniel S. Rizzuto and Michael J. Kahana. An Autoassociative Neural Network Model of Paired-Associate Learning. *Neural Computation*, 13(9):2075–2092, September 2001.

[70] Eric-Jan Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804, October 2007.

[71] Andrew F. Jarosz and Jennifer Wiley. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, 7(1), November 2014.

[72] David M. Green and Franklin L. Moses. On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, 66(3):228–234, 1966.

[73] John A. Swets and Ronald M. Pickett. *Evaluation of diagnostic systems: methods from signal detection theory*. Academic Press series in cognition and perception. Academic Press, New York, 1982.

[74] Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, March 1999.

[75] Charan Ranganath and Maureen Ritchey. Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*, 13(10):713–726, October 2012.

[76] Larry R. Squire, Craig E.L. Stark, and Robert E. Clark. The medial temporal lobe. *Annual Review of Neuroscience*, 27(1):279–306, July 2004.

[77] E. A. Buffalo, P. S.F. Bellgowan, and A. Martin. Distinct roles for medial temporal lobe structures in memory for objects and their locations. *Learning & Memory*, 13(5):638–643, September 2006.

[78] Edmund T. Rolls. The storage and recall of memories in the hippocampo-cortical system. *Cell and Tissue Research*, 373(3):577–604, September 2018.

[79] Randall C. O'Reilly and James L. McClelland. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, 4(6):661–682, December 1994.

[80] David Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 262(841):23–81, July 1971.

[81] B.L. McNaughton and R.G.M. Morris. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10(10):408–415, January 1987.

[82] Iain A. Wilson, Michela Gallagher, Howard Eichenbaum, and Heikki Tanila. Neurocognitive aging: prior memories hinder new hippocampal encoding. *Trends in Neurosciences*, 29(12):662–670, December 2006.

[83] Iva K. Brunec, Morris Moscovitch, and Morgan D. Barense. Boundaries Shape Cognitive Representations of Spaces and Events. *Trends in Cognitive Sciences*, 22(7):637–650, July 2018.

[84] Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman. Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, 95(3):709–721.e5, August 2017.

[85] Zachariah M. Reagh, Angelique I. Delarazan, Alexander Garber, and Charan Ranganath. Aging alters neural activity at event boundaries in the hippocampus and Posterior Medial network. *Nature Communications*, 11(1):3980, August 2020.

[86] Zachariah M. Reagh and Charan Ranganath. Flexible reuse of cortico-hippocampal representations during encoding and recall of naturalistic events. *Nature Communications*, 14(1):1279, March 2023.

[87] James W. Tanaka and Tim Curran. A Neural Basis for Expert Object Recognition. *Psychological Science*, 12(1):43–47, January 2001.

[88] Isabel Gauthier and Michael J. Tarr. Becoming a "Greeble" Expert: Exploring Mechanisms for Face Recognition. *Vision Research*, 37(12):1673–1682, June 1997.

[89] Isabel Gauthier, Pepper Williams, Michael J. Tarr, and James Tanaka. Training 'greeble' experts: a framework for studying expert object recognition processes. *Vision Research*, 38(15-16):2401–2428, August 1998.

[90] Isabel Gauthier, Michael J. Tarr, Adam W. Anderson, Pawel Skudlarski, and John C. Gore. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2(6):568–573, June 1999.

[91] Isabel Gauthier and Nikos K. Logothetis. IS FACE RECOGNITION NOT SO UNIQUE AFTER ALL? *Cognitive Neuropsychology*, 17(1-3):125–142, February 2000.

[92] Isabel Gauthier, Pawel Skudlarski, John C. Gore, and Adam W. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2):191–197, February 2000.

[93] Michael L. Mack, Bradley C. Love, and Alison R. Preston. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, 113(46):13203–13208, November 2016.

[94] Colin T Kyle, Jared D Stokes, Jennifer S Lieberman, Abdul S Hassan, and Arne D Ekstrom. Successful retrieval of competing spatial environments in humans involves hippocampal pattern separation mechanisms. *eLife*, 4:e10499, November 2015.

[95] Jared Stokes, Colin Kyle, and Arne D. Ekstrom. Complementary Roles of Human Hippocampal Subfields in Differentiation and Integration of Spatial Context. *Journal of Cognitive Neuroscience*, 27(3):546–559, March 2015.

[96] Jared M. Roberts, Maria Ly, Elizabeth Murray, and Michael A. Yassa. Temporal discrimination deficits as a function of lag interference in older adults: Temporal Discrimination and Interference in Aging. *Hippocampus*, 24(10):1189–1196, October 2014.

[97] Zachariah M. Reagh, Huy D. Ho, Stephanie L. Leal, Jessica A. Noche, Amanda Chun, Elizabeth A. Murray, and Michael A. Yassa. Greater loss of object than spatial mnemonic discrimination in aged adults: selective object memory deficits in aging. *Hippocampus*, 26(4):417–422, April 2016.

[98] Maria E Montchal, Zachariah M Reagh, and Michael A Yassa. Precise temporal memories are supported by the lateral entorhinal cortex in humans. *Nature Neuroscience*, 22(2):284–288, February 2019.

[99] Rita E. Loiotile and Susan M. Courtney. A signal detection theory analysis of behavioral pattern separation paradigms. *Learning & Memory*, 22(8):364–369, August 2015.

[100] Daniela J. Palombo, Lynne J. Williams, Hervé Abdi, and Brian Levine. The survey of autobiographical memory (SAM): A novel measure of trait mnemonics in everyday life. *Cortex*, 49(6):1526–1540, June 2013.

[101] E. Hoddes, V. Zarcone, H. Smythe, R. Phillips, and W. C. Dement. Quantification of Sleepiness: A New Approach. *Psychophysiology*, 10(4):431–436, July 1973.

[102] Michael J. Kahana. *Foundations of human memory*. Oxford University Press, New York, 2012. OCLC: ocn744297060.

[103] Paul Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing. Volume 1: Foundations*, pages 282–317. MIT Press, Cambridge, MA, US, 1986.

[104] Randall C. O'Reilly and Yuko Munakata. *Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain*. MIT Press, Cambridge, Mass, 2000.

[105] R. Bogacz. High capacity neural networks far familiarity discrimination. In *9th International Conference on Artificial Neural Networks: ICANN '99*, volume 1999, pages 773–778, Edinburgh, UK, 1999. IEE.

[106] Rafal Bogacz, Malcolm W. Brown, and Christophe Giraud-Carrier. Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, 10(1):5–23, 2001.

[107] Nikolaus Kriegeskorte. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008.

[108] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1), 2015.

[109] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, April 2013.

[110] Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. Parsimonious Mixed Models. 2015. Publisher: arXiv Version Number: 2.

[111] Donald D. Dorfman and Edward Alf. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology*, 6(3):487–496, October 1969.

[112] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[113] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, March 1970.

[114] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[115] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[116] Donald D. Dorfman. RSCORE II. In John A. Swets and Ronald M. Pickett, editors, *Evaluation of diagnostic systems: Methods from signal detection theory.*, pages 212–232. Academic Press, New York, 1982.