A neural network model of hippocampal contributions to category learning

Jelena Sučević*[1] & Anna C. Schapiro[2]

[1]Department of Experimental Psychology, University of Oxford
[2]Department of Psychology, University of Pennsylvania

*Author for correspondence: jelena.sucevic@psy.ox.ac.uk

Abstract

In addition to its critical role in encoding individual episodes, the hippocampus is capable of extracting regularities across experiences. This ability is central to category learning, and a growing literature indicates that the hippocampus indeed makes important contributions to this kind of learning. Using a neural network model that mirrors the anatomy of the hippocampus, we investigated the mechanisms by which the hippocampus may support novel category learning. We simulated three category learning paradigms and evaluated the network's ability to categorize and to recognize specific exemplars in each. We found that the trisynaptic pathway within the hippocampus—connecting entorhinal cortex to dentate gyrus, CA3, and CA1—was critical for remembering individual exemplars, reflecting the rapid binding and pattern separation functions of this circuit. The monosynaptic pathway from entorhinal cortex to CA1, in contrast, was responsible for detecting the regularities that define category structure, made possible by the use of distributed representations and a slower learning rate. Together, the simulations provide an account of how the hippocampus and its constituent pathways support novel category learning.

Introduction

Learning how the entities in our environment cluster into groups with overlapping properties, names, and consequences allows us to communicate and act adaptively. This category learning often unfolds over long periods of time, for example when learning about the many species of dogs across development, but can also occur within a few minutes or hours, as when learning about different kinds of penguins on a first visit to the zoo. Much is known about how neocortical areas represent categories of information learned over long time scales (Martin, 2007; Miller, Freedman, & Wallis, 2002; Miller, Nieder, Freedman, & Wallis, 2003; Spiridon & Kanwisher, 2002), but less is understood about the mechanisms by which the brain learns quickly in initial encounters. Given the ability of the hippocampus to learn rapidly (McClelland, McNaughton, & O'Reilly, 1995) combined with its ability to learn regularities (Schapiro, Kustner, & Turk-Browne, 2012), this brain area seems well-suited to make a contribution to rapid category learning. Indeed, neuroimaging studies provide strong evidence that the hippocampus is engaged in novel category learning (Bowman & Zeithamova, 2018; Mack, Love, & Preston, 2016; Zeithamova, Maddox, & Schnyer, 2008). Studies with hippocampal amnesics tend to find partial but not complete deficits in category learning (Knowlton & Squire, 1993; Kolodny, 1994; Reber, Knowlton, & Squire, 1996; Reed, Squire, Patalano, Smith, & Jonides, 1999), indicating that the hippocampus—though not the sole region involved—does make an important causal contribution.

In the present work, we ask what computational properties of the hippocampus might allow it to contribute to category learning. Using a neural network model of the hippocampus named C-HORSE (Complementary Hippocampal Operations for Representing Statistics and Episodes), we previously demonstrated how the hippocampus might contribute to learning temporal regularities embedded in sequences of stimuli and to inference over pairwise associations (Schapiro, Turk-Browne, et al., 2017; Zhou, Singh, Tandoc, & Schapiro, 2021). We showed that the heterogeneous properties of the two main pathways within the hippocampus may support complementary learning systems—a microcosm of hippocampus-neocortex relationship (McClelland, McNaughton, & O'Reilly, 1995), with one pathway specializing in the rapid encoding of individual episodes and another in extracting regularities over time. The present work evaluates whether this ability to extract statistical regularities may also support learning the structure of novel categories.

C-HORSE comes from a lineage of models developed to account for episodic memory phenomena (Ketz, Morkonda, & O'Reilly, 2013; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). It instantiates the broad anatomical structure of the hippocampus: hippocampal subfields dentate gyrus (DG), cornu ammonis (CA3) and CA1 are represented as three hidden layers; they receive input and process output through entorhinal cortex (EC; Figure 1). The subfields are connected via two main pathways: the trisynaptic pathway (TSP) and the monosynaptic pathway (MSP). The TSP runs from EC to DG, CA3, and then CA1. The projections within the TSP are sparse, enabling the formation of orthogonalized representations even with highly similar input patterns (i.e., pattern separation). It has a high learning rate, which supports rapid, even one-shot learning. The TSP also contributes to the process of retrieving previously encoded patterns from partial cues (i.e., pattern completion) via recurrent connections in CA3. The TSP is thus critical for carrying out the episodic memory function of the hippocampus.

The MSP connects EC directly to CA1. These projections do not have the specialized sparsity of those in the TSP, allowing for more overlapping representations to emerge. In addition, the MSP seems to learn more slowly (Lee, Rao, & Knierim, 2004; Nakashiba, Young, McHugh, Buhl, & Tonegawa, 2008). These properties of relatively more overlapping representations and more incremental learning mirror those of neocortex (McClelland et al., 1995). In earlier versions of this model (Norman & O'Reilly, 2003), the MSP was seen as merely a translator between the TSP representations and neocortex, but we have argued that its properties may make the MSP well-suited to learning structured information across episodes (Schapiro, Turk-Browne, et al., 2017).

To investigate the role of the hippocampus in category learning, we tested how the MSP and TSP of C-HORSE contribute to forming category representations across three different types of categories. First, we evaluated the network's ability to learn simple nonoverlapping categories of exemplars consisting of multiple discrete features, with some features shared among the members of a category and others unique to each exemplar (Schapiro, McDevitt, et al., 2017; Schapiro, McDevitt, Rogers, Mednick, & Norman, 2018). We assessed the model's memory for these different kinds of features as well as its ability to generalize to novel exemplars. Second, we simulated the probabilistic Weather Prediction Task (Djonlagic et al., 2009; Eldridge, Masterman, & Knowlton, 2002; Knowlton, Mangels, & Squire, 1996; Knowlton, Squire, & Gluck, 1994; Reber et al., 1996). In this task, four different cards with shapes are each probabilistically associated with one of two categories: On each trial, a prediction about the weather (sun or rain) is made based on a combination of one, two, or three presented cards. We assessed the model's categorization ability as well as recognition of particular card combinations. Third, we tested the network's ability to learn categories with varying typicality defined along a continuum of overlapping features (Zeithamova et al., 2008). Prototypes of two categories have no features in common, and category exemplars then fall on a continuum between the two prototypes. Exemplars that share more features with the prototype are more typical category members. We assessed the model's categorization and recognition, as a function of typicality.

Across the three category learning tasks, C-HORSE was able to both determine the category membership of exemplars and recognize individual studied exemplars. There was a division of labor across the two pathways of the hippocampus in these functions: The MSP was critical for learning the regularities underlying category structure and was responsible for generalization of knowledge to novel exemplars. The TSP also contributed to behavior across the tasks, but only to the extent that memorizing unique properties of exemplars was useful. The rapid binding and pattern separation abilities of the TSP that make the pathway well-suited to episodic memory are also advantageous for encoding arbitrary relationships in category learning. The findings together motivate a theory of hippocampal contributions to category learning, with the MSP responsible for true understanding of category structure and the TSP for encoding the specifics of individual exemplars.

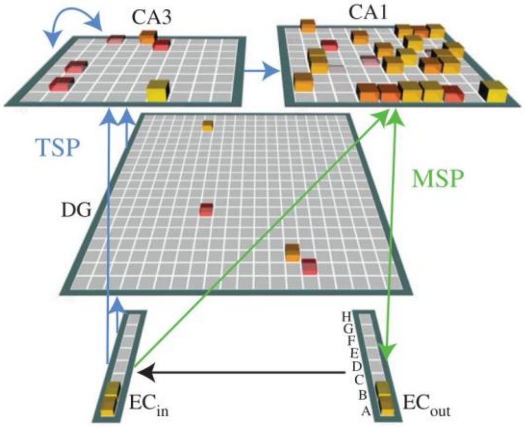

Figure 1. C-HORSE architecture: The model consists of dentate gyrus (DG), CA3 and CA1 subfields which map inputs from superficial to deep layers of the entorhinal cortex (EC). The trisynaptic pathway (TSP) connects EC to CA1 via DG and CA3 (blue arrows), and the monosynaptic pathway (MSP) connects EC directly with CA1 (green arrows).

Methods

We adopted a neural network model of the hippocampus developed after a lineage of models used to explain episodic memory phenomena (Ketz et al., 2013; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). This variant, C-HORSE, was developed recently to account for the role of the hippocampus in statistical learning (Schapiro, Turk-Browne, et al., 2017; Zhou et al., 2021). Simulations were performed in the Emergent simulation environment (version 7.0.1, Aisa, Mingus, & O'Reilly, 2008). Files for running the model can be found at github.com/schapirolab/hip-cat. The following section provides a brief description of the model; for more information on the model see Schapiro et al. (2017), and for full parameters and stimulus specifics for these simulations see the Supplementary Material.

Model architecture. The model has three hidden layers, representing DG, CA3, and CA1 hippocampal subfields, which learn to map input from superficial to deep layers of entorhinal cortex (EC$_{in}$ and EC$_{out}$; Figure 1). There is also a separate Input layer (not shown in Figure 1) with the same dimensionality as EC$_{in}$, where external input was clamped, allowing activity in EC$_{in}$ to vary as a function of external input as well as EC$_{out}$ activity. There were one-to-one connections between Input and EC$_{in}$ and between EC$_{in}$ and EC$_{out}$. Each layer contains units (400 in DG, 80 in CA3, 100 in CA1, and other layer sizes varying as a function of the task) with activity levels ranging from 0 to 1, implementing a rate code. A unit's activity is proportional to the activity of all units connected to it, weighed by connection weights between them. Unit activity is also modulated by inhibition between units within a layer. The inhibition is implemented using a set-point inhibitory current with k-winner-take-all dynamics, and simulates the action of inhibitory interneurons (O'Reilly, Munakata, Frank, Hazy, & Contributors, 2014).

The trisynaptic pathway (TSP) connects EC to CA1 via DG and CA3. Connections are sparse, reflecting known physiological properties of the hippocampus: DG and CA3 units receive input from 25% of units in the EC$_{in}$ layer, and CA3 receives directly from 5% of DG. Both DG and CA3 have high levels of within-layer inhibition. CA3 also has a full recurrent projection (every unit connected to every other). Finally, CA3 is fully connected to CA1.

The monosynaptic pathway (MSP) is formed by a direct connection from EC$_{in}$ to CA1, and bidirectional connections between CA1 and EC_{out}. CA1 has lower inhibition than CA3 and DG, allowing a higher proportion of units in the layer to be simultaneously active.

We ran 100 networks for each simulation. Each network had a randomized configuration of the sparse projections in the TSP and randomly initialized weights throughout the network. All analyses were conducted within each network and the results averaged across the networks.

Learning. The model was trained as an autoencoder, adjusting connection weights to reproduce patterns presented to EC_{in} on EC_{out}. Weights were updated via Contrastive Hebbian Learning, with two 'minus' phases each contrasted to a 'plus' phase on every training trial (Ketz et al., 2013). One of the minus phases simulates the trough of the hippocampal theta oscillation, when EC has a strong influence on CA1, and the connection from CA3 to CA1 is inhibited. The second minus phase simulates the theta peak, when CA3 has a stronger influence on CA1, and connections from EC_{in} to CA1 are inhibited. During the plus phase, the target output is directly clamped on EC_{out}. Weights are adjusted after each trial to reduce the local differences in unit coactivities between each of the two minus phases and the plus phase. The learning rate on the TSP was set to be 10 times higher than the MSP (Ketz et al., 2013; Schapiro, Turk-Browne, et al., 2017), except in the third simulation, where the MSP learning rate was even smaller to accommodate stimuli with high degrees of overlap (which can lead to degenerate learning at relatively higher learning rates).

Simulations had a fixed number of training trials except in the Weather Prediction Task, where we used a stopping rule: after a minimum of 25 training trials, the model had to achieve five consecutive trials with sum squared error below 1.2. The stopping rule was introduced because of the probabilistic nature of the categories, where it is not possible to eliminate all error.

Testing. Connection weights were not changed during test. Networks were tested before any training (epoch 0) and after every training epoch. For each set of simulations, we assessed the model's categorization ability and its ability to remember item-specific information.

Lesions. We simulated lesions of the MSP and TSP in order to assess the contributions of each pathway to performance. The MSP lesion was performed by setting the strength of the projection from EC_{in} to CA1 to 0. We did not lesion connections between CA1 and EC_{out} because they are necessary for producing output. The TSP lesion was performed by setting weights from CA3 to CA1 to 0. The lesions were implemented in a version of the model that did not use the theta-inspired learning scheme described above, as only one minus phase is appropriate with only one pathway intact. Lesions were implemented during both learning and testing.

Statistical analysis. To compare performance in the intact and lesioned networks, the mean accuracy was submitted to an ANOVA with a between-network factor *Condition* (intact, MSP-only, and TSP-only network), a within-network factor *Trial* (number of training trials prior to test) and *Network initialization* as a random effects factor (100 random initializations). Following the omnibus ANOVA, to determine which conditions may differ, we ran three separate ANOVAs with a between-network factor *Condition* which included two out of three conditions (intact vs MSP-only, intact vs TSP-only, and MSP-only vs TSP-only). Data visualization and statistical analyses were performed in R, version 3.6.1 (R Core Team, 2019).

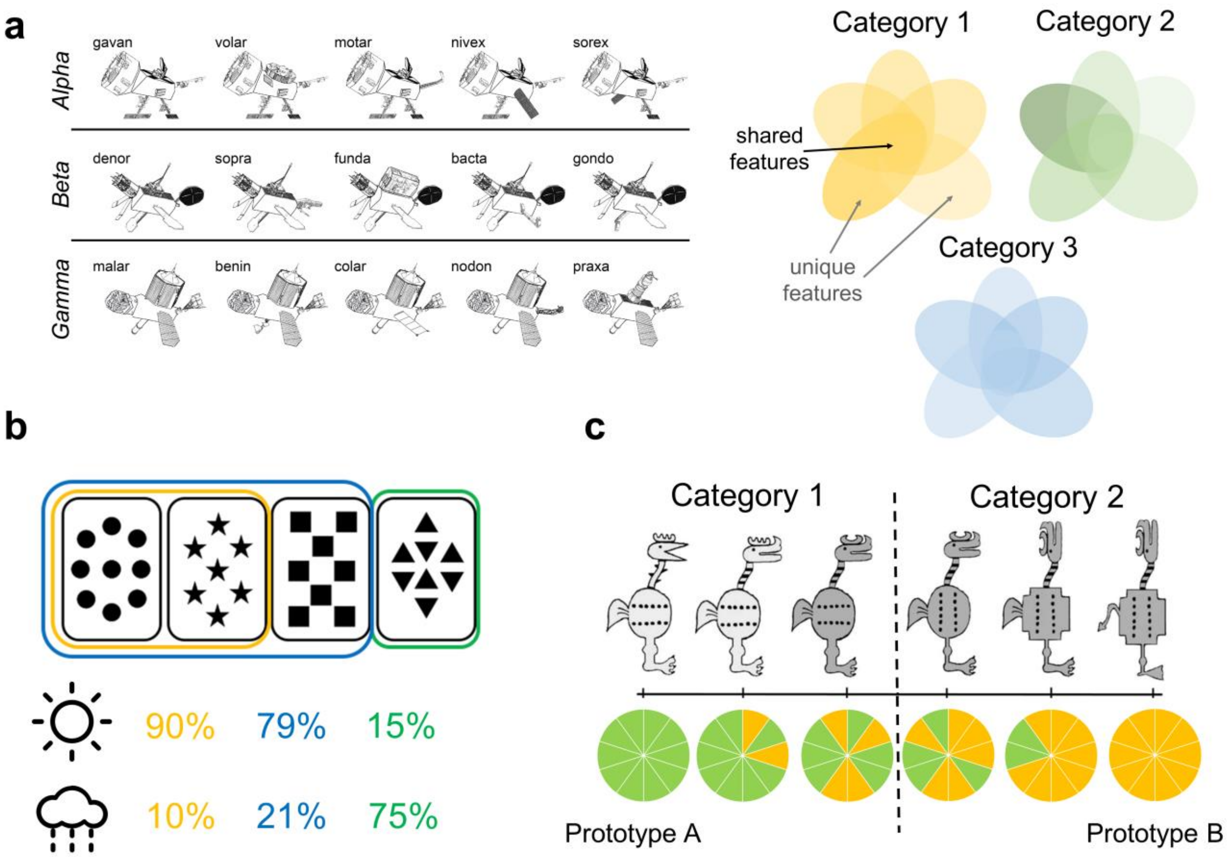

Figure 2. An overview of categories simulated. (a) Satellite categories: Distinct categories of novel "satellites" consisting of unique and shared features (Schapiro, Turk-Browne, et al., 2017). (b) Weather Prediction Task: each abstract card is probabilistically related to a category (sun or rain), and on a given trial, category must be guessed from a combination of cards (Knowlton et al., 1994). (c) Intermixed categories with varying typicality: Categories where each item consists of 10 binary features. The two prototypes on opposite sides of the feature space have no features in common, and the rest of the exemplars have a varying number of features in common with the prototypes (Zeithamova et al., 2008).

Results

Learning distinct categories of items with unique and shared features

First, we examined C-HORSE's ability to learn categories of items that consist of multiple discrete features, with some features unique to individual items and others shared amongst members of the same category, and no features overlapping across categories. To test the network's ability to learn these categories, we presented a set of novel objects representing three categories of "satellites," with five satellites in each category, following empirical work with this paradigm (Schapiro, McDevitt, et al., 2017; Schapiro et al., 2018). The model and humans were given a comparable number of training trials: 140 for the model and on average 122 for humans (Schapiro et al., 2018).

Each category had a prototype, defining the shared features for that category. Four other exemplars of the category had one out of five features swapped away from the prototype, such that they had one unique feature and four shared features (Figure 2a). This structure means that one of the shared features is present across all exemplars in a category, which effectively serves as a category name/indicator and will be used to assess categorization ability. Each feature was assigned to one unit in the input layer. If the feature was present, the input unit representing the feature took on a value of 1, and otherwise 0. Thus, there were 27 input units in total, 9 per category. Within each category, there were 5 units representing shared features and 4 units for unique features (Supplementary Table 1).

To characterize the network's behavior, we investigated its ability to recognize unique features of individual trained satellites (unique feature recognition), recognize the prototypical feature shared across all trained members of a category (categorization), and fill in shared features for novel satellites not presented during training (generalization). For unique feature recognition, we presented the network with the unique feature of a trained satellite as input and evaluated the network's ability to activate that feature on the output, compared to unique features of other members of the same category (Supplementary Table 2). Accuracy was determined by dividing the activation of the correct unit by the total activation in the four units representing unique features for that category (chance accuracy expected to be around 0.25). As shown in Figure 3a, the network learned to recognize unique features of individual satellites. With 140 total training trials, the intact model reached an accuracy of .51, similar to the levels observed empirically in humans after one learning session (.53; Schapiro, McDevitt, et al., 2017). A version of the network with access only to the MSP was completely unable to output the correct unique features, whereas a version with only the TSP could do this well above a chance, and even slightly better than the intact model (accuracy of .55). This reveals that the TSP is responsible for the network's ability to remember unique features. Differences between model types across time were all highly reliable (all $p_s < .001$).

To test categorization ability, we examined the network's ability to indicate the correct category of a satellite, operationalized as activating the category-prototypical feature shared across all members of the satellite's category. We presented the network with the unique feature of a satellite and divided output activity for the correct prototypical feature by the sum of activation of the three prototypical features for the three categories (with chance expected to be around 0.33). Across 140 trials, the intact network reached categorization accuracy of .79 (Figure 3b). We re-analyzed the published human data (Schapiro, McDevitt, et al., 2017) to calculate the analogous measure of accuracy and found a comparable accuracy level of 0.68. The MSP-only network exhibited much better performance than the intact network (1 at the end of training). The TSP-only network had poorer performance (.71), but still well above chance. Because this test involved trained satellites, categorization could be solved using either a memorization strategy or an extraction of regularities, leading to relatively good performance even for the TSP-only network. Because the MSP is unable to remember unique features (Figure 3a), it expresses knowledge only of the prototypical features, leading to excellent categorization performance. The intact network combines information from both sources, resulting in intermediate performance (see below for discussion of the idea that a control mechanism might allow selective enhancement of pathways depending on task). All differences between model types were again highly reliable (all $p_s < .001$).

The strongest test of category understanding is the ability to generalize to novel instances. To test generalization, we presented the network with a set of 18 satellites (6 per category) that were not presented during training (Supplementary Table 3). Each input satellite consisted of two shared features (not including the category-prototypical feature) and two unique features, and we tested the network's ability to output the category-prototypical feature. A similar pattern was observed as with the categorization of familiar items, but the TSP-only network showed poorer performance (.53) in comparison to the intact network (.94) while the MSP-only network was even better (Figure 3c). The MSP was able to ignore the unique features of these novel satellites, resulting in perfect generalization behavior relatively early in training (Figure 3c). All differences were reliable ($p_s < .001$).

To assess network representations, we performed representational similarity analysis for each hidden layer of the network. We used Pearson correlation to relate the patterns of unit activities evoked by presentation of each satellite's unique feature (for the 12 satellites with a unique features). There was no structure in the representations prior to training, and the representations that emerged with training revealed sensitivity to the category structure. This was particularly evident in CA1 (Figure 3d), with items from the same category represented much more similarly than items from different categories. This result is consistent with our recent neuroimaging findings using this paradigm, where CA1 was the only subfield of the hippocampus to show significant within versus between category multivoxel pattern similarity (Schapiro et al., 2018). There was different representational similarity in the initial response, after 45 cycles of processing, versus the settled response, after there was time for recurrent activity to spread throughout the network. In the initial response, there was no sensitivity at all to category structure in DG and CA3—items were represented orthogonally. CA1, in contrast, was immediately sensitive to the category structure. The settled response revealed sensitivity to category structure in all three hidden layers, as the structure in CA1 had time to influence the rest of the network via the EC_{out} to EC_{in} "big loop" connection (Kumaran & Maguire, 2007; Schapiro, Turk-Browne, et al., 2017). All sensitivity to the category structure in this network was thus driven by the learned representations in CA1.

In sum, these results suggest that the network is capable of learning categories and generalizing to novel instances. To achieve this, the MSP and the TSP take on complementary roles: the MSP extracts regularities and learns information that defines category structure, while the TSP encodes individual exemplars and handles unique feature recognition. These properties map directly to our prior simulations, where we found that the MSP detects statistical structure while the TSP encodes episode-unique information (Schapiro, Turk-Browne, et al., 2017). There are two key properties that differ between the pathways that lead to these results: 1) slower learning in the MSP than TSP, which allows integration of information overall longer periods of time in the MSP and quick learning in the TSP, and 2) more overlapping representations in the MSP than TSP, which helps the MSP see commonalities across experiences and helps the TSP separate experiences to avoid interference.

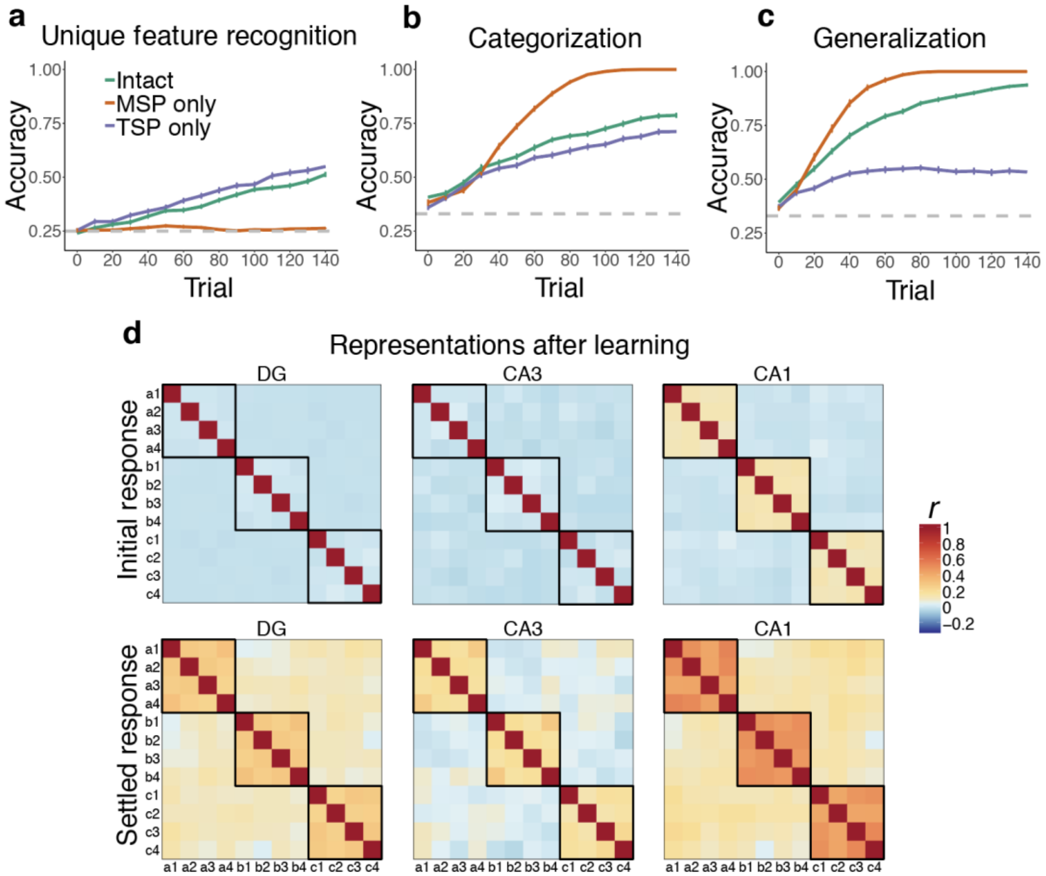

Figure 3. Satellite task. Performance of the intact network, a version of the network with only the MSP, and a version with only the TSP: (a) unique feature recognition, (b) categorization of trained items, and (c) categorization of novel items (generalization). (d) Representational similarity for the initial and settled response. Each item appears in the rows and columns of the heatmaps. The diagonals are always 1, as this reflects items correlated to themselves, and the off-diagonals are symmetric. Black boxes delineate categories. All plots represent mean performance averaged across random network initializations. Error bars denote ±1 s.e.m. across network initializations (some are too small to be visible).

Learning probabilistic categories

While the previous set of simulations focused on deterministic categories, i.e. categories in which each item could belong to only one category, in this section we test the network's ability to learn probabilistic categories. The Weather Prediction Task is a canonical probabilistic category learning task (Djonlagic et al., 2009; Eldridge et al., 2002; Knowlton et al., 1996, 1994; Reber et al., 1996). In this task, there are a total of four cards with abstract shapes, and a combination of one, two, or three cards is presented on each trial (Figure 2b). The combination of cards predicts a weather outcome, sunshine or rain. To learn these weather outcome categories, the network needs to keep track of the probability of each card being associated with each category and combine information about the probability of the cards presented together.

As in Knowlton et al. (1994), all 14 possible card combinations were presented. The number of times a particular combination of cards was presented and frequency of its association to each category was identical to the experimental procedure used in Knowlton et al. (1994; Supplementary Table 4). Two combinations of cards that had an equal probability of being associated with each category were removed from analysis. Each card was represented by one unit in the input and output, and each weather outcome (category) was represented by two units. As in prior simulations, the model was trained as an autoencoder, meaning that both the cards and category information were presented as input and the model was asked to reconstruct all of these features on the output layer. This training regimen is more akin to an "observational" than "feedback" mode of the task, which is appropriate given evidence that the medial temporal lobe is more engaged by observational variants (Poldrack et al., 2001; Shohamy et al., 2004). We trained the network for 50 trials, simulating Task 2 in Knowlton et al. (1994), where patients also saw 50 trials.

To examine the network's performance, we tested its ability to reconstruct individual combinations of cards (recognition) and to predict category based on the presented cards (categorization). For recognition performance, we evaluated reconstruction of the correct card output units given a set of input cards. Since all possible card combinations are presented during training, this does not involve discriminating old from new combinations, but is rather a simple measure of the network's ability to process information about each distinct card configuration. Recognition score was calculated by dividing the mean activation of correct card units by the mean activation across all card units. Given the stopping criteria used during training, networks were trained for different numbers of trials, with networks performing better stopping earlier. We ran as many networks as needed to obtain data from 100 networks in each of the three lesion conditions at trial 50. As a result, there were 739 networks at trials 0, 10 and 20, 646 networks at trial 30, 451 networks at trial 40, and 300 networks in the final test trial (100 per condition). The results indicated that the network was able to recognize individual combinations of cards (

Figure 4a). An ANOVA revealed significant main effects of trial, lesion type, and their interaction (all p_s < .001). While the intact and TSP-only network showed equivalent performance (p = .636), both showed significantly higher recognition accuracy than the MSP-only network (p_s < .001). In sum, consistent with the prior simulations, the TSP-only network demonstrated better recognition than the MSP-only network, and performed in this case virtually identically to the intact network. For this form of recognition, the MSP-only network was able to perform above chance.

Categorization performance was assessed by presenting sets of cards without any category input and testing the network's ability to output the correct category. The intact and the MSP-only network were able to categorize the sets of cards more effectively than the TSP-only network (Figure 4b). The observed accuracy levels for the intact and MSP-only network were similar to the performance levels typically observed in healthy participants (e.g. Knowlton et al., 1994). The TSP-only network performed close to chance on this task. An ANOVA revealed significant main effects of trial, lesion type, and their interaction (all p_s < .001). Further analyses revealed significant differences between all three lesion conditions (intact vs MSP-only network: p = .0017, others: p_s < .001).

In sum, the network successfully learned probabilistic categories, while also being able to process the individual combinations of cards. The MSP was necessary for the network to learn categories, whereas the TSP contributed more to encoding individual combinations of cards.

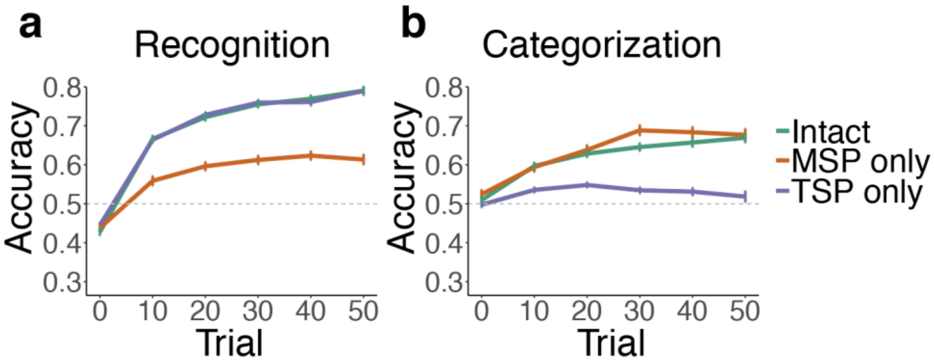

Figure 4. Weather Prediction Task. The model's (a) recognition and (b) categorization performance. Dashed line at 0.5 indicates chance level performance.

Learning intermixed categories with varying typicality

The third set of simulations tested the network's ability to acquire categories with intermixed features and varying typicality. The network was exposed to a set of novel creatures belonging to two categories (Figure 2c; Zeithamova et al., 2008). Each creature had ten binary features, and prototypes of the two categories had no features in common. The rest of the items spanned a continuum between the two prototypes: some items have nine features in common with one prototype and one feature in common with the other prototype; other items have eight features shared with one prototype and two features shared with the other prototype, and so on. If an item has more than five features in common with one prototype, it is considered to belong to the prototype's category (Figure 2c). Each feature was represented by 2 units (one unit for each of the two possible feature values), and each category label was represented by 5 units.

During training, the network learned 20 items, 10 from each category. The model saw each item 5 times for a total of 100 trials, similar to Zeithamova et al. (2008), where participants were presented with 4 runs of 20 items, 10 from each category (80 total trials). Within each category, there were 2 items that shared 9 features with the prototype, 3 items with 8 shared features, 3 items with 7 shared features, and 2 items with 6 shared features (Supplementary Table 5). At test, the network was presented with the training set and a test set consisting of 42 novel items (Supplementary Table 6): the 2 untrained prototypes and 5 items at each distance from the prototype (Zeithamova et al., 2008). We tested the network's ability to remember the atypical features of the training items (atypical feature recognition) and the ability to predict the correct category for the novel items (generalization).

Atypical feature recognition was assessed by testing the ability to activate the correct atypical features in the output layer when presented with trained category exemplars. For each item, we compared the activation of features that did not match the prototypical item (atypical features) to the total activation in the atypical units. The proportion of activation in the correct atypical features was compared against chance (0.1 for items that had only 1 atypical feature, 0.2 for items that had 2 atypical features, etc.). As shown in Figure 5a, the network showed good atypical feature recognition performance. The level of recognition accuracy depended on the level of similarity of the item to its prototype. Atypical features of less typical category members were recognized more easily than atypical features for items very similar to the prototype. The intact network exhibited better performance than the lesioned networks in this task. The TSP-only network performed better than the MSP-only network, which was virtually unable to recognize atypical category members (3 or 4 atypical features), but showed somewhat better performance on items more similar to the prototype (1 or 2 atypical features). The MSP can thus contribute to atypical feature to some extent, when the item is overall very similar to the prototype. The more arbitrary the item, the more the TSP is needed. Effects of lesion, time, number of shared features, and their interactions were all significant ($p_s < .001$). Follow-up analysis confirmed that performance of the three learning conditions differed at all levels of feature overlap. Initial below-chance performance for the exemplars with only 1 atypical feature reflects the tendency to pattern-complete these items to the highly similar prototype.

Generalization was assessed by testing the network's ability to predict the correct category for a set of novel category exemplars based on their features only (no category information was inputted). The mean activation in units representing the correct category was divided by the mean activation across units representing both the correct and incorrect categories. The intact network was able to categorize novel items, the MSP-only network performed better than the intact network, and the TSP-only network performed worse (Figure 5b). Again, all main effects and interactions were significant ($p_s < .001$).

These results are convergent with the simulations above, with the TSP contributing more to recognition than categorization and the MSP contributing more to categorization than recognition. As in the satellite simulation, there was a clear trade-off across pathways in categorization behavior, with the MSP-only network performing better without the influence of the TSP. The recognition results showed an interesting new dimension of behavior as a function of exemplar typicality: the TSP is better than the MSP at remembering the unique features of more atypical exemplars. The more features an exemplar has that depart from the category prototype, the more important the arbitrary binding ability of the TSP.

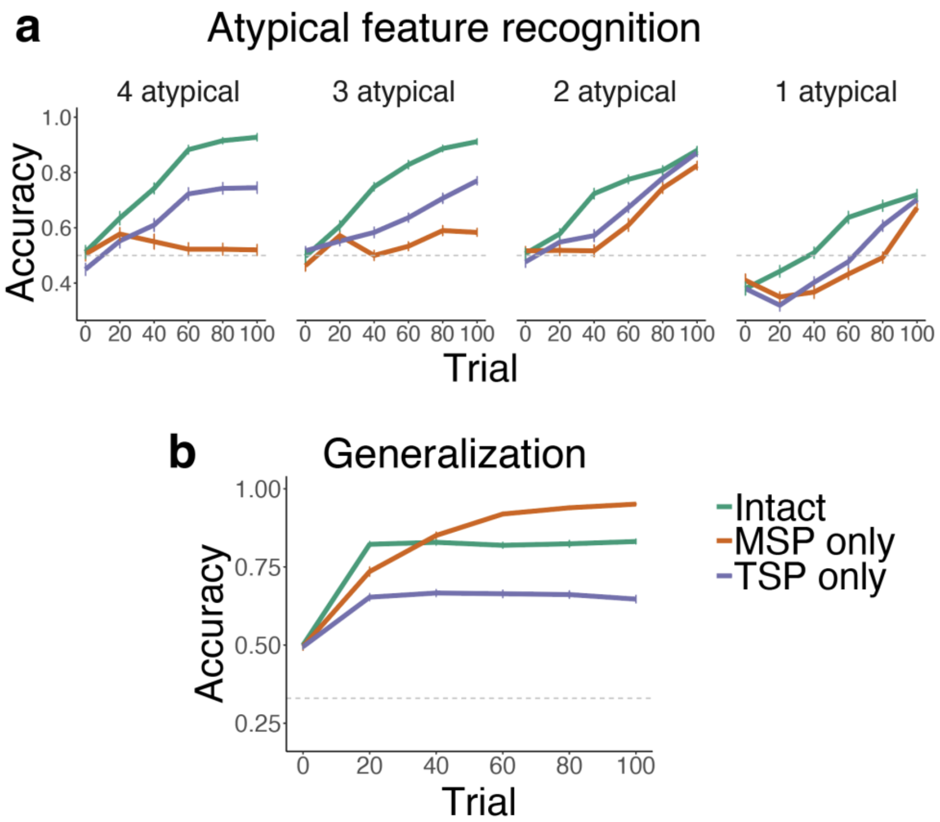

Figure 5. Intermixed categories with varying typicality. The network's (a) recognition performance, and (b) generalization performance.

Discussion

We found that a neural network model of the hippocampus was readily able to learn three different types of categories, providing an account of how the hippocampus may contribute to category learning. Across paradigms, the MSP was critical for detecting the regularities that define category structure. Lower sparsity in this pathway enables distributed (overlapping) representations (Hinton, 1984), which facilitates the detection of commonalities across exemplars, and a relatively lower learning rate helps to integrate this information gracefully over time. In addition to enabling the network to categorize familiar exemplars, the MSP also supports categorization of novel exemplars. After learning, representations of items from the same category were more similar than items from different categories, and this was driven by and especially true in subfield CA1. This is consistent with our recent fMRI finding that CA1 shows stronger within- than across-category representational similarity (Schapiro et al., 2018). The work thus demonstrates that the principles that allowed C-HORSE to detect regularities in structured temporal input (Schapiro, Turk-Browne, et al., 2017) also apply to detecting regularities in multidimensional category spaces.

In contrast to the MSP's capacity for detecting shared structure and generalizing, the main contribution of the TSP to category learning was encoding information about individual category exemplars. Higher sparsity in this pathway allowed the TSP to orthogonalize similar inputs and encode the details of individual exemplars. The ability to quickly bind together arbitrary information that is so useful for episodic memory (e.g., Norman & O'Reilly, 2003) translates into a specialization for remembering the details of individual exemplars in the domain of category learning. This ability proved especially useful for atypical exemplars. Lesioning the TSP resulted in poor recognition with preserved categorization ability. Consistent with these behaviors, a recent study found that strong TSP white matter integrity predicts the ability to learn category exceptions (Schlichting, Gumus, Zhu, & Mack, 2021). We thus propose that the properties of the TSP should make it useful beyond its traditional domain of episodic memory—it should contribute to any new learning that requires memory for arbitrary, as opposed to systematic, information.

We will consider below how our results relate to other models of categorization, the development of categorization ability, and recruitment of different brain regions in category learning in healthy and patient populations.

Relationship to other models of categorization

Our goal was to take a model with an architecture inspired by the anatomy and properties of the hippocampus and explore how the model might accomplish category learning. We did not endeavor to build in any particular strategies for categorization. Interestingly, the behaviors of the model components that emerged from these investigations bear resemblance to existing models of categorization.

The classic exemplar model proposes that people store memory representations of individual category instances and perform similarity judgments on these separate representations at test in order to come to a categorization decision (Medin & Schaffer, 1978; Nosofsky, 2011; Nosofsky & Johansen, 2000). This model has been able to account for many findings across categorization and recognition paradigms (Nosofsky, 1988, 1991; Nosofsky & Zaki, 1998; Palmeri, 1997). The TSP of our model is similar to the exemplar model in that it stores separate traces of individual exemplars. In fact, our model provides an account of how a neural circuit might implement exemplar-style representations: the machinery that leads to pattern separation of individual episodic memories in the TSP, sparse connectivity and a high learning rate, similarly leads to pattern separation across exemplars. The consequence in our model is high fidelity memory for the details of particular exemplars. Unlike the exemplar model, however, our model's TSP exhibited relatively poor categorization. There may be modifications to the model that would allow the TSP to behave more like an exemplar model. For example, the present version of the model does not modulate the influence of the DG during encoding and retrieval, but it is possible that reducing the influence of DG during retrieval would bias the TSP toward pattern completion at test (Lee & Kesner, 2004; Rolls, 1995, 2018), which might enhance certain kinds of categorization. REMERGE (Kumaran & McClelland, 2012) is a model of how the hippocampus might support inference and generalization that relies on pattern separated, conjunctive representations, as in our TSP. The model can accomplish categorization in a manner closely analogous to exemplar models (Kumaran & McClelland, 2012, Appendix), suggesting that there may indeed be ways to increase the categorization ability of a TSP-style representation. Regardless, and across these models, the unique expertise of the TSP-style representation is in its ability to retain the details of individual exemplars.

The classic prototype model postulates that categories are represented by the central tendency across exemplars in a category, without retaining traces of the individual observed exemplars (Minda & Smith, 2011). The prototype model explains categorization behavior well in the context of well-defined, high-coherence categories (Bowman & Zeithamova, 2020; Minda & Smith, 2001). The MSP of our model behaves similarly to a prototype model, in that it tends to abstract across the details of individual exemplars and represent the central tendency. However, this is not true in an absolute sense—the representation in the MSP is sensitive to individual exemplars to some extent.

McClelland and Rumelhart (1985) showed how specifics and generalities can coexist in a neural network model with distributed representations. Our MSP uses distributed representations and shows some degree of this dual sensitivity. However, there *is* a tension between the representation of specifics and generalities in the way that the hidden layers in our model behave. In a hidden layer with very large capacity and a very slow learning rate, distributed internal representations can be carefully and gradually shaped to faithfully reflect the statistics of the environment, which can include representation of both arbitrary and systematic information, to the extent that each is present in the inputs. Neocortical areas of the brain likely have this property of representing arbitrary and general information in harmonious superposition, as in the representations described by McClelland and Rumelhart (1985). But in the case of our hippocampal system, capacity is somewhat more limited and, critically, learning rates are necessarily fast, in order to support behavior on the timescale of a few minutes to hours. The fast learning rate forces trade-offs: Representations can either tend to emphasis the specifics or tend to emphasize the generalities. Whether the hippocampus indeed operates in this parameter space that requires the trade-offs we observe here is a matter for empirical test. Existing data already points to qualitative differences in the behavior and representations of these pathways (e.g. Leutgeb, Leutgeb, Treves, Moser, & Moser, 2004; Nakashiba et al., 2008), but we outline below some specific predictions that will directly test the theory.

Our model assumes that every item is encoded in two different ways, one representation focusing on its details, separating it from other similar items, and the other glossing over the details, emphasizing its similarity to other items. This idea is consistent with neuroimaging data showing coexisting neural representations that are more prototype- and exemplar-like (Bowman, Iwashita, & Zeithamova, 2020). This perspective avoids the kind of discrete category decision making that occurs in a category learning model like SUSTAIN, where a new exemplar either merges with an existing category or separates into a new one (Love, Medin, & Gureckis, 2004). We propose that the brain may have it both ways, solving the tension between representing details and generalities by maintaining both representations in different systems. The solution is closely analogous to that proposed by the Complementary Learning Systems theory, which argued that the hippocampus and neocortex take on complementary roles in memory for encoding the specifics of new items and generalizing across them over time (McClelland et al., 1995). The MSP in our model has properties similar to the neocortex in that framework, with relatively more overlapping representations and a relatively slower learning rate, allowing it to behave as a miniature semantic memory system. The TSP and MSP in our model are thus a microcosm of the broader Complementary Learning Systems dynamic, with the MSP playing the role of a *rapid* learner of novel semantics, relative to the slower learning of neocortex.

Coordinating the contributions of the MSP and TSP

Having two different representations of the same item leads to a problem at retrieval: which representation should be used? In our current work, we have assumed that both representations contribute, and the retrieved information reflects basically an average of the two. But in many cases, there is a trade-off in the utility of the representations, depending on the task. Such trade-offs between representing specifics and regularities have been documented in the literature (e.g. Sherman & Turk-Browne, 2020). We found several cases of trade-offs playing out in our simulations. For example, generalization in the satellite categories is strong in the intact model, which uses both pathways, but much stronger in the version of the model that only uses the MSP. This suggests that a control mechanism that enhances one pathway over another depending on the task could be beneficial for behavior. In a recent paper, we adopted a version of C-HORSE that implemented such a control function in order to explain behavior across tasks with different demands in an associative inference paradigm (Zhou et al., 2021). Medial prefrontal cortex could potentially carry out a control function of this kind, as it participates in category learning (Mack, Preston, & Love, 2020) and is known to modulate CA1 representations as a function of task (Eichenbaum, 2017; Guise & Shapiro, 2017). As the TSP and MSP are both routed through CA1, mPFC control over CA1 could conceivably help coordinate information flow there for optimal behavior. This will be an interesting hypothesis to explore in future modeling and empirical work.

Hippocampal maturation and development of categorization abilities

In humans, the hippocampus has a protracted development, with hippocampal subfields exhibiting different maturations rates (Lavenex & Banta Lavenex, 2013). While the CA1 subfield develops during the first two years of life and reaches adult-like volume around two years, the DG and CA3 subfields develop at a slower pace (Bachevalier, 2013; Gómez & Edgin, 2016; Lavenex & Banta Lavenex, 2013). The projection from the EC to CA1, i.e. the MSP, develops prior to the projection from EC to DG in the TSP (Hevner & Kinney, 1996; Jabes, Lavenex, Amaral, & Lavenex, 2011).

Given the MSP's role in detecting regularities, early maturation of CA1 suggests that the ability to detect regularities should emerge early in development. Indeed, even before their first birthday infants show evidence of categorization (Eimas & Quinn, 1994; Mareschal & Quinn, 2001; Younger & Cohen, 1983) and statistical learning abilities (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002; Saffran, Aslin, & Newport, 1996). There is evidence for involvement of the anterior hippocampus in statistical learning as young as three months (Ellis et al., 2021). Our model predicts that infants should struggle with learning categories that require greater involvement of the TSP (categories with more atypical exemplars or arbitrary features), and that infants should have a poor memory for category exceptions. In line with these predictions, infants' ability to learn categories is found to be affected by the level of category coherence, with less coherent categories being more difficult (Gómez & Lakusta, 2004; Younger, 1990; Younger & Gotlieb, 1988). Moreover, young children demonstrate poorer memory for category exceptions than for typical category members (Savic & Sloutsky, 2019).

Our model may resolve the puzzle in the developmental literature about the discrepancy between infants' precocious performance on categorization tasks on the one hand, and poor episodic memory abilities on the other hand (Keresztes, Ngo, Lindenberger, Werkle-Bergner, & Newcombe, 2018). To the extent that infants have access to the MSP and not TSP, early stages of development would correspond to our MSP-only simulations, where we find poor recognition performance (especially for atypical category instances) but intact categorization and even enhanced generalization. A fully operating basic hippocampal circuitry is eventually needed for learning low-coherence categories and for successful episodic memory functions which emerge later in development (Gómez & Edgin, 2016).

Neuropsychological accounts of hippocampal contributions to category learning

Initial accounts of the role of the hippocampus in category learning came from studies of patients with medial temporal lobe (MTL) damage. Patients have been tested on a range of category learning tasks, including random dot patterns, probabilistic categories, faces, scenes, and painting categorization (Kéri, Kálmán, Kelemen, Benedek, & Janka, 2001; Knowlton & Squire, 1993; Kolodny, 1994; Reber et al., 1996; Reed et al., 1999; Zaki, Nosofsky, Jessup, & Unverzagt, 2003). Knowlton and Squire (1993) tested amnesics' ability to learn abstract novel categories of random dot patterns and observed similar categorization performance as in healthy controls, but impaired recognition, leading to the proposal that the MTL is not involved in category learning. However, amnesics do show impairment on a more difficult version of this task (learning categories A vs. B, as opposed to simply A vs. not-A; Zaki et al., 2003). Amnesics are also impaired on categorizing paintings by artist (Kolodny, 1994), and while they succeed in a categorization task with faces, they fail with scenes (Graham et al., 2006). When learning probabilistic categories, amnesic patients show similar performance to control participants initially (first 50 trials), but fail to reach accuracy levels observed in healthy controls with more exposure (Knowlton et al., 1994). In addition, amnesics are impaired in flexibly using new category knowledge (Reber et al., 1996).

Studies with Alzheimer's disease patients have revealed a similar pattern of performance as in amnesic patients. Alzheimer's patients show intact performance on the A/not-A task (Kéri et al., 2001; Zaki et al., 2003) but poor performance on the A/B task (Zaki et al., 2003). Categorization performance deteriorates as the disease progresses (Kéri et al., 2001). Overall, patients with MTL damage clearly have some ability to learn novel categories, indicating that the hippocampus is not the only region involved in category learning, but they also show clear deficits, especially when aggregating evidence across studies (Zaki, 2004), indicating that the hippocampus makes a causal contribution.

Neuroimaging evidence of hippocampal involvement in category learning

Neuroimaging studies provide strong additional evidence for hippocampal involvement in category learning (Bowman & Zeithamova, 2018; Mack et al., 2016; Mack, Love, & Preston, 2018; Zeithamova et al., 2008). This evidence has motivated the proposal that categorization has computational needs in common with episodic memory (Mack et al., 2018) and decision making (Seger & Peterson, 2013), with the hippocampus as a central neural substrate. The hippocampus and medial prefrontal cortex appear to work together in learning new categories (Bowman & Zeithamova, 2018; Mack et al., 2020), with the hippocampus perhaps playing an especially important role in generalizing knowledge to novel situations (Kumaran, Summerfield, Hassabis, & Maguire, 2009). Using a model-based fMRI approach, Mack and colleagues (2016) showed that the object representations within the hippocampus reflect dynamic updating of category knowledge, and that the representation of an object can change as a function of categorization rules. In the paradigm used in our third simulation, Zeithamova and colleagues (2008) found that activation of the hippocampus correlated with behavioral performance during the categorization task, and Bowman and Zeithamova (2018) found that the hippocampus contributes to generalization in this paradigm.

Hippocampal subfields have not generally been investigated directly in imaging studies of category learning, with one exception being our finding that CA1 (but not CA3/DG) represented category structure in the satellite stimuli (Schapiro, Turk-Browne, Norman, & Botvinick, 2016). However, there is a more indirect way to assess our theory of a division of labor within the hippocampus: Anterior hippocampus has a much larger proportion of the CA1 subfield than posterior hippocampus (Poppenk, Evensmoen, Moscovitch, & Nadel, 2013), so our account predicts that anterior hippocampus should preferentially reflect the learning of category structure. Indeed, several studies have found that activation in the anterior hippocampus is related to category learning (Mack et al., 2016, 2018; Zeithamova et al., 2008) and to prototype-style learning specifically (Bowman & Zeithamova, 2018). Further, activation in the hippocampal body and tail is associated with learning categories that involve exceptions (Davis, Love, & Preston, 2012), which is also consistent with the finding that TSP white matter integrity predicts exception learning (Schlichting et al., 2021). Thus, neuroimaging evidence suggests a strong involvement of the hippocampus in category learning, with some evidence consistent with our account of the nature of this involvement.

Recruitment of multiple neural systems during rapid category learning

As described above, we know that the hippocampus is not the only region contributing to category learning, with the basal ganglia and various regions of the neocortex known to be critically involved (Ashby & Maddox, 2005; Seger & Miller, 2010). While the hippocampus and basal ganglia seem especially important for rapid category learning, on the timescale of minutes to hours, cortical regions likely support slower learning, across days, weeks, and months (Seger & Miller, 2010).

The extent to which different neural systems are involved in category learning relates to properties of the learning task. For example, an fMRI study found that A vs. B category learning primarily engages the MTL, while A vs. not-A task recruits the striatum (Zeithamova et al., 2008), providing an explanation for the pattern of neuropsychological results described above (with MTL damage leading to deficits on A/B but not A/not-A). The need to bind features together to differentiate two separate categories (A vs. B) may be better suited to the computational abilities of the hippocampus. Another important task property that influences the involvement of different systems is the presence of feedback during learning. For example, feedback preferentially engages the basal ganglia in the weather prediction task, whereas the observational version of the same task tends to mainly engage MTL (Poldrack et al., 2001).

Important insights into the flexible recruitment of different neural systems in category learning come from studies with patients suffering from Parkinson's disease (PD; affecting basal ganglia function) relative to amnesic patients (primarily affecting MTL function). PD patients recruit the MTL in category learning to a larger extent than do controls (Moody, Bookheimer, Vanek, & Knowlton, 2004). In addition, PD patients show deficits when learning categories in the feedback-based version of the Weather Prediction Task, but show preserved performance in the observational version of the task (Shohamy et al., 2004). On the other hand, MTL amnesic patients tend to have relatively intact performance in the initial stages of learning, but show deficits as learning progresses (Knowlton et al., 1994; Poldrack et al., 2001).

A neuroimaging study with healthy controls found negatively correlated activity in the MTL and basal ganglia during learning, leading to the proposal that the MTL-based and striatal-based memory systems compete during category learning (Poldrack et al., 2001). There is also evidence, however, that the systems can be recruited in parallel, indicating the possibility for independent or cooperative contributions (Cincotta & Seger, 2007).

In sum, the hippocampus (and MTL more broadly) and basal ganglia are both involved in rapid category learning, with some hints of situations where one or the other may be more important, though more work is needed to flesh this out. Our view, based on the above literature as well as literatures outside the domain of category learning, is that the hippocampus should be especially important in situations that involve more neutral, observational learning, with less motor response, feedback, or reward. Our current model has provided an account of the potential contributions of the hippocampus to category learning, but future modeling and empirical work should expand to explore the interactions with the basal ganglia (as well as mPFC, as described above) in this learning.

Consolidation of category knowledge

The hippocampus plays a role in the rapid learning of novel categories, but ultimately these new representations need to be integrated into neocortical knowledge structures. Offline hippocampal-cortical replay, especially during sleep, may play an important role in this integration process (Klinzing, Niethard, & Born, 2019; Marshall & Born, 2007; McClelland et al., 1995; Rogers & McClelland, 2004). We found that a night of sleep after learning the satellite stimuli improves memory for the category-relevant information, i.e. shared object features, and preserves exemplar-specific information (Schapiro, McDevitt, et al., 2017). There is also evidence that sleep benefits categorization in the Weather Prediction Task (Djonlagic et al., 2009) and the classic dot pattern task (Graveline & Wamsley, 2017). Hippocampal replay is often thought to be a relatively veridical record of experience, though generalized replay has been observed as well (Gupta, van der Meer, Touretzky, & Redish, 2010). The hippocampal replay and behavioral sleep findings are at this point consistent with replay driven by the TSP or a combination of the TSP and MSP (replay driven only by the MSP should not benefit unique features). Regardless of whether hippocampal representations tend to emphasize the specific or the general, new neocortical representations of recent information should help to bring out the shared, general category structure (McClelland et al., 1995), consistent with this literature.

Predictions and conclusions

We have put forward an account of the possible contributions of the hippocampus to rapid, novel category learning. We propose that the TSP, known for its rapid binding and pattern separation computations, contributes to remembering the *arbitrary* aspects of categories — the specifics of individual exemplars or observations and the exceptions to the category rules. The MSP, with a relatively slower learning rate and more distributed representations, contributes to the *systematic* aspects of categories — the structure shared across category exemplars. This proposal for two systems within the hippocampus with complementary expertise makes specific predictions about the response properties in the two pathways during learning as well as consequences of damage or anatomical variation in the pathways.

The neural responses to exemplars from the same category should tend to be more similar in CA1 than in CA3 and DG. This should be especially true immediately after presentation of a stimulus, as the circuit is recurrent, so information in CA1 can spread through EC back to DG and CA3 with more processing time (as in Fig. 3d). To be concrete, assume a human or animal has learned that stimuli A, B, and C belong to the same category. In an experiment recording from neurons from several subfields, the model predicts that when viewing A, the initial pattern of activity in DG and CA3 should look dissimilar to that for B and C, whereas the initial pattern of activity in CA1 should look similar to B and C. With additional processing time, DG and CA3 may start to show some of that similarity structure as well. The degree of within-category similarity structure in CA1 should predict the ability to remember shared structure and to generalize to novel exemplars, and the degree of separation in DG and CA3 should predict the ability to remember unique aspects of the exemplars.

There are many changes to the human brain that are known to, or could plausibly result in, differential strength of the two pathways, including development, aging, psychiatric disorders, and neurological disease. In rodent models, it is possible to separately lesion the two pathways (Nakashiba et al., 2008). There is also variance across people (or animals) in the normal anatomical integrity of the two pathways that can be measured (Schlichting et al., 2021). In general, we expect double dissociations in the behavior resulting from strength vs. weakness of the TSP vs. MSP: Weakness specifically in the TSP should result in poor memory for specific, arbitrary features of exemplars but preserved memory for structure shared across exemplars. Weakness specifically in the MSP should result in poor memory for shared structure and relatively preserved memory for specifics. We predict behavioral consequences to be stronger in paradigms that involve more passive, observational learning, as the basal ganglia is more likely to be able to pick up the slack in tasks involving motor responses and feedback.

There are several empirical datapoints that already fit these predictions, including category-related similarity structure in CA1 (Schapiro et al., 2016), TSP white matter integrity predicting exception learning (Schlichting et al., 2021), and behavioral exception learning that unfolds in accordance with MSP and TSP properties (Heffernan, Schlichting, & Mack, 2021). But more work is needed to establish the extent to which, and the conditions under which, this account correctly characterizes the contribution of the hippocampus to category learning. We hope the model inspires new empirical theoretically diagnostic work, which will in turn inform model development and expansion.

Acknowledgments

We are grateful to Dhairyya Singh for helpful discussions of this work. This work was supported by a St Hugh's College Travel Grant, University of Oxford, to JS.

References

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The Emergent neural modeling system. *Neural Networks*, *21*, 1146–1152. https://doi.org/10.1016/j.neunet.2008.06.016

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol*, *56*, 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Bachevalier, J. (2013). The Development of Memory from a Neurocognitive and Comparative Perspective. In P. J. Bauer & R. Fivush (Eds.), *The Wiley Handbook on the Development of Children's Memory* (pp. 109–125). Wiley-Blackwell. https://doi.org/10.1002/9781118597705.ch6

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife*, *9*(e59360). https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract Memory Representations in the Ventromedial Prefrontal Cortex and Hippocampus Support Concept Generalization. *The Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2020). Training Set Coherence and Set Size Effects on Concept Generalization and Recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *46*(8), 1442–1464. https://doi.org/10.1037/xlm0000824

Cincotta, C. M., & Seger, C. A. (2007). Dissociation between Striatal Regions while Learning to Categorize via Feedback and via Observation. *Journal of Cognitive Neuroscience*, *19*(2), 249–265. https://doi.org/doi: 10.1162/jocn.2007.19.2.249.

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members. *Cerebral Cortex*, *22*(2), 260–273. https://doi.org/10.1093/CERCOR/BHR036

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., & Stickgold, R. (2009). Sleep enhances category learning. *Learning & Memory*, *16*, 751–755. https://doi.org/10.1101/lm.1634509.but

Eichenbaum, H. (2017). On the Integration of Space, Time, and Memory. *Neuron*, *95*(5), 1007–1018. https://doi.org/10.1016/j.neuron.2017.06.036

Eimas, P. D., & Quinn, P. C. (1994). Studies on the Formation of Perceptually Based Basic-Level Categories in Young Infants Based Basic-Level Categories in Young Infants. *Child Development*, *65*(3), 903–917. https://doi.org/10.2307/1131427

Eldridge, L. L., Masterman, D., & Knowlton, B. J. (2002). Intact implicit habit learning in Alzheimer's disease. *Behavioral Neuroscience, 116*(4), 722–726. https://doi.org/10.1037/0735-7044.116.4.722

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2021). Evidence of hippocampal learning in human infants. *Current Biology*, *31*(15), 3358-3364.e4. https://doi.org/10.1016/j.cub.2021.04.072

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899

Gómez, R. L., & Edgin, J. O. (2016). The extended trajectory of hippocampal development: Implications for early memory development and disorder. *Developmental Cognitive Neuroscience*, *18*, 57–69. https://doi.org/10.1016/j.dcn.2015.08.009

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science*, *7*(5), 567–580. https://doi.org/10.1111/j.1467-7687.2004.00381.x

Graham, K. S., Scahill, V. L., Hornberger, M., Barense, M. D., Lee, A. C. H., Bussey, T. J., & Saksida, L. M. (2006). Abnormal Categorization and Perceptual Learning in Patients with Hippocampal Damage. *Journal of Neuroscience*, *26*(29), 7547–7554. https://doi.org/10.1523/JNEUROSCI.1535-06.2006

Graveline, Y. M., & Wamsley, E. J. (2017). The impact of sleep on novel concept learning. *Neurobiology of Learning and Memory*, *141*, 19–26. https://doi.org/10.1016/j.nlm.2017.03.008

Guise, K. G., & Shapiro, M. L. (2017). Medial Prefrontal Cortex Reduces Memory Interference by Modifying Hippocampal Encoding. *Neuron*, *94*(1), 183-192.e8. https://doi.org/10.1016/j.neuron.2017.03.011

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal Replay Is Not a Simple Function of Experience. *Neuron*, *65*(5), 695–705. https://doi.org/10.1016/j.neuron.2010.01.034

Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. *Scientific Reports*, *11*(1), 1–14. https://doi.org/10.1038/s41598-021-00864-9

Hevner, R. F., & Kinney, H. C. (1996). Reciprocal entorhinal-hippocampal connections established by human fetal midgestation. *Journal of Comparative Neurology*, *372*(3), 384–394. https://doi.org/10.1002/(SICI)1096-9861(19960826)372:3<384::AID-CNE4>3.0.CO;2-Z

Hinton, G. (1984). *Distributed representations*. *Technical Report. Carnegie-Mellon University*. Pittsburgh, PA.

Jabes, A., Lavenex, P. B., Amaral, D. G., & Lavenex, P. (2011). Postnatal development of the hippocampal formation: A stereological study in macaque monkeys. *Journal of Comparative Neurology*, *519*(6), 1051–1070. https://doi.org/10.1002/CNE.22549

Keresztes, A., Ngo, C. T., Lindenberger, U., Werkle-Bergner, M., & Newcombe, N. S. (2018). Hippocampal Maturation Drives Memory from Generalization to Specificity. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2018.05.004

Kéri, S., Kálmán, J., Kelemen, O., Benedek, G., & Janka, Z. (2001). Are Alzheimer's disease patients able to learn visual prototypes? *Neuropsychologia*, *39*(11), 1218–1223. https://doi.org/10.1016/S0028-3932(01)00046-X

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus. *PLoS Computational Biology*, *9*(6), e1003067. https://doi.org/10.1371/journal.pcbi.1003067

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Klinzing, J. G., Niethard, N., & Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, *22*, 1598–1610. https://doi.org/10.1038/s41593-019-

0467-3

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A Neostriatal Habit Learning System in Humans. *Science*, *273*, 1399–1402. https://doi.org/10.1126/science.273.5280.1399

Knowlton, B. J., & Squire, L. R. (1993). The Learning of Categories : Parallel Brain Systems for Item Memory and Category Knowledge. *Science*, *262*(5140), 1747–1749. https://doi.org/10.1126/science.8259522

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic Classification Learning in Amnesia. *Learning & Memory*, *1*, 106–120.

Kolodny, J. A. (1994). Memory processes in classification learning: An Investigation of Amnesic Performance in Categorization of Dot Patterns and Artistic Styles. *Psychological Science*, *5*(3), 164–169. https://doi.org/10.1111/j.1467-9280.1994.tb00654.x

Kumaran, D., & Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, *17*(9), 735–748. https://doi.org/10.1002/hipo.20326

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. https://doi.org/10.1037/a0028681

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, *63*(6), 889–901. https://doi.org/10.1016/J.NEURON.2009.07.030

Lavenex, P., & Banta Lavenex, P. (2013). Building hippocampal circuits to learn and remember: Insights into the development of human memory. *Behavioural Brain Research*, *254*, 8–21. https://doi.org/10.1016/J.BBR.2013.02.007

Lee, I., & Kesner, R. P. (2004). Encoding versus retrieval of spatial memory: Double dissociation between the dentate gyrus and the perforant path inputs into CA3 in the dorsal hippocampus. *Hippocampus*, *14*(1), 66–76. https://doi.org/10.1002/hipo.10167

Lee, I., Rao, G., & Knierim, J. J. (2004). A double dissociation between hippocampal subfields: Differential time course of CA3 and CA1 place cells for processing changed environments. *Neuron*, *42*(5), 803–815. https://doi.org/10.1016/j.neuron.2004.05.010

Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M. B., & Moser, E. I. (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science*, *305*(5688), 1295–1298. https://doi.org/10.1126/science.1100265

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *PNAS*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-019-13930-8

Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, *5*(10), 443–450. https://doi.org/10.1016/S1364-6613(00)01752-6

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences*, *11*(10), 442–450. https://doi.org/10.1016/j.tics.2007.09.001

Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, *58*, 25–45. https://doi.org/10.1146/ANNUREV.PSYCH.57.102904.190143

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed Memory and the Representation of General and Specific Information. *Journal of Experimental Psychology. General*, *114*(2), 159–188. article. https://doi.org/10.1037/0096-3445.114.2.159

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1123–1136. https://doi.org/10.1098/RSTB.2002.1099

Miller, E. K., Nieder, A., Freedman, D. J., & Wallis, J. D. (2003). Neural correlates of categories and concepts. *Current Opinion in Neurobiology*, *13*(2), 198–203. https://doi.org/10.1016/S0959-4388(03)00037-0

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799. https://doi.org/10.1037//0278-7393.27.3.775

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 40–64). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.003

Moody, T. D., Bookheimer, S. Y., Vanek, Z., & Knowlton, B. J. (2004). An Implicit Learning Task Activates Medial Temporal Lobe in Patients with Parkinson's Disease. *Behavioral Neuroscience*, *118*(2), 438–442. https://doi.org/10.1037/0735-7044.118.2.438

Nakashiba, T., Young, J. Z., McHugh, T. J., Buhl, D. L., & Tonegawa, S. (2008). Transgenic Inhibition of Synaptic Transmission Reveals Role of CA3 Output in Hippocampal Learning. *Science, 319*, 1260–1264. https://doi.org/10.1126/science.1151120

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Nosofsky, R. M. (1988). Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1991). Tests of an Exemplar Model for Relating Perceptual Classification and Recognition Memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M. (2011). The generalized context model: an exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 18–39). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*(3), 375–402. https://doi.org/10.1007/BF03543066

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between Categorization and Recognition in Amnesic and Normal Individuals: An Exemplar-Based Interpretation. *Science*, *9*(4), 247–255. https://doi.org/10.1111/1467-9280.00051

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311–345. https://doi.org/10.1037/0033-295X.108.2.311

O'Reilly, R., Munakata, Y., Frank, M. J., Hazy, T., & Contributors. (2014). *Computational cognitive neuroscience, 2nd edn., ch. 8.* See https://grey. colorado.edu/CompCogNeuro/index.php/CCNBook/ MainAugust.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning Memory and Cognition*, *23*(2), 324–354. https://doi.org/10.1037/0278-7393.23.2.324

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550. https://doi.org/10.1038/35107080

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, *17*(5), 230–240. https://doi.org/10.1016/j.tics.2013.03.005

Reber, P. J., Knowlton, B. J., & Squire, L. R. (1996). Dissociable properties of memory systems: Differences in the flexibility of declarative and nondeclarative knowledge. *Behavioral Neuroscience*, *110*(5), 861–871. https://doi.org/10.1037/0735-7044.110.5.861

Reed, J. M., Squire, L. R., Patalano, A. L., Smith, E. E., & Jonides, J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience*, *113*(3), 411–419. https://doi.org/10.1037/0735-7044.113.3.411

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition : a parallel distributed processing approach*. MIT Press.

Rolls, E. T. (1995). A model of the operation of the hippocampus and entorhinal cortex in memory. *International Journal of Neural Systems*, *6*, 51–70.

Rolls, E. T. (2018). The storage and recall of memories in the hippocampo-cortical system. *Cell and Tissue Research*, *373*, 577–604. https://doi.org/10.1007/s00441-017-2744-3

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926. https://doi.org/doi: 10.1126/science.274.5294.1926

Savic, O., & Sloutsky, V. M. (2019). Assimilation of exceptions? Examining representations of regular and exceptional category members across development. *Journal of Experimental Psychology: General*, *148*(6), 1071–1090. https://doi.org/10.1037/XGE0000611

Schapiro, A. C., Kustner, L. V, & Turk-Browne, N. B. (2012). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Current Biology*, *22*, 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., & Rogers, T. T. (2017). Sleep Benefits Memory for Semantic Category Structure while Preserving Exemplar-Specific Information. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-12884-5

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018).

Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*(3920). https://doi.org/10.1038/s41467-018-06213-1

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modeling approach to recomciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*, 20160049. https://doi.org/10.1098/rstb.2016.0049

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Schlichting, M. L., Gumus, M., Zhu, T., & Mack, M. L. (2021). The structure of hippocampal circuitry relates to rapid category learning in humans. *Hippocampus*, *31*(11), 1179–1190. https://doi.org/10.1002/hipo.23382

Seger, C. A., & Miller, E. K. (2010). Category Learning in the Brain. *Annu. Rev. Neurosci*, *33*, 203–219. https://doi.org/10.1146/annurev.neuro.051508.135546

Seger, C. A., & Peterson, E. J. (2013). Categorization = decision making + generalization. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1187–1200. https://doi.org/10.1016/J.NEUBIOREV.2013.03.015

Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(37), 22760–22770. https://doi.org/10.1073/pnas.2013291117

Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. *Brain*, *127*(4), 851–859. https://doi.org/10.1093/brain/awh100

Spiridon, M., & Kanwisher, N. (2002). How Distributed Is Visual Category Information in Human Occipito-Temporal Cortex? An fMRI Study a low level baseline such as a fixation point in an other-wise blank screen. The existence of partial responses to nonpreferred. *Neuron*, *35*, 1157–1165. https://doi.org/10.1016/S0896-6273(02)00877-2

Team, R. C. (2019). R: A Language and Environment for Statistical Computing. *Vienna, Austria*.

Younger, B. A. (1990). Infants ' Detection of Correlations among Feature Categories. *Child Development*, *61*(3), 614–620. https://doi.org/10.2307/1130948

Younger, B. A., & Cohen, L. B. (1983). Infant Perception of Correlations among Attributes. *Child Development*, *54*(4), 858–867. https://doi.org/10.2307/1129890

Younger, B. A., & Gotlieb, S. (1988). Development of categorization skills: Changes in the nature or structure of infant form categories? *Developmental Psychology*, *24*(5), 611–619. https://doi.org/10.1037/0012-1649.24.5.611

Zaki, S. R. (2004). Is categorization performance really intact in amnesia? A meta-analysis. *Psychonomic Bulletin and Review*, *11*(6), 1048–1054. https://doi.org/10.3758/BF03196735

Zaki, S. R., Nosofsky, R. M., Jessup, N. M., & Unverzagt, F. W. (2003). Categorization and recognition performance of a memory-impaired group: Evidence for single-system models. *Journal of the International Neuropsychological Society*, *9*(3), 394–406. https://doi.org/10.1017/S1355617703930050

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable Prototype Learning Systems: Evidence from Brain Imaging and Behavior. *The Journal of Neuroscience*, *28*(49), 13194–13201. https://doi.org/10.1523/JNEUROSCI.2915-08.2008

Zhou, Z., Singh, D., Tandoc, M. C., & Schapiro, A. C. (2021). Distributed representations for human inference. *BioRxiv Preprint*. https://doi.org/10.1101/2021.07.29.454337

Supplementary material for:

A neural network model of hippocampal contributions to category learning

Jelena Sučević & Anna C. Schapiro

Supplementary Table 1. Satellites: Input patterns used in training and for categorization test (based on Schapiro, McDevitt, Rogers, Mednick, & Norman, 2018). If a feature is present in an item, the input is 1, otherwise 0. Each of the three categories has one prototype (items 1, 6, and 11) that consists of entirely shared features, and exemplars that have four shared features and one unique feature. C1=Category 1; C2=Category 2; C3=Category 3.

		Features																										
Cat	**Item**	**Shared** (C1)					**Unique** (C1)				**Shared** (C2)					**Unique** (C2)				**Shared** (C3)					**Unique** (C3)			
1	**1**	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	5	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	**6**	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	8	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
2	9	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	10	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
3	**11**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0
3	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
3	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	1	0	0
3	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	0
3	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	1

Supplementary Table 2. Satellites: Input patterns for the unique and shared feature memory test.

		Features																									
Cat	Item	Shared (C1)				Unique (C1)				Shared (C2)					Unique (C2)				Shared (C3)					Unique (C3)			
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
3	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Supplementary Table 3. Satellites: Input patterns for the generalization test set. Each exemplar consists of two shared and two unique features.

		Features																										
Cat	**Item**	**Shared** (C1)					**Unique** (C1)				**Shared** (C2)					**Unique** (C2)				**Shared** (C3)					**Unique** (C3)			
1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	5	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	6	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	7	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
2	8	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
2	9	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
2	10	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
2	11	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
2	12	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
3	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0
3	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1
3	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1
3	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0
3	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1
3	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0

Supplementary Table 4. Weather prediction task: input patterns (based on Knowlton, Squire, & Gluck, 1994). There are four cards and each pattern consists of a combination of 1, 2 or 3 cards. Each card is represented by one unit (if the card is presented in a pattern, the input is 1, otherwise it is 0).

Pattern	Card 1	Card 2	Card 3	Card 4	P(outcome 1)	Pattern probability
1	0	0	0	1	0.15	0.14
2	0	0	1	0	0.38	0.084
3	0	0	1	1	0.1	0.087
4	0	1	0	0	0.62	0.084
5	0	1	0	1	0.18	0.064
6	0	1	1	0	0.5	0.047
7	0	1	1	1	0.21	0.041
8	1	0	0	0	0.85	0.14
9	1	0	0	1	0.5	0.058
10	1	0	1	0	0.82	0.064
11	1	0	1	1	0.43	0.032
12	1	1	0	0	0.9	0.087
13	1	1	0	1	0.57	0.032
14	1	1	1	0	0.79	0.041

Supplementary Table 5. Intermixed categories with varying typicality: training set (based on Zeithamova, Maddox, & Schnyer, 2008). Each exemplar consists of ten binary features, and each feature is represented by two units (one unit for each possible feature value).

		Features																			
		1		2		3		4		5		6		7		8		9		10	
Cat	Item	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
1	2	1	0	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1	3	1	0	1	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0	1	0
1	4	1	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0
1	5	1	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0	1	0	1	0
1	6	1	0	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1	0	0	1
1	7	0	1	1	0	1	0	0	1	1	0	1	0	1	0	1	0	0	1	1	0
1	8	1	0	1	0	1	0	0	1	0	1	1	0	1	0	0	1	1	0	1	0
1	9	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	1	0
1	10	0	1	0	1	1	0	1	0	1	0	0	1	1	0	1	0	1	0	0	1
2	11	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1
2	12	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0	1
2	13	0	1	1	0	0	1	0	1	0	1	0	1	0	1	1	0	0	1	1	0
2	14	0	1	0	1	1	0	0	1	1	0	0	1	0	1	1	0	0	1	0	1
2	15	0	1	0	1	0	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1
2	16	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0
2	17	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	1	0
2	18	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1
2	19	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
2	20	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	1	0	1

Supplementary Table 6. Intermixed categories with varying typicality: test set. Items 1 and 42 represent prototypes of the two categories and have no overlapping features, and the rest of the items span the continuum between the two prototypes.

		Features																			
		1		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**		**10**	
Cat	**Item**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
1	**1**	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1	2	1	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1	3	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	1	0
1	4	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1	5	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1	0
1	6	1	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	1	0	1	0
1	7	1	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	0	1	1	0
1	8	1	0	1	0	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0
1	9	0	1	1	0	1	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0
1	10	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	0	1
1	11	1	0	0	1	1	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0
1	12	0	1	1	0	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1	0
1	13	1	0	0	1	1	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0
1	14	1	0	1	0	1	0	1	0	0	1	1	0	1	0	0	1	1	0	0	1
1	15	1	0	0	1	0	1	1	0	1	0	1	0	1	0	1	0	0	1	1	0
1	16	1	0	1	0	1	0	0	1	1	0	0	1	1	0	1	0	1	0	0	1
1	17	0	1	1	0	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1
1	18	1	0	1	0	0	1	1	0	1	0	0	1	0	1	1	0	1	0	0	1
1	19	1	0	0	1	1	0	1	0	0	1	0	1	1	0	1	0	0	1	1	0
1	20	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	1	0	1	0
1	21	0	1	1	0	1	0	0	1	0	1	1	0	0	1	1	0	1	0	1	0
2	22	1	0	1	0	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	1
2	23	0	1	0	1	0	1	0	1	1	0	1	0	1	0	0	1	1	0	0	1
2	24	1	0	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1	1	0
2	25	0	1	1	0	0	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1
2	26	0	1	0	1	1	0	0	1	0	1	1	0	0	1	1	0	0	1	1	0
2	27	0	1	1	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1	0	1
2	28	0	1	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0
2	29	1	0	1	0	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	1
2	30	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	1	0
2	31	0	1	0	1	1	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1
2	32	0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	1	0	1	0	1
2	33	0	1	0	1	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	1
2	34	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1
2	35	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0

2	36	0	1	0	1	0	1	0	1	0	1	1	0	0	1	1	0	0	1	0	1
2	37	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1
2	38	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0	1
2	39	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1
2	40	0	1	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1
2	41	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	1
2	**42**	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Supplementary Table 7. Layer size and inhibition parameters. All values the same as Schapiro, Turk-Browne, Botvinick, & Norman, 2017 except EC size and corresponding kWTA pct.

Area	**# Units**	**kWTA type**	**Proportion activity (kWTA pct)**	**kWTA pt**
EC_{in} and EC_{out}	27 / 8 / 30	kWTA Avg Inhib	K = 5 / K = 6 / K = 10	0.5
DG	400	kWTA Avg Inhib	0.01	0.9
CA3	80	kWTA Avg Inhib	0.06	0.7
CA1	100	kWTA Avg Inhib	0.25	0.7

Supplementary Table 8. Parameters for projections between layers. All values the same as Schapiro, Turk-Browne, Botvinick, & Norman, 2017 except where underlined.

Projection	**Weight range**	**Scale** **(abs / rel)**	**Connectivity**	**lrate** **sim1 / sim2 / sim3**
Input → EC_{in}	0.25 – 0.75	1 / 1	1 to 1	0
EC_{in} → DG	0.25 – 0.75	1 / 1	25%	0.2
EC_{in} → CA3	0.25 – 0.75	1 / 1	25%	0.2
DG → CA3 (*mossy fiber*)	0.89 – 0.91	1 / 8	25%	0
CA3 → CA3	0.25 – 0.75	1 / 1	5%	0.2
CA3 → CA1 (*Schaffer*)	0.25 – 0.75	1 / 1	100%	0.05
EC_{in} → CA1	0.25 – 0.75	3 / 1	100%	0.02 / 0.02 / 0.002
CA1 → EC_{out}	0.25 – 0.75	1 / 1	100%	0.002 / 0.002 / 0.02
EC_{out} → CA1	0.25 – 0.75	1 / 1	100%	0.002
EC_{out} → EC_{in}	0.49 – 0.51	2 / .5	1 to 1	0