Neurophysiological and computational mechanisms of non-associative and associative memories during complex human behavior

Yuchen Xiao[1,2*], Paula Sánchez López[3,4*], Ruijie Wu[5,6*], Ravi Srinivasan[7], Peng-Hu Wei[8,9], Yong-Zhi Shan[8,9], Daniel Weisholtz[10], Garth Rees Cosgrove[10], Joseph R Madsen[11], Scellig Stone[11], Guo-Guang Zhao[8,9‡], Gabriel Kreiman[11,12‡]

[1]Harvard University, Cambridge, MA, USA

[2]Westlake University, Hangzhou, Zhejiang, China

[3]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[4]Harvard Medical School, Boston, MA, USA

[5]State Key Laboratory of Brain and Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

[6]Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei, China

[7]Eidgenössische Technische Hochschule (ETH), Zurich, Switzerland.

[8]Department of Neurosurgery, Xuanwu Hospital, Capital Medical University, Beijing, China

[9]National Center for Neurological Disorders, Beijing, China

[10]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[11]Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

[12]Center for Brains, Minds and Machines, Cambridge, MA, USA

*These authors contributed equally to this work.

‡To whom correspondence should be addressed: Gabriel.kreiman@tch.harvard.edu

Competing Interest Statement: The authors have no competing interests.

Keywords: Working memory, human neurophysiology, recognition memory, associative memory, natural behavior

This PDF file includes: Main Text, Figures 1 to 9**,** Table 1

Supplementary material: 13 supplementary figures, 9 supplementary tables

Abstract

The ability to transiently remember what happened where and when is a cornerstone of cognitive function. Forming and recalling working memories depends on detecting novelty, building associations to prior knowledge, and dynamically retrieving context-relevant information. Previous studies have scrutinized the neural machinery for individual components of recognition or associative memory under laboratory conditions, such as recalling elements from arbitrary lists of words or pictures. In this study, we implemented a well-known card-matching game that integrates multiple components of memory formation together in a naturalistic setting to investigate the dynamic neural processes underlying complex natural human memory. We recorded intracranial field potentials from 1,750 depth or subdural electrodes implanted in 20 patients with pharmacologically-intractable epilepsy while they were performing the task. We leveraged generalized linear models to simultaneously assess the relative contribution of neural responses to distinct task components. Neural activity in the gamma frequency band signaled novelty and graded degrees of familiarity, represented the strength and outcome of associative recall, and finally reflected visual feedback on a trial-by-trial basis. We introduce an attractor-based neural network model that provides a plausible first-order approximation to capture the behavioral and neurophysiological observations. The large-scale data and models enable dissociating and at the same time dynamically tracing the different cognitive components during fast, complex, and natural human memory behaviors.

Introduction

Working memory serves as a fundamental component of our cognitive abilities, enabling us to store and retrieve immediate information. In stark contrast to most efforts in current artificial intelligence algorithms, the transient storage of memories occurs in a largely unsupervised fashion, with single or limited exposure. The formation and recall of memories require assessing novelty versus familiarity, building bridges between sensory inputs and prior knowledge, connecting spatial and temporal cues and effectively retrieving information in the context of current task demands. While substantial literature exists on neural responses in laboratory-based tasks for separate components of working memory, our understanding of how these components are integrated and coordinated in real-life tasks remains limited.

Non-associative recognition memory refers to the ability to judge the prior occurrence of a stimulus. Judging whether an item is novel or not is necessary for its successful memory encoding[1], and recognizing an item as familiar facilitates memory retrieval[2]. Several studies have documented correlates of recognition memory for novelty versus familiarity, primarily but not exclusively, in medial temporal lobe (MTL) structures in rodents, monkeys, and humans[3-14]. Many studies have focused on tasks that involve presenting a list of items, such as words, pictures, or video clips, and either recalling items from these lists or assessing recognition memory for those items (e.g.,[6,8-12,15-20]). Both novel and familiar items need to be incorporated into the body of prior knowledge by forming novel associations. Associative memory refers to the ability to link items or evaluate the correctness of such associations (e.g.,[21-30]). Associative memory has been commonly investigated by having participants learn pairs of items and recalling one of the two items given the other or assessing whether a given association is correct or not.

Although recognition memory and associative memory have been largely studied separately, they are not independent in real-world memory tasks. The successful implementation of associative memory is contingent on basic recognition processes. To understand the connections and dissociations between different components of memory formation during natural and complex behavior, here we recorded intracranial field potentials from 20 patients with pharmacologically intractable epilepsy while they played a classical card matching game,

colloquially known as the "Memory" game (**Figure 1, Movie S1**). We focused on the neural activity in the gamma frequency band (30-150 Hz)[31-35]. Participants thrived in the task, demonstrating dependence on the task memory demands and temporal recency effects. Using generalized linear models, we characterized how neural responses are modulated by the different behavioral components involved in the task. Our results demonstrate that neural circuits can represent novelty and familiarity independently of the sensory content, along with the strength and outcome of associative recall on a trial-by-trial basis.

To better understand the mechanisms underlying memory formation and retrieval, we turn to computational models[36]. Models rooted in persistent neuronal activity[37-45] provide important insights into working memory's neural basis. Recent perspectives, including those involving attractor networks[46-52], have also highlighted the significance of Hebbian synaptic plasticity and short-term depression and facilitation as means to enhance memory encoding[53-57]. As a proof-of-principle, we introduce a simple attractor-based neural network model that provides a first-order approximation to describe the behavioral and neurophysiological observations.

Results

We recorded intracranial field potentials (IFPs) from 20 patients with pharmacologically intractable epilepsy implanted with depth electrodes (**Table S1**, one participant also had subdural surface electrodes). Participants played a memory-matching game (**Figure 1, Movie S1, Methods**). Each trial consisted of two self-paced clicks. Clicking on a tile revealed an image (**Figure 1A**). Image categories included person, animal, food, vehicle, and indoor scenes. If the two tiles in a trial contained the same image (*match*, **Figure 1B**), the two tiles turned green and could not be clicked again for the remainder of the block. If the two images were different (*mismatch*, **Figure 1A** and **1C**), the two tiles turned black and could be clicked again. Participants started in a 3×3 tile board block like the one shown in **Figure 1** and progressed to more difficult blocks (4×4, 5×5, 6×6, or 7×7 tiles). All tiles had a corresponding match, except for one tile in the boards with an odd number of tiles (3×3, 5×5, and 7×7).

Mismatch trials showed longer reaction times and were associated with less frequent and less recent exposure to matching pairs

The average number of clicks per tile increased with difficulty (board size), as expected (**Figure 2A**). All participants performed much better than a memoryless model (random clicking, $p<0.001$, here and in subsequent tests unless stated otherwise: permutation test, 5,000 iterations, one-tailed) and performed worse than a model assuming perfect memory ($p<0.001$, **Figure 2A**). The reaction time (RT) was defined as the time interval between the first and second clicks within a trial (**Figure 1A**). The reaction time was longer for mismatch than match trials for all board sizes ($p<0.007$, **Figure 2B**).

For a tile in a given trial, we defined *n-since-last-click* (nslc) as the number of clicks elapsed since the last time *the same tile* was clicked (**Figure 1A**-**B**). As expected, nslc increased with board size ($p<0.001$, linear regression, F-test, **Figure 2C**-**D**). For the 2nd tile, nslc was larger in mismatch compared to match trials for all board sizes except the 3×3 case ($p<0.001$, **Figure 2D**), a reflection of the decay in memory for tiles that were not seen recently. The larger nslc in mismatch trials also held for the 1st tile only for the 7×7 board size ($p<0.001$, **Figure 2C**). If the participants believe that they know the locations of both tiles in a matching pair, a reasonable strategy is to click first the tile they are less sure about, likely because they have seen this tile earlier rather than later in the block. This strategy accounts for the differences between the 1st tile (**Figure 2C**) and 2nd tile (**Figure 2D**).

For a tile in a given trial, we defined *n-since-pair* (nsp) as the number of clicks since the last time when *its matching pair* was seen (**Figure 1A**, **C**). As expected, nsp increased with board size given the increased difficulty ($p<0.001$, linear regression, F-test, **Figure 2E-2F**). Additionally, the more recent the tile's matching pair was seen, the more likely the trial was a match. Thus, nsp was larger in mismatch compared to match trials in all cases except the 3x3 board size for the 1st tile ($p<0.001$, **Figure 2E**). For the 2nd tile, nsp for any match trial was always one because the matching pair would have been revealed in the previous click, by definition. Thus, there was a large difference between nsp between match and mismatch trials ($p<0.001$, **Figure 2F**).

We performed infrared eye-tracking on ten healthy participants while they performed the same task. Participants fixated on the tile they clicked, both for the first and second tiles, and both for match and mismatch trials (**Figure S1**). For the first tile, there was no difference in the dynamics of saccades towards and away from the target tile between match and mismatch trials (**Figure S1A**) after equalizing the RT and the distances between the 1st and the 2nd tiles. For the second tile, there was no difference in the dynamics of saccades toward the target tile before the click. However, within the 1 second window after the 2nd click, the distance to the center of the tile was, on average, 1.76 dva (degrees of visual angle) larger for match than mismatch trials (**Figure S1B**). This small difference may be attributed to participants' lingering slightly longer during mismatch trials, arguably in an effort to remember the tile.

Neural signals reflect novelty and familiarity

We recorded intracranial field potentials from 1,750 electrodes (**Table S2**). We excluded 582 electrodes due to bipolar referencing, locations in pathological sites, or signals containing artifacts (**Methods**). We included in the analyses 676 bipolarly referenced electrodes in the gray matter (**Figure S2**) and 492 in the white matter (**Figure S3**). **Table S2** describes electrode locations separated by brain region and hemisphere. Although the white matter is presumed to contain mostly myelinated axons, previous studies have shown that intracranial field potential signals from the white matter can demonstrate biologically meaningful information[58,59]. Such signals could reflect small errors in electrode localization on the order of ~2 millimeters and also the spread of intracranial field potential signals over 1 to 5 millimeters[32,33,60-62], implying that white matter electrodes may still capture activity from gray matter. Indeed, we show here that electrodes in the white matter reveal task-relevant properties and therefore included electrodes in the white matter in our analyses. To avoid confusion about the origin of the signals, we focused on the gray matter electrodes in the main text and reported results from electrodes in the white matter in the Supplementary Material. None of the conclusions in this study would change if we were to report the results from electrodes in the gray matter exclusively.

We built two generalized linear models (GLM) to characterize how the neural responses depended on the cognitive demands of each trial. The first model focused on the neural responses to the 1st tile, and the second model on the 2nd tile. In both cases, we focused on predicting the area under the curve (AUC) of the gamma band power (30-150 Hz) in each trial (**Methods**). For the first GLM, the time window started when the 1st tile was clicked and ended at a time corresponding to the 90th percentile of the distribution of reaction times (time difference between the 1st and the 2nd clicks, **Figure 1A**). This criterion was a reasonable tradeoff between minimizing overlap with responses after the 2nd tile and maximally capturing information before the 2nd tile. For the second model, the time window started with the 2nd click and ended one second afterward.

We considered 15 predictors for the GLM models, including whether a trial was a match or not, reaction time, n-since-last-click (nslc), and n-since-pair (nsp), the variables introduced in **Figures 1**-**2**. We also included additional predictors: first-click (whether a tile was clicked for the first time), n-times-seen (number of times an image had been seen), next-match (whether the subsequent trial was a match), board size, x and y position of the clicked tile within the board, distance between the first and the second tiles, and whether the image contained a person, animal, food, or vehicle. **Table 1** lists all the predictors and their definitions. Since several predictors were correlated with each other (**Figure S4A**-**B**), we computed the variance inflation factor (VIF)[63], a metric commonly used to account for correlations between predictors in generalized linear models. The VIF of each predictor was smaller than 3 for all participants (**Figure S4C**-**D**). Therefore, the correlations between predictors did not harm the performance of the models (**Methods**).

When the first tile in a trial was clicked, its status in memory guided the following actions. If it was a new image, the participant needed to encode it in memory for future retrieval. Thus, the ability to detect novelty is the first step for successful encoding. The predictor *first-click* described novelty and had a value of 1 whenever a tile was seen for the first time and 0 otherwise. If a tile had been viewed before, it would appear familiar to the participant, and the degree of familiarity depended on how long ago that tile had been seen last. The predictor n-since-last-click (**Figure 1A**-**B, Figure 2C**-**D**) captures the notion of familiarity; the smaller the nslc value, the more

familiar the tile is because that same tile was seen more recently and there were fewer competing stimuli encountered in between.

Figure 3A-D shows the neural activity of an electrode located in the right lateral orbitofrontal cortex (arrow in **Figure 3D**), whose responses to the first tile correlated with novelty. The GLM analysis indicated that first-click was a significant predictor of the neural responses (**Figure 3A**). The neural responses to the first tile showed a decrease in activity for novel tiles compared to tiles that had been seen before (**Figure 3B**), which could also be readily seen in individual trials (**Figure 3C**). This decrease in activity is reflected by the negative sign in the GLM first-click predictor (**Figure 3A**).

Novelty was a significant predictor of the neural responses after the first tile ($p<0.01$, GLM) for 50 electrodes in the gray matter (7.4% of the total, **Table S3A**, **Figure 3D**) and 33 electrodes in the white matter (6.7% of the total, **Table S3B**). The lateral orbitofrontal (LOF) cortex and pars opercularis contained significantly more electrodes than expected by chance ($p<0.01$, **Methods**).

Figure 3E-H shows the neural activity of an electrode located in the left pars opercularis (arrow in **Figure 3H**), whose responses to the first tile correlated with familiarity. The GLM analysis indicated that both n-since-last-click (nslc) and first-click were significant predictors of the neural responses ($p<0.001$, GLM, **Figure 3E**). Novel tiles (completely unfamiliar tiles, **Figure 3F**, blue) elicited strong responses, followed by less familiar tiles (higher nslc, **Figure 3F**, yellow). Familiar tiles (nslc=1, i.e., tiles that had just been seen in the preceding trial) elicited almost no response (**Figure 3F**, red). The strong correlation between the neural responses, novelty, and familiarity can also be readily appreciated in individual trials (**Figure 3G**).

The reaction time was also a significant predictor for the neural responses recorded from this electrode (**Figure 3E**). However, the differences in neural responses signaling novelty and distinct degrees of familiarity *cannot* be explained by differences in reaction time. The differences in neural responses associated with novelty and familiarity persisted after reaction time equalization (see vertical dashed lines indicating equalized RT in **Figure 3F**).

The nslc predictor was statistically significant ($p<0.01$, GLM) in 45 gray matter electrodes (6.7% of the total, **Table S4A**, **Figure 3H**) and 32 white matter electrodes (6.5%, **Table S4B**, **Figure S5D**). **Figure S5** shows an example electrode located in the white matter whose responses

correlated with novelty and familiarity. The majority of electrodes (82.2%) showed a positive correlation between the neural responses and nslc as illustrated in **Figure 3E-G**. The remaining electrodes (17.8%) showed a negative correlation, i.e., stronger neural responses for more familiar items. **Figure S6** depicts an example electrode located in the right pars opercularis showing a negative correlation between familiarity and neural responses.

Both the electrode in **Figure 3E**-**H** and the one in **Figure S6** revealed first-click as a significant predictor in addition to nslc, meaning that their responses not only reflected the familiarity gradient but also represented novelty. The electrodes that showed both first-click and n-since-last-click as significant predictors (20 electrodes) are denoted by red circles in **Figure 3D** and **Figure 3H**. Among these 20 electrodes, the signs of the t-statistic for the first-click and n-since-last-click predictors were consistent for 18 electrodes (16 positive and 2 negative). Only two electrodes exhibited opposite signs. These results indicate that novelty largely resembles extremely low familiarity in terms of the underlying neural responses.

Neural signals show anticipation of the trial's outcome

After seeing the 1st tile, participants attempt to find the tile's pair. If the 1st tile's pair was never encountered before, this is a random choice among the unseen tiles. If the tile's pair is unfamiliar, recalling a match is error-prone and often leads to mismatches (**Figure 2E**). For highly familiar cases, participants can retrieve the correct location to find the match tile. Therefore, we asked whether the neural responses after exposure to the first tile and before seeing the 2nd tile could predict successful retrieval.

Figure 4 shows the neural activity of an electrode located in the right lateral orbitofrontal cortex (arrow in **Figure 4E**), whose responses were predictive of successful retrieval. The GLM analysis indicated that match was a significant predictor of the neural responses ($p<0.001$, **Figure 4A**). This electrode showed stronger responses during match trials (**Figure 4B**, green) than during mismatch trials (**Figure 4B**, black). These differences can even be appreciated in single trials (compare **Figure 4C** versus **Figure 4D**). Of note, these differences are evident shortly after visualization of the first tile, with a peak at 500 ms after clicking the first tile, well before clicking

the second tile, when the participant did not know for certain yet whether the trial would be a match or not. Thus, the strong neural differences between match and mismatch trials reflect the participant's internal retrieval of the correct pairs' locations.

Whether a trial was a match or not was a significant predictor of the neural responses for 32 electrodes in the gray matter (4.7% of the total, **Table S5A**, **Figure 4E**) and 30 electrodes in the white matter (6% of the total, **Table S5B**, **Figure S7**). For an example electrode in the white matter see **Figure S7**. In most cases (91%), neural activity was higher during match trials than during mismatch trials, as illustrated in **Figure 4**. The locations of all these electrodes, shown in **Figure 4E** (gray matter) and **Figure S7** (white matter), reveal that the majority were located in the lateral orbitofrontal (LOF) cortex, the medial temporal lobe, and the insula. The LOF cortex contained significantly more electrodes than expected by chance ($p<0.01$, **Methods**).

The peak in neural activity occurred at approximately 500 ms after the 1st click (**Figure 4B**). As discussed in the previous section, the match predictor correlated with several other predictors (**Figure S4**). However, the GLM analysis shows that the match's presence, but not other predictors, accounts for the neural responses (**Figure 4A**). To further establish this point, **Figure S8A** shows the responses of this same electrode, in the same format as **Figure 4B**, after equalizing the n-since-last-click (nslc) distributions for match and mismatch trials by subsampling the data. The same conclusions hold in this case. Furthermore, **Figure S8B** shows each match trial's gamma power AUC versus the value of n-since-last-click and **Figure S8C** displays the same data from mismatch trials. The variable nslc did *not* account for the neural responses in either case ($p>0.18$, linear regression). Similar conclusions hold for the other predictors.

Figure S9 shows another example electrode located in the left middle temporal gyrus where the match was a significant predictor for the gamma band activity between the 1st and 2nd tiles. Similar to the LOF electrode in **Figure 4**, the gamma power during match trials was higher than during mismatch trials. However, the pattern of modulation in this electrode was sustained rather than transient (compare **Figure S9** versus **Figure 4B**-**D**). The change in gamma power was also evident in individual trials (**Figure S9B**-**C**). These observations suggest that the middle temporal and lateral orbitofrontal regions might be functionally distinct during memory retrieval.

In sum, these results indicate that even before the actual realization of whether a trial was a match or mismatch (i.e., before the onset of the 2nd tile), there were distinct neural responses that were predictive of the trial's outcome.

Neural signals reflect the strength of memory retrieval

In addition to reflecting the outcome of a given trial (match versus mismatch), we considered the *n-since-pair* (nsp) predictor as a proxy for the degree of confidence or strength of memory retrieval. The smaller the nsp, the more recently *the tile's pair* had been seen (**Figure 2E**-**F**). This predictor is different from n-since-last-click, which indicates how recently *the same tile*, rather than its pair, had been seen (**Figure 1**). We considered only match trials for this predictor (n-since-pair*match) because there was no successful retrieval of the tile's pair in mismatch trials.

Figure 5 shows an example electrode in the left middle temporal gyrus (arrow in **Figure 5E**). In contrast with the electrode in **Figure 4**, both the match predictor and the nsp predictor were significant in the GLM analysis (**Figure 5A**). The t-statistic for nsp was negative, indicating a *decrease* in gamma band power for matching pairs that were more distant in memory. Indeed, responses were strongest for those tiles whose pairs had been seen less than 2 clicks ago (**Figure 5B**, red) and weakest when matching pairs had been seen more than 10 clicks ago (**Figure 5B**, purple). There was a negative correlation between the area under the curve (AUC) of the gamma band power and nsp (**Figure 5C**, $p<0.001$, linear regression). This correlation disappeared when considering mismatch trials (**Figure 5D**, $p=0.66$, linear regression), suggesting that the relationship between the neural signals and memory strength was contingent on successful retrieval.

The nsp predictor was statistically significant ($p<0.01$, GLM) in 15 electrodes in the gray matter (2.2% of the total, **Table S6A**, **Figure 5E**) and 9 electrodes in the white matter (1.8% of the total, **Table S6B**, **Figure S10E**). For an example electrode located in the white matter, see **Figure S10**. Most of these electrodes showed a negative t-statistic, as in the example in **Figure 5**, and three electrodes (20%) showed the reverse effect (i.e., an increase in the neural signal for more

distant associative memories). For most of these electrodes (73.3%), *match* was also a significant predictor, as illustrated by the example in **Figure 5A**, indicating that the neural signals encoded *both* successful retrieval *and* memory strength.

Neural signals reflect feedback after the second tile

We have thus far focused on describing the responses elicited by the first tile in each trial. Next, we evaluated the neural responses triggered by the click of the second tile. We first asked whether novelty and familiarity were also encoded in the neural responses after the 2nd tile. We built a separate GLM using the same 15 predictors except for n-since-pair*match (**Table 1**) to describe the AUC of the gamma power during one second after clicking the 2nd tile. We excluded n-since-pair*match here because it would always be 1 during match trials (by definition, the first click was the pair of the second click). For the second tile, first-click was a significant predictor in 24 electrodes in the gray matter (3.6% of the total, **Table S7A**, **Figure S11E**) and 12 in the white matter (2.4% of the total, **Table S7B**). Thus, less than half the number of electrodes reflected novelty during the 2nd tile compared to the first tile (cf. **Table S7A** versus **Table S3A** and **Table S7B** versus **Table S3B**). For the second tile, *n-since-last-click* was a significant predictor in 9 electrodes in the gray matter (1.3% of the total, **Table S8A**, **Figure S11D**) and 24 in the white matter (4.9% of the total, **Table S8B**), again, less than half the number of electrodes reflecting familiarity during the first tile (cf. **Table S8A** versus **Table S4A** and **Table S8B** versus **Table S4B**).

An example electrode in the LOF region whose responses correlated with familiarity after the 2nd tile is shown in **Figure S11**. The LOF cortex contained significantly more electrodes than expected by chance ($p<0.01$, **Methods**). Among all the 9 electrodes where n-since-last-click was a significant predictor during the 2nd tile, 7 electrodes also had first-click as a significant predictor (**Figure S11D-E**, red circles). In sum, novelty and familiarity of a tile were still encoded in the neural responses to the second tile, but to a lesser degree than during the responses to the first tile. This reduction may be due to the fact that two images were presented simultaneously, and the neural signals might reflect a weighted combination of the responses to each[64]. Moreover,

for match trials, the information about the 2nd tile does not need to be encoded in memory anymore to thrive in the task.

Among the 83 electrodes that had first-click as a significant predictor during the 1st tile and the 36 electrodes during the 2nd tile (including both gray and white matter), 15 electrodes (11 gray matter + 4 white matter) overlapped, i.e., first-click was a significant predictor during both the 1st and the 2nd tiles. Among the 77 electrodes that had n-since-last-click as a significant predictor during the 1st tile and the 33 electrodes during the 2nd tile (including both gray and white matter), 20 electrodes (5 gray matter + 15 white matter) overlapped, i.e., nslc was a significant predictor during both the 1st and the 2nd tiles. These electrodes may reflect general rather than specific novelty or familiarity mechanisms, irrespective of tile order, image content, or location.

Next, we asked whether the differences between match and mismatch trials were also manifested *after* the 2nd tile was revealed, i.e., after the participant became explicitly aware of whether the trial was a match or not. **Figure 6** shows an example electrode located in the left insula (see arrow in **Figure 6E**) where match was a significant predictor for the neural responses after the 2nd tile ($p<0.001$, GLM, **Figure 6A**). The neural signals during match trials were larger than during mismatch trials (**Figure 6B**) and could be readily observed even in single trials (**Figure 6C** vs. **7D**).

After the 2nd tile, the match predictor was statistically significant ($p<0.01$, GLM) for 112 electrodes in the gray matter (16.6% of the total, **Table S9A**, **Figure 6E**) and 66 electrodes in the white matter (13.4% of the total, **Table S9B**, **Figure S12E**). For an example electrode in the white matter, see **Figure S12**. The locations of all these electrodes, shown in **Figure 6E** (gray matter) and **Figure S12E** (white matter), reveal that the majority were circumscribed to the LOF and the insula. The proportions of significant electrodes in both regions were higher than expected by chance ($p<0.01$, **Methods**).

There were 17 electrodes in the gray matter and 15 in the white matter where the match predictor was significant for both the 1st and the 2nd tiles. These electrodes represented 53.1% (gray matter) and 50% (white matter) of the electrodes that were significant according to the 1st tile, and 15.2% (gray matter) and 22.8% (white matter) of the electrodes that were significant

according to the 2nd tile. These electrodes were located in the lateral orbitofrontal cortex, the medial temporal lobe, and the insula (**Figure 4E**, **S6E**, **7E**, and **S11E**, red circles). The electrode in **Figure S12** exemplifies such responses (compare the difference between match and mismatch before the onset of the 2nd tile in **Figure S12B** versus **Figure 6A**). The electrode in **Figure S9** reveals a continuous enhancement for match trials after the 1st tile that was sustained and continued after the onset of the 2nd tile.

A machine learning classifier could predict matches in single trials

We evaluated whether the neural responses within a brain region could predict if a trial was a match or a mismatch (**Methods**). For this analysis, we considered only those brain regions with more than 12 electrodes combining all participants (**Methods**). **Figure 7** shows the average decoding performance of an SVM classifier after 200 iterations of 5-fold cross-validation. At each iteration, the SVM consisted of a binary classifier (match versus mismatch). The predictors were the PCA features extracted from the concatenated neural responses of electrodes within a particular brain region. The superior parietal gyrus and insula exhibited decoding accuracy above chance ($p<0.01$, **Methods**) during the 1st tile (**Figure 7A**). The lateral orbitofrontal and middle temporal cortex also showed accuracy above chance (>60%, **Figure 7A**), albeit not statistically significant. As expected based on the responses of individual electrodes, the decoding accuracy was higher after the 2nd tile compared to the 1st tile. After the 2nd tile, multiple brain regions showed accuracy above chance ($p<0.01$, **Figure 7B**). The lateral orbitofrontal cortex, insula, middle temporal gyrus, and pars opercularis showed the highest accuracy (>75%). Similar results were obtained when subsampling 12 electrodes during each iteration (**Figure S13**).

A computational model provides a first-order approximation to the behavioral and neural measurements

To further understand the mechanisms at play during the task, we built a computational model that focused on the storage and retrieval of information (**Figure 8**, **Methods**). The computational model consists of a Hebbian attractor neural network with all-to-all connectivity.

The units are divided into position units (the number equaling the number of tiles on the board) and label units (the number equaling the number of images on the board) (**Figure 8A**). The model has two main modes of operation: learning (**Figure 8B**), and inference (**Figure 8C-D**). After the first click, the model receives as input the label of the tile and its position. The activity of each unit evolves over time based on the external input and the weighted input from other units followed by a rectifying non-linearity and normalization (**Methods**, Equation 1). Concomitantly, the weights are updated in a Hebbian manner (**Methods**, Equation 2). During inference, the model selects the position unit with the maximum activation for the second click. The model proceeds in this manner until all matches have been found.

We computed the same performance evaluators from **Figure 2A**,**B**,**D,F** for the model. We did not compute the metrics for **Figure 2C,E** because the model chooses the first tile randomly among the available tiles (**Methods**). We defined the reaction time as the number of steps needed for the selected unit to reach 0.9 of its maximum value (**Methods**). The number of clicks per tile increased with the board size, approximating the participants' behavior (compare **Figure 9A** versus **Figure 2A**). The reaction time for the model was longer for mismatch trials than for match trials for all board sizes ($p<0.001$, compare **Figure 9B** versus **Figure 2B**). The nslc value increased with board size and was much larger for mismatch trials compared to match trials, consistent with the participants' behavior ($p<0.001$, compare **Figure 9C** versus **Figure 2D**). Similarly, the nsp value increased with board size and was also significantly larger in mismatch trials compared to match trials for all board sizes ($p<0.001$, compare **Figure 9D** versus **Figure 2F**).

To investigate the model's inner workings, we defined two metrics based on the unit activations. First, to compare with the match related signals in **Figure 3**, we computed an overall maximum energy (**Methods**, Equation 3). This maximum energy was smaller for trials with nslc=1 ($p<0.001$, **Figure 9E**), reflecting a strong correlate of memory for recently seen tiles (compare to **Figure 3B** and especially **Figure 3F**). Second, we defined a confidence metric by assessing the relative activation for the strongest unit with respect to the other units during the inference step (**Methods**). The confidence metric was significantly larger for match trials compared to non-match trials ($p<0.01$, **Figure 9F**), which was qualitatively similar to the neural responses described in **Figure 5**.

Discussion

In this study, we investigated the dynamics of neural signals during a natural memory task where participants engaged in a classical card-matching game. Participants performed the task well, slightly worse than expected by a perfect memory model (**Figure 2A**), showing increased reaction times during mismatch trials (**Figure 2B**) as well as decay of memory traces with time since encoding (**Figure 2C-F**).

Many studies have focused on studying memory for a list of sequentially presented stimuli [1,7,12,19,24,65,66] or examined memory in naturalistic or real-world scenarios at the behavioral level [67-71]. The task introduced here strikes a balance between these two approaches, presenting a more realistic and complex setting that involves associative and non-associative memories within the same task and introducing task-dependent complexity compared to word lists. Yet, our task allows for a high level of control over stimulus timing and experimental parameters that are difficult to achieve when studying memory in real-world scenarios.

Complex and natural tasks necessarily depend on the interplay of multiple intercorrelated variables. To tame the complexity of these different variables, we used generalized linear models (GLMs) to quantitatively assess the influence of distinct predictors on the neural responses. Through these GLM analyses, we could characterize neurophysiological responses that were largely governed by individual predictors after accounting for the correlations among predictors (**Figure S4**). While focusing on any one predictor, these results were corroborated by subsampling the data to equalize other predictor variables that could potentially affect the neural responses. The extensive sampling of brain regions, including neural activity from more than 1,000 electrodes across 20 participants (**Figure S2-S3**), allowed us to track neural responses during each of the steps required in the task with broad brain coverage.

The first step to encode tile information is to correctly determine whether a tile is novel or not. We refer to the ability to detect novelty and familiarity as non-associative recognition memory[16,24]. Assessment of novelty and familiarity orchestrate strategies of encoding and maintenance of information in memory[1,72]. We found strong neural responses that signal novelty

and familiarity (**Figure 3**, **Figure S5-6**, **S10-11**, **Table S3**, **S4**, **S7** and **S8**). Several electrodes responded both to novelty and familiarity, showing similar responses to novel and highly unfamiliar items (**Figure 3F-G** and **Figure S5B-C**). The lateral orbitofrontal cortex, the pars opercularis, and the medial temporal lobe contained a preponderance of electrodes signaling novelty and familiarity. These responses are reminiscent of novelty and familiarity signals that have typically been described in tasks involving a sequence of images presented with occasional repetitions[6,16,18,66] but are not restricted to the medial temporal lobe. While some electrodes may also encode information about the image content (**Figure 3E**), most of the electrodes signaling novelty and familiarity were not content-specific, consistent with the observation that neurons involved in memory formation are rarely sharply tuned to particular sensory features[28].

After seeing the first tile in a trial, participants could estimate whether they remembered the location of its pair and, thus, whether they would get a match or not. Participants link the first tile with its pair to internally predict where to click next. Indeed, neural responses strongly reflected not only these predictions (**Figure 4**, **Figures S7**-**9**) but also the internal estimate of the memory strength or the confidence of these predictions (**Figure 5**). Even though participants could not know for sure yet whether the trial would be a match or not, the neural signals exhibited large differences between match and mismatch trials after the first tile and before the revelation of the second tile. These differences were either transient (**Figure 4**) or sustained (**Figure S9**). We speculate that transient increases in activity during match trials might signal the sudden realization and high confidence about the trial outcome (match or mismatch). In contrast, sustained responses may correspond to the active retrieval processes. It is possible that sustained activity could arise due to averaging across temporally shifted transient activities from individual neurons[73]; however, it has been reported that single neurons in the hippocampus can also show sustained firing rate increase for successful associative retrieval[74]. Extensive work has documented the importance of the hippocampus and surrounding structures in the MTL in associative memory (e.g.,[16,26,74,75]). In addition to the MTL, the current results show that other areas, such as the lateral orbitofrontal cortex, also play a critical role in associative recall.

Several studies have highlighted the potential for attractor-based models to characterize working memory processes[47,49,76,77]. Here we show as a proof-of-principle that a simple

instantiation of an attractor-based neural network model can qualitatively capture the properties of human participants both at the behavioral level (**Figure 9A-D**) and also at the neural level (**Figure 9E-F**). This basic neural network architecture can be readily linked to a visual neural network backbone to further examine the underlying representation of visual signals in working memory. Additionally, the model could be extended to examine even more complex tasks that involve multiple-way associations and dynamic changes in the structure of the environment over time. The high temporal resolution, extensive spatial sampling, and computational models provide an opportunity to characterize the dynamics of complex naturalistic tasks. These observations provide initial steps to further our understanding of how different components of encoding and retrieval interact during the formation of natural memory events.

Methods

Task paradigm

Participants performed our implementation of the classical memory matching game (**Figure 1, Movie S1**). The game involves remembering the location and content of a set of tiles to find all the matching pairs. A square board containing $n \times n$ tiles was shown throughout each block. In the beginning, all tiles were shown in black. In each trial, participants chose one tile, and then a second tile, by clicking on them in a self-paced fashion. Upon clicking, the tile revealed a common object like a cat or an indoor scene like a kitchen. At the end of each trial, either the two tiles revealed the same content (match) or not (mismatch). If the tiles matched, then the two tiles turned green 1,000 ms after the second click, and the two tiles could not be clicked again for the remainder of the block. If the tiles did not match, they turned black 1,000 ms after the second click and could be clicked again in subsequent trials. When all tiles turned green, i.e., all matches were found, the block ended, and another block began. During each block, the map between positions and objects was fixed. The game always started with a block of size 3×3 and progressed to more difficult blocks (4×4, 5×5, 6×6, and finally 7×7). Blocks with an odd number of tiles (3×3, 5×5, and 7×7) contained one distractor object (a human face) with no corresponding pair. For each block except the 3×3 board, there was a limit for the total time elapsed (2 minutes for 4×4,

3.3 min for 5×5, 4.8 min for 6×6, and 8.2 min for 7x7). If a participant did not complete a block within the time limit, the block ended, and a new, easier block started by reducing the board size *n* by 1, except when *n*=7, where it was reduced by 2. Conversely, when participants successfully completed a block with a board of size *n* within the allotted time limit, they moved on to a more difficult block by increasing *n* by 1. When participants completed an *n*=7 block, they performed further *n*=7 blocks. There was no image repetition across blocks.

All the images were from the Microsoft COCO 2017 validation dataset [78] and were rendered in grayscale and square shape. We included a balanced number of pictures from 5 categories: person, animal, food, vehicle, and indoor scenes. All the images were rendered on a 13-inch Apple MacBook Pro laptop. The size of each tile was 0.75×0.75 inches (approximately 2x2 degrees of visual angle, dva) and the separation between two adjacent tiles was 0.125 inch (0.33 dva) for board size *n*=7 and 0.25 inch (0.67 dva) for the others. The game implementation was written and presented using the Psychtoolbox extension [79,80] in Matlab_2016b (Mathworks, Natick, MA).

Epilepsy participants and recording procedures

We recorded intracranial field potentials from 20 patients with pharmacologically intractable epilepsy (12-52 years old, 9 female, **Table S1**) undergoing monitoring at Boston Children's Hospital (Boston, US), Brigham and Women's Hospital (Boston, US), and Xuanwu Hospital (Beijing, China). All recording sessions were seizure-free. All patients had normal or corrected-to-normal vision. The study protocol was approved by each hospital's institutional review board. Experiments were run under patients' or their legal guardians' informed consent. One patient at Brigham and Women's Hospital (BWH) was implanted with both stereo encephalography (sEEG) and electrocorticography (ECoG) electrodes, while all other patients had only sEEG electrodes (Ad-tech, USA; ALCIS, France). Intracranial field potentials were recorded with Natus (Pleasanton, CA) and Micromed (Italy). The sampling rate was 2048 Hz at Boston Children's Hospital (BCH), 512 Hz or 1024 Hz at BWH, and 512 Hz at Xuanwu Hospital (XWH). Electrode trajectories were determined based on clinical purposes for precisely localizing suspected epileptogenic foci and surgically treating epilepsy[81].

Eye tracking procedures

Ten non-epilepsy healthy participants (23-35 years old, 9 female) performed the same task while their eye movements were tracked with the EyeLink 1000 plus system (SR Research, Canada) at a sampling rate of 500 Hz. The task paradigm was the same as the one for epilepsy participants except that, before each block began, participants fixated on a center cross to ensure that the EyeLink eye-tracking system was well-calibrated. Otherwise, a re-calibration session ensued. The task was presented on a 19-inch CRT monitor (Sony Multiscan G520), and participants sat about 21 inches away from the monitor screen. The tile size was 1x1 inches (approximately 2.7x2.7 degrees of visual angle) as appeared on the screen. The study protocol was approved by the institutional review board at Boston Children's Hospital, and each participant completed the task with informed consent. All participants had normal or corrected-to-normal vision. All participants completed 16 blocks.

Behavioral analyses

We created two computational models to simulate behavior assuming perfect memory or no memory (chance performance, **Figure 2A**). The perfect memory model remembered all revealed tiles without forgetting. The random model simulated random clicking. We calculated the reaction time (RT, time between two clicks in a trial), n-since-pair (number of clicks since the last time when a tile's matching pair was seen), n-since-last-click (the number of clicks since the same tile was clicked), and n-times-seen (number of times the same image had been seen). For n-since-pair and n-since-last-click, we excluded trials in which any tile was seen for the first time, i.e., when a tile's matching pair had never been revealed, or there was no previous click. We compared these variables for match and mismatch trials at each board size (**Figure 2**, permutation test, 5,000 iterations, α=0.01). We defined random matches as a match trial where the second tile had never been seen before; such trials were excluded from both the behavioral and neurophysiological analyses. We used the F-test for linear regression models to assess whether RT, n-since-pair, n-since-last-click, and n-times-seen significantly covary with board size. The linear regression models' predictors were these four behavioral parameters and the

dependent variable the board size. We created separate models for match and mismatch trials and 1st and 2nd tiles.

Electrode localization

Electrodes were localized using the iELVis[82] toolbox. We used Freesurfer[83] to segment the preimplant magnetic resonance (MR) images, upon which post-implant CT was rigidly registered. Electrodes were marked in the CT aligned to preimplant MRI using Bioimage Suite[84]. Each electrode was assigned to an anatomical location using the Desikan-Killiany[85] atlas for subdural grids or strips or FreeSurfer's volumetric brain segmentation for depth electrodes. For white matter electrodes, we also reported their closest gray matter locations. Out of 1,750 electrodes in total, we included 676 bipolarly referenced electrodes in the gray matter (**Figure S2, Table S2**) and 492 bipolarly referenced electrodes in the white matter (**Figure S3, Table S2**). Five hundred and eighty-two electrodes were not considered for analyses due to bipolar referencing, locations in pathological sites, or electrodes containing large artifacts. Electrode locations were mapped onto the MNI305 average brain via affine transformation[86] for display purposes (e.g., **Figure S2-S3**).

Preprocessing of intracranial field potential data

Bipolar subtraction was applied to each pair of neighboring electrodes on each shank of depth electrodes or subdural grids/strips[87]. A zero-phase digital notch filter (Matlab function "filtfilt") was applied to the bipolarly subtracted broadband signals to remove the line frequency at 60 Hz (BCH, BWH) or 50 Hz (Xuanwu) and their harmonics. For each electrode, trials whose amplitudes ($Voltage_{max}$-$Voltage_{min}$) were larger than 5 standard deviations from the mean amplitude across all trials were considered potential artifacts and discarded from further analyses [88]. For the first tile, the time window for artifact rejection was from 400 ms before the click until 1 second after the average RT. For the second tile, the time window was [400 ms + average RT] before the second click until 1 second after the second click. Across all electrodes, we rejected 1.75% of all trials for the 1st tile and 1.73% for the second tile.

Time-frequency decomposition

The gamma band (30-150 Hz) power was computed using the Chronux toolbox[89]. We used a time-bandwidth product of 5 and 7 leading tapers, a moving window size of 200 ms, and a step size of 10 ms[90]. For each trial, the power was normalized by subtracting the mean gamma band power during the baseline (400 ms before 1st tile) and dividing by the standard deviation of the gamma power during the baseline. For all the participants, there were more mismatch than match trials. In the raster plots, we subsampled the mismatch trials, keeping those trials whose reaction times were closest to the mean reaction time of match trials. All random matches were excluded from analyses.

Generalized linear model

We used generalized linear models (GLM)[91,92] to analyze the relationship between the gamma band power and behavioral parameters. We used two different GLMs, one using neural responses between the 1st and the 2nd tiles and the other using neural responses after the 2nd tile. For the first GLM, the time window started when the 1st tile was clicked and ended at a time corresponding to the 90th percentile of the distribution of reaction times (time difference between the 1st and the 2nd click, **Figure 1A**) for each participant. This criterion was a reasonable tradeoff between minimum overlap with responses after the 2nd tile and the maximum amount of information captured. For the second GLM, we used 1 second after the 2nd tile click as the analysis window. The response variable to be fit by the GLM analyses was defined as the area under the curve (AUC) of the gamma band power over the specified time windows. **Table 1** describes the behavioral parameters that were considered as predictors in the models.

We performed a multicollinearity analysis to assess the presence of highly correlated predictors that could impair the model's performance[93,94]. We calculated the variance inflation factor (VIF) for each predictor to detect the presence of multicollinearities. A VIF of 1 indicates that there is no correlation with other predictors. The larger the VIF, the higher the correlation. A VIF greater than 5 indicates a very high correlation that could significantly harm the model's performance. For all participants in our analysis, the VIFs of all predictors were smaller than 3 (**Figure S4C-D**).

For n-since-pair, we included the interaction term between this predictor and match (n-since-pair*match) to test the hypothesis that when a trial was a match, the strength of the neural response after the first tile was modulated by how recently the tile's matching pair was seen for the last time. The neurophysiological responses confirmed that this is a reasonable way to model the data (**Figure 5B**). We represented image categories as predictor variables in the GLM by including four out of the five categories (animal, food, vehicle, and person). We dropped the "indoor" category to avoid falling into the "dummy variable trap"[95]. For each predictor, we calculated the parameter estimate (beta coefficient) from the least mean squares fit of the model to the data, the t-statistic (beta divided by its standard error), and the p-value to test the effect of each predictor on the neural responses. A beta coefficient or t-statistic of zero indicated that the predictor did not affect the neural responses. A predictor was considered statistically significant if the GLM model differed from a constant model ($p<0.01$) and the p-value for that predictor was smaller than 0.01.

To determine if any brain region contained significantly more electrodes than expected by chance considering any GLM predictor, we randomly sampled the same number (n) of electrodes as those where that predictor was significant, from the total electrode population (separately for gray and white matter electrodes) for 5,000 iterations. Taking the *match* predictor for gray matter electrodes as an example, from the total of 676 gray matter electrode locations, we randomly sampled 32 electrodes (the number of match-significant gray matter electrodes) 5,000 times, and calculated the p-value of any location, say insula, as the number of times when n for insula and match-significant was less than n for insula sampled. If $p<0.01$, that brain region was considered to have significantly more electrodes than expected by chance.

Decoding of match

We used a machine learning decoding approach to evaluate whether the neural responses from a given brain region could predict whether the trial was a match or a mismatch. For this analysis, we selected only those brain regions in which we had at least 12 electrodes. We used two different decoders for each brain region, one using neural responses between the 1st tile and 800 ms after the 1st tile click, and the other using the neural responses between the 2nd

tile and 800 ms after the 2nd tile click. We performed 200 iterations of 5-fold cross-validation for each brain region and tile to split the trials into independent train and test sets (1,000 splits in total). We concatenated the neural responses of electrodes within the same brain region, using two different approaches: (i) taking all electrodes in each brain region, and (ii) randomly subsampling 12 electrodes at each iteration. The number of match and mismatch trials was normalized by random subsampling. To reduce the dimensionality of the neural responses, we used principal component analysis (PCA). The PCA parameters were computed using only the training data, and we selected the number of components that could explain 70% of the training neural responses variance. We used support vector machines (SVM) with a linear kernel for the binary classification (match or mismatch), using as the model inputs the PCA features computed from the neural responses. We followed the same procedure to test whether the classification performance was above chance, but we randomized the response variable (match or mismatch) at every iteration. We calculated the p-value as the number of times when the accuracy using random labels was above the average accuracy using the actual labels. If $p<0.01$, neural responses of that brain region could predict match and mismatch above chance.

Computational model

We developed an attractor network model consisting of a fully connected recurrent network with the number of units n equal to the number of tiles in the grid plus the number of different images. For example, the model for the 3x3 board shown in **Fig 1A** was an attractor network with n=3x3+5=14 units (**Figure 8**).

The units in the network were designed to model "where" and "what", i.e., position and image labels. Let $\boldsymbol{x}_p$ be a vector of length equal to the number of tiles in the grid, $\boldsymbol{x}_l$ be a vector of length equal to the number of different images in the grid, and $\boldsymbol{x}$ denote the concatenation $[\boldsymbol{x}_p, \boldsymbol{x}_l]$ (**Figure 8A**). The input to the network is $\boldsymbol{x}$. Each entry in $\boldsymbol{x}_p$ and $\boldsymbol{x}_l$ can take the values -1, 0, or 1. The state of the network at time t is denoted by the vector $\boldsymbol{h}_t = [\boldsymbol{p}_t, \boldsymbol{l}_t]$ of size n, where $\boldsymbol{p}_t$ and $\boldsymbol{l}_t$ are the vectors of activations of the position and label units, respectively. Each entry in $\boldsymbol{h}_t$ is a scalar value. The units in the network are connected in an all-to-all fashion and the matrix $\boldsymbol{M}_t$ indicates the weights at time t ($\boldsymbol{M}_t \in \mathbb{R}^{n \times n}$).

The network stores memories in both persistent activities (active representations) and weights (silent representations)[49]. In contrast to the approach in ref.[49], which incorporates a bottleneck in the model to restrict its capacity, our model is devoid of any such bottleneck. Given an input $\boldsymbol{x}$ at time t, the network state and weights were updated similarly to ref.[96], according to:

$$\boldsymbol{h_t} = f\left(\mathcal{N}(\boldsymbol{x} + \boldsymbol{M}_{t-1}\boldsymbol{h}_t)\right) \quad \text{Equation 1}$$

$$\boldsymbol{M}_t = \lambda \boldsymbol{M}_{t-1} + \eta \boldsymbol{h}_t \boldsymbol{h}_t^T \quad \text{Equation 2}$$

Here $f(\cdot)$ is the LeakyReLU activation function and $\mathcal{N}(\cdot)$ is activation normalization. λ and η represent a decay rate for the previously stored memories and the learning rate for new memories, respectively. In line with ref.[96], activation normalization is expected to make the network more robust to the choice of the decay and learning rates. We note that the Hebbian learning is computed on the state of the network $\boldsymbol{h}_t$ rather than on the input $\boldsymbol{x}$. This means that the update of the memory matrix $\boldsymbol{M}_t$ is influenced by the interference between active and silent representations, thus limiting the network capacity. The hyperparameters were chosen by fitting the number of clicks per tile of the model to the participants' number of clicks per tile (**Figure 2A**). The results presented in this paper were obtained with $\lambda = 0.6$ and $\eta = 0.9$. Before the start of each board, the network weights were initialized uniformly at random in $[0,1]$, while the state of the network was initialized to 0. Changes in weights in neural networks are often interpreted as structural modifications to synaptic strengths. However, given the time scales involved in working memory tasks such as the one studied here, changes in $\boldsymbol{M}_t$ are more likely to reflect transient biophysical mechanisms such as synaptic facilitation rather than permanent structural synaptic modifications.

The model operates in two distinct regimes, which we refer to as *learning* (**Figure 8B**) and *inference* (**Figure 8C-D**). For each trial, the 1st tile was chosen at random among the available tiles. To simulate the task, for each trial the model performs learning→inference→learning. First, the model represents the position and label of the 1st tile. Second, the model performs inference on the label of the 1st tile. At the end of the inference regime, the most active neuron in $\boldsymbol{p}_t$ determines which tile to click (**Figure 8D**). Last, the model learns the position and label of the 2nd tile.

During learning (**Figure 8B**), the corresponding position entry of $\boldsymbol{x}_p$ is set to 1 and all other units are set to -1. Similarly, the corresponding label entry of $\boldsymbol{x}_l$ is set to 1 and all other units are set to -1. The network dynamics goes through 10 steps according to the two equations above. During inference (**Figure 8C-D**), the corresponding label of $\boldsymbol{x}_l$ is set to 1 and all the other units are set to 0. All the units of $\boldsymbol{x}_p$ corresponding to the available tiles are set to 0, while the ones corresponding to the unavailable tiles (those that have already been matched or already clicked in that trial) are set to -1. The network dynamics goes through 10 steps according to Equations 1-2. After these 10 steps, we select the unit with the maximum activation within the units of $\boldsymbol{x}_p$ corresponding to available tiles. If the second tile is a match, then those two tiles become unavailable in the next trials. The weight matrix $\boldsymbol{M_t}$, however, continues to include all the connections among all the units. The model proceeds until all tiles have been matched.

The number of clicks per tile, n-since-last-click and n-since-pair click for the 2nd tile were calculated for the model and compared to the participants' behavior (**Figure 9A-D**). To compute a proxy for the reaction time in the model, we used the same approach as in ref.[49], whereby the unit in $\boldsymbol{x}_p$ with the strongest activation during the inference time was selected and the reaction time was computed as the number of steps the unit takes to reach 0.9 of its maximum value.

To compare the inner workings of the model with the neural data, we defined two new metrics based on the unit activations. First, we defined the *max-energy* metric computed during the 1st learning phase of each trial, in analogy to the memory signals in **Figure 3**. The energy of the network was computed as:

$$E_t = -\boldsymbol{h}_t \boldsymbol{M}_t \boldsymbol{h}_t^T \qquad \text{Equation 3}$$

Min-max normalization was applied to the energy in each trial, and the maximum value in each trial was reported. The max-energy metric is shown in **Figure 9E**. Second, we defined a *confidence* metric that reflected the evidence for a match in a given trial, in analogy with the predictive signals shown in **Figure 5**. The confidence metric was defined by selecting the strongest activation in $\boldsymbol{p}_t$ during inference, subtracting the mean value of $\boldsymbol{p}_t$, applying min-max normalization to the difference, and then taking the maximum over time t in each trial. The confidence metric is shown in **Figure 9F**.

Data availability

We share all the deidentified psychophysics data, electrode location information, and neural data, together with all the code generated to model and analyze the data through the following public site: https://klab.tch.harvard.edu/resources/HowToGetAMatch.html

Author contributions: The task was designed by YX and GK. All the data were collected by RJW (XWH) and YX (BWH and BCH) with the help of PHW (XWH) and DW (BWH). GGZ (XWH), YZS (XWH), CRG (BWH), JRM (BCH), and SS (BCH) performed the surgeries on the patients. All the data were curated by YX and analyzed by YX and PSL, with frequent discussions with GK. The computational model was developed by RS, with frequent discussions with GK. The manuscript was written by YX, PSL, RS, and GK and approved by all authors.

Acknowledgments

This work was supported by the McKnight Foundation, NIH Grant R01026025, and NSF Grant CCF-1231216.

References

1 Tulving, E. & Kroll, N. Novelty assessment in the brain and long-term memory encoding. *Psychon Bull Rev* **2**, 387-390 (1995). https://doi.org:10.3758/BF03210977

2 Duncan, K. D. & Shohamy, D. Memory states influence value-based decisions. *J Exp Psychol Gen* **145**, 1420-1426 (2016). https://doi.org:10.1037/xge0000231

3 Montaldi, D., Spencer, T. J., Roberts, N. & Mayes, A. R. The neural system that mediates familiarity memory. *Hippocampus* **16**, 504-520 (2006). https://doi.org:10.1002/hipo.20178

4 Mehrpour, V., Meyer, T., Simoncelli, E. P. & Rust, N. C. Pinpointing the neural signatures of single-exposure visual recognition memory. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org:10.1073/pnas.2021660118

5 Park, J. *et al.* Role of low- and high-frequency oscillations in the human hippocampus for encoding environmental novelty during a spatial navigation task. *Hippocampus* **24**, 1341-1352 (2014). https://doi.org:10.1002/hipo.22315

6 Yassa, M. A. & Stark, C. E. Multiple signals of recognition memory in the medial temporal lobe. *Hippocampus* **18**, 945-954 (2008). https://doi.org:10.1002/hipo.20452

7 Rutishauser, U., Ross, I. B., Mamelak, A. N. & Schuman, E. M. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature* **464**, 903-907 (2010). https://doi.org:10.1038/nature08860

8 Daselaar, S. M., Fleck, M. S. & Cabeza, R. Triple dissociation in the medial temporal lobes: recollection, familiarity, and novelty. *J Neurophysiol* **96**, 1902-1911 (2006). https://doi.org:10.1152/jn.01029.2005

9 Fried, I., MacDonald, K. A. & Wilson, C. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* **18**, 753-765 (1997).

10 Murray, R. J., Brosch, T. & Sander, D. The functional profile of the human amygdala in affective processing: insights from intracranial recordings. *Cortex* **60**, 10-33 (2014). https://doi.org:10.1016/j.cortex.2014.06.010

11 Zaehle, T. *et al.* Nucleus accumbens activity dissociates different forms of salience: evidence from human intracranial recordings. *J Neurosci* **33**, 8764-8771 (2013). https://doi.org:10.1523/JNEUROSCI.5276-12.2013

12 Viskontas, I. V., Knowlton, B. J., Steinmetz, P. N. & Fried, I. Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal regions. *J Cogn Neurosci* **18**, 1654-1662 (2006). https://doi.org:10.1162/jocn.2006.18.10.1654

13 Brown, M. & Aggleton, J. Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience* **2**, 51-61 (2001). https://doi.org:10.1038/35049064

14 Kucewicz, M. T. *et al.* High frequency oscillations are associated with cognitive processing in human recognition memory. *Brain* **137**, 2231-2244 (2014). https://doi.org:10.1093/brain/awu149

15 Zheng, J. *et al.* Neurons detect cognitive boundaries to structure episodic memories in humans. *Nature Neuroscience* **25**, 358-368 (2022).

16 Rutishauser, U. Testing Models of Human Declarative Memory at the Single-Neuron Level. *Trends Cogn Sci* **23**, 510-524 (2019). https://doi.org:10.1016/j.tics.2019.03.006

17 Johnson, E. L. & Knight, R. T. Intracranial recordings and human memory. *Curr Opin Neurobiol* **31**, 18-25 (2015). https://doi.org:10.1016/j.conb.2014.07.021

18 Knight, R. Contribution of human hippocampal region to novelty detection. *Nature* **383**, 256-259 (1996). https://doi.org:10.1038/383256a0

19 Sederberg, P. B. *et al.* Hippocampal and neocortical gamma oscillations predict memory formation in humans. *Cereb Cortex* **17**, 1190-1196 (2007). https://doi.org:bhl030 [pii]
10.1093/cercor/bhl030

20 Roediger, H. L., 3rd & Tekin, E. Recognition memory: Tulving's contributions and some new findings. *Neuropsychologia* **139**, 107350 (2020). https://doi.org:10.1016/j.neuropsychologia.2020.107350

21 Ison, M. J., Quian Quiroga, R. & Fried, I. Rapid Encoding of New Memories by Individual Neurons in the Human Brain. *Neuron* **87**, 220-230 (2015). https://doi.org:10.1016/j.neuron.2015.06.016

22 Sheehan, T. C., Sreekumar, V., Inati, S. K. & Zaghloul, K. A. Signal Complexity of Human Intracranial EEG Tracks Successful Associative-Memory Formation across Individuals. *J Neurosci* **38**, 1744-1755 (2018). https://doi.org:10.1523/JNEUROSCI.2389-17.2017

23 Wirth, S. *et al.* Single neurons in the monkey hippocampus and learning of new associations. *Science* **300**, 1578-1581. (2003).
24 Kirwan, C. B. & Stark, C. E. Medial temporal lobe activation during encoding and retrieval of novel face-name pairs. *Hippocampus* **14**, 919-930 (2004). https://doi.org:10.1002/hipo.20014
25 Ranganath, C., Cohen, M. X., Dam, C. & D'Esposito, M. Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. *J Neurosci* **24**, 3917-3925 (2004). https://doi.org:10.1523/JNEUROSCI.5053-03.2004
26 Sakai, K. & Miyashita, Y. Neural organization for the long-term memory of paired associates. *Nature* **354**, 152-155 (1991).
27 Zhou, Y. D., Ardestani, A. & Fuster, J. M. Distributed and associative working memory. *Cereb Cortex* **17 Suppl 1**, i77-87 (2007). https://doi.org:10.1093/cercor/bhm106
28 Rutishauser, U., Reddy, L., Mormann, F. & Sarnthein, J. The Architecture of Human Memory: Insights from Human Single-Neuron Recordings. *J Neurosci* **41**, 883-890 (2021). https://doi.org:10.1523/JNEUROSCI.1648-20.2020
29 Mayes, A., Montaldi, D. & Migo, E. Associative memory and the medial temporal lobes. *Trends Cogn Sci* **11**, 126-135 (2007). https://doi.org:10.1016/j.tics.2006.12.003
30 Van Petten, C., Luka, B. J., Rubin, S. R. & Ryan, J. P. Frontal brain activity predicts individual performance in an associative memory exclusion test. *Cereb Cortex* **12**, 1180-1192 (2002). https://doi.org:10.1093/cercor/12.11.1180
31 Jensen, O., Kaiser, J. & Lachaux, J. P. Human gamma-frequency oscillations associated with attention and memory. *Trends Neurosci* **30**, 317-324 (2007). https://doi.org:10.1016/j.tins.2007.05.001
32 Buzsaki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents - EEG, ECoG, LFP and spikes. *Nat Rev Neurosci* **13**, 407-420 (2012). https://doi.org:nrn3241 [pii]
10.1038/nrn3241
33 Kreiman, G. *et al.* Object selectivity of local field potentials and spikes in the inferior temporal cortex of macaque monkeys. *Neuron* **49**, 433-445 (2006).
34 Mukamel, R. *et al.* Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* **309**, 951-954 (2005).
35 Fries, P., Reynolds, J., Rorie, A. & Desimone, R. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **23**, 1560-1563 (2001).
36 Baddeley, A. Working memory. *Curr Biol* **20**, R136-140 (2010). https://doi.org:10.1016/j.cub.2009.12.014
37 Barak, O. & Tsodyks, M. Working models of working memory. *Curr Opin Neurobiol* **25**, 20-24 (2014). https://doi.org:10.1016/j.conb.2013.10.008
38 Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb Cortex* **10**, 910-923 (2000). https://doi.org:10.1093/cercor/10.9.910
39 Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn Sci* **7**, 415-423 (2003). https://doi.org:10.1016/s1364-6613(03)00197-9

40 Duarte, R., Seeholzer, A., Zilles, K. & Morrison, A. Synaptic patterning and the timescales of cortical dynamics. *Curr Opin Neurobiol* **43**, 156-165 (2017). https://doi.org:10.1016/j.conb.2017.02.007

41 Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. Neurocomputational models of working memory. *Nat Neurosci* **3 Suppl**, 1184-1191 (2000). https://doi.org:10.1038/81460

42 Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol* **61**, 331-349 (1989). https://doi.org:10.1152/jn.1989.61.2.331

43 Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652-654 (1971). https://doi.org:10.1126/science.173.3997.652

44 Nakamura, K. & Kubota, K. Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. *J Neurophysiol* **74**, 162-178 (1995). https://doi.org:10.1152/jn.1995.74.1.162

45 Watanabe, K. & Funahashi, S. Prefrontal delay-period activity reflects the decision process of a saccade direction during a free-choice ODR task. *Cereb Cortex* **17 Suppl 1**, i88-100 (2007). https://doi.org:10.1093/cercor/bhm102

46 Almeida, R., Barbosa, J. & Compte, A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J Neurophysiol* **114**, 1806-1818 (2015). https://doi.org:10.1152/jn.00362.2015

47 Lansner, A., Marklund, P., Sikstrom, S. & Nilsson, L. G. Reactivation in working memory: an attractor network model of free recall. *PLoS One* **8**, e73776 (2013). https://doi.org:10.1371/journal.pone.0073776

48 Macoveanu, J., Klingberg, T. & Tegner, J. A biophysical model of multiple-item working memory: a computational and neuroimaging study. *Neuroscience* **141**, 1611-1618 (2006). https://doi.org:10.1016/j.neuroscience.2006.04.080

49 Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P. & Husain, M. Neural mechanisms of attending to items in working memory. *Neurosci Biobehav Rev* **101**, 1-12 (2019). https://doi.org:10.1016/j.neubiorev.2019.03.017

50 Roggeman, C., Klingberg, T., Feenstra, H. E., Compte, A. & Almeida, R. Trade-off between capacity and precision in visuospatial working memory. *J Cogn Neurosci* **26**, 211-222 (2014). https://doi.org:10.1162/jocn_a_00485

51 Seeholzer, A., Deger, M. & Gerstner, W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Comput Biol* **15**, e1006928 (2019). https://doi.org:10.1371/journal.pcbi.1006928

52 Wei, Z., Wang, X. J. & Wang, D. H. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *J Neurosci* **32**, 11228-11240 (2012). https://doi.org:10.1523/JNEUROSCI.0735-12.2012

53 Fiebig, F. & Lansner, A. A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *J Neurosci* **37**, 83-96 (2017). https://doi.org:10.1523/JNEUROSCI.1989-16.2016

54 Mi, Y., Katkov, M. & Tsodyks, M. Synaptic Correlates of Working Memory Capacity. *Neuron* **93**, 323-330 (2017). https://doi.org:10.1016/j.neuron.2016.12.004

55 Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543-1546 (2008). https://doi.org:10.1126/science.1150769

56 Romani, S. & Tsodyks, M. Short-term plasticity based network model of place cells dynamics. *Hippocampus* **25**, 94-105 (2015). https://doi.org:10.1002/hipo.22355

57 York, L. C. & van Rossum, M. C. Recurrent networks with short term synaptic depression. *J Comput Neurosci* **27**, 607-620 (2009). https://doi.org:10.1007/s10827-009-0172-4

58 Mercier, M. R. *et al.* Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: A glimpse on white matter signal. *Neuroimage* **147**, 219-232 (2017). https://doi.org:10.1016/j.neuroimage.2016.08.037

59 McCarthy, G., Nobre, A. C., Bentin, S. & Spencer, D. D. Language-related field potentials in the anterior-medial temporal lobe: I. Intracranial distribution and neural generators. *J Neurosci* **15**, 1080-1089 (1995).

60 Bansal, A. *et al.* Neural Dynamics Underlying Target Detection in the Human Brain. *Journal of Neuroscience* **34**, 3042-3055 (2014). https://doi.org: 10.1523/JNEUROSCI.3781-13.2014

61 Mitzdorf, U. Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. *Physiological Reviews* **65**, 37-99 (1985).

62 Logothetis, N. K. The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philos Trans R Soc Lond B Biol Sci* **357**, 1003-1037 (2002).

63 O'Brien, R. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity: International Journal of Methodology* **41**, 673-690 (2007). https://doi.org:10.1007/s11135-006-9018-6

64 Agam, Y. *et al.* Robust selectivity to two-object images in human visual cortex. *Current Biology* **20**, 872-879 (2010).

65 Habib, R., McIntosh, A. R., Wheeler, M. A. & Tulving, E. Memory encoding and hippocampally-based novelty/familiarity discrimination networks. *Neuropsychologia* **41**, 271-279 (2003). https://doi.org:10.1016/s0028-3932(02)00160-4

66 Jiang, Y., Haxby, J. V., Martin, A., Ungerleider, L. G. & Parasuraman, R. Complementary neural mechanisms for tracking items in human working memory. *Science* **287**, 643-646 (2000). https://doi.org:10.1126/science.287.5453.643

67 Chow, T. E. & Rissman, J. Neurocognitive mechanisms of real-world autobiographical memory retrieval: insights from studies using wearable camera technology. *Ann N Y Acad Sci* **1396**, 202-221 (2017). https://doi.org:10.1111/nyas.13353

68 Duff, M. C., Wszalek, T., Tranel, D. & Cohen, N. J. Successful life outcome and management of real-world memory demands despite profound anterograde amnesia. *J Clin Exp Neuropsychol* **30**, 931-945 (2008). https://doi.org:10.1080/13803390801894681

69 Nielson, D. M., Smith, T. A., Sreekumar, V., Dennis, S. & Sederberg, P. B. Human hippocampus represents space and time during retrieval of real-world memories. *Proc Natl Acad Sci U S A* **112**, 11078-11083 (2015). https://doi.org:10.1073/pnas.1507104112

70 Misra, P., Marconi, A., Peterson, M. & Kreiman, G. Minimal memory for details in real life events. *Sci Rep* **8**, 16701 (2018). https://doi.org:10.1038/s41598-018-33792-2

71 Tang, H. *et al.* Predicting episodic memory formation for movie events. *Scientific Reports* **6**, 30175 (2016). https://doi.org:10.1038/srep30175

72 Tulving, E., Markowitsch, H. J., Craik, F. E., Habib, R. & Houle, S. Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cereb Cortex* **6**, 71-79 (1996). https://doi.org:10.1093/cercor/6.1.71

73 Lundqvist, M. *et al.* Gamma and Beta Bursts Underlie Working Memory. *Neuron* **90**, 152-164 (2016). https://doi.org:10.1016/j.neuron.2016.02.028

74 Staresina, B. P. *et al.* Recollection in the human hippocampal-entorhinal cell circuitry. *Nat Commun* **10**, 1503 (2019). https://doi.org:10.1038/s41467-019-09558-3

75 Bergmann, H. C., Rijpkema, M., Fernandez, G. & Kessels, R. P. Distinct neural correlates of associative working memory and long-term memory encoding in the medial temporal lobe. *Neuroimage* **63**, 989-997 (2012). https://doi.org:10.1016/j.neuroimage.2012.03.047

76 Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nat Rev Neurosci* **23**, 744-766 (2022). https://doi.org:10.1038/s41583-022-00642-0

77 Spalla, D., Cornacchia, I. M. & Treves, A. Continuous attractors for dynamic memories. *Elife* **10** (2021). https://doi.org:10.7554/eLife.69499

78 Lin, T. *et al.* in *arxiv:1405.0312v3* (2015).

79 Brainard, D. The Psychophysics Toolbox. *Spatial Vision* **10**, 433-436 (1997).

80 Pelli, D. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* **10**, 437-442 (1997).

81 Fried, I., Cerf, M., Rutishauser, U. & Kreiman, G. *Single neuron studies of the human brain. Probing cognition.*, 408 (MIT Press, 2014).

82 Groppe, D. M. *et al.* iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *J Neurosci Methods* **281**, 40-48 (2017). https://doi.org:10.1016/j.jneumeth.2017.01.022

83 Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179-194 (1999).

84 Joshi, A. *et al.* Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics* **9**, 69-84 (2011). https://doi.org:10.1007/s12021-010-9092-8

85 Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968-980 (2006). https://doi.org:10.1016/j.neuroimage.2006.01.021

86 Wu, J. *et al.* Accurate nonlinear mapping between MNI volumetric and FreeSurfer surface coordinate systems. *Hum Brain Mapp* **39**, 3793-3808 (2018). https://doi.org:10.1002/hbm.24213

87 Wang, J., Tao, A., Anderson, W. S., Madsen, J. R. & Kreiman, G. Mesoscopic physiological interactions in the human brain reveal small-world properties. *Cell Rep* **36**, 109585 (2021). https://doi.org:10.1016/j.celrep.2021.109585

88 Bansal, A. *et al.* Temporal stability of visually selective responses in intracranial field potentials recorded from human occipital and temporal lobes. *Journal of Neurophysiology* **108**, 3073-3086 (2012). https://doi.org:10.1152/jn.00458.2012

89 Mitra, P. & Bokil, H. *Observed brain dynamics* (Oxford University Press, 2008).

90 Xiao, Y. *et al.* Cross-task specificity and within-task invariance of cognitive control processes. *Cell Rep* **42**, 111919 (2023). https://doi.org:10.1016/j.celrep.2022.111919

91 Nelder, J. & Wedderburn, R. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370-384 (1972). https://doi.org:doi:10.2307/2344614
92 Hastie, T. & Tibshirani, R. *Generalized additive models*. (Chapman and Hall/CRC, 1990).
93 Dormann, C. F. *et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27-46 (2013). https://doi.org:https://doi.org/10.1111/j.1600-0587.2012.07348.x
94 Welsch, R. E. & Kuh, E. Linear regression diagnostics. *National Bureau of Economic Research* (1977).
95 Gujarati, D. Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Linear Regressions: A Generalization. *The American Statistician* **24**, 18-22 (1970).
96 Ba, J., Hinton, G., Mnih, V., Leibo, J. & Ionescu, C. Using Fast Weights to Attend to the Recent Past. *arXiv* **1610.06258** (2016).

Figures and Tables

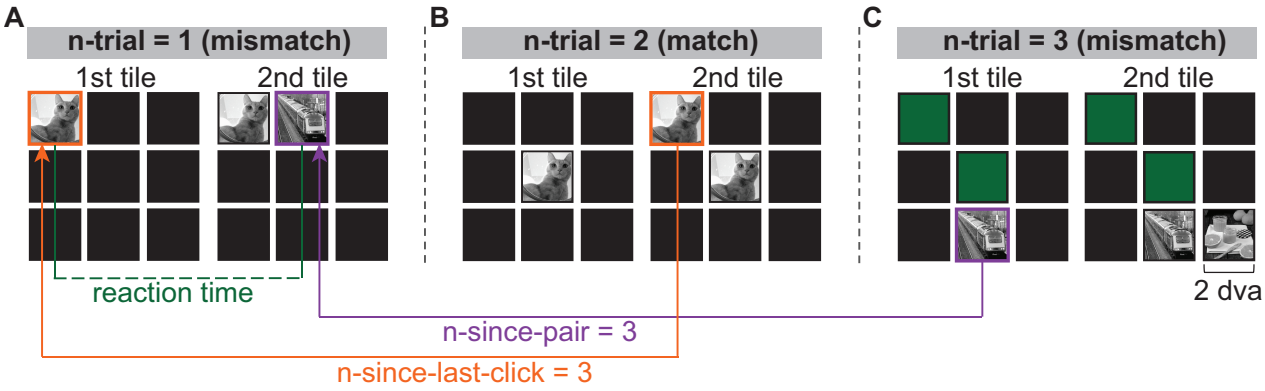

Figure 1. Experimental paradigm

A-**C.** Three consecutive trials in a 3×3 board. In each trial, two tiles were flipped sequentially in a self-paced manner (1st tile, then 2nd tile). If the two tiles contained different images (**A**, **C**, **mismatch**), both tiles reset to their original active (black) state after 1 second. If both tiles contained the same image (**B**, **match**), they turned green after 1 second and stayed green for the remainder of the block. Three behavioral predictors used in the generalized linear models (GLM) are defined here: reaction time (the time between the 1st and 2nd tile within a trial), n-since-last-click (the number of clicks elapsed since the same tile was clicked last), and n-since-pair (the number of clicks elapsed since the last time a given tile's matching pair was clicked). Each tile spanned approximately 2 degrees of visual angle (dva) in size. See also **Movie S1**.

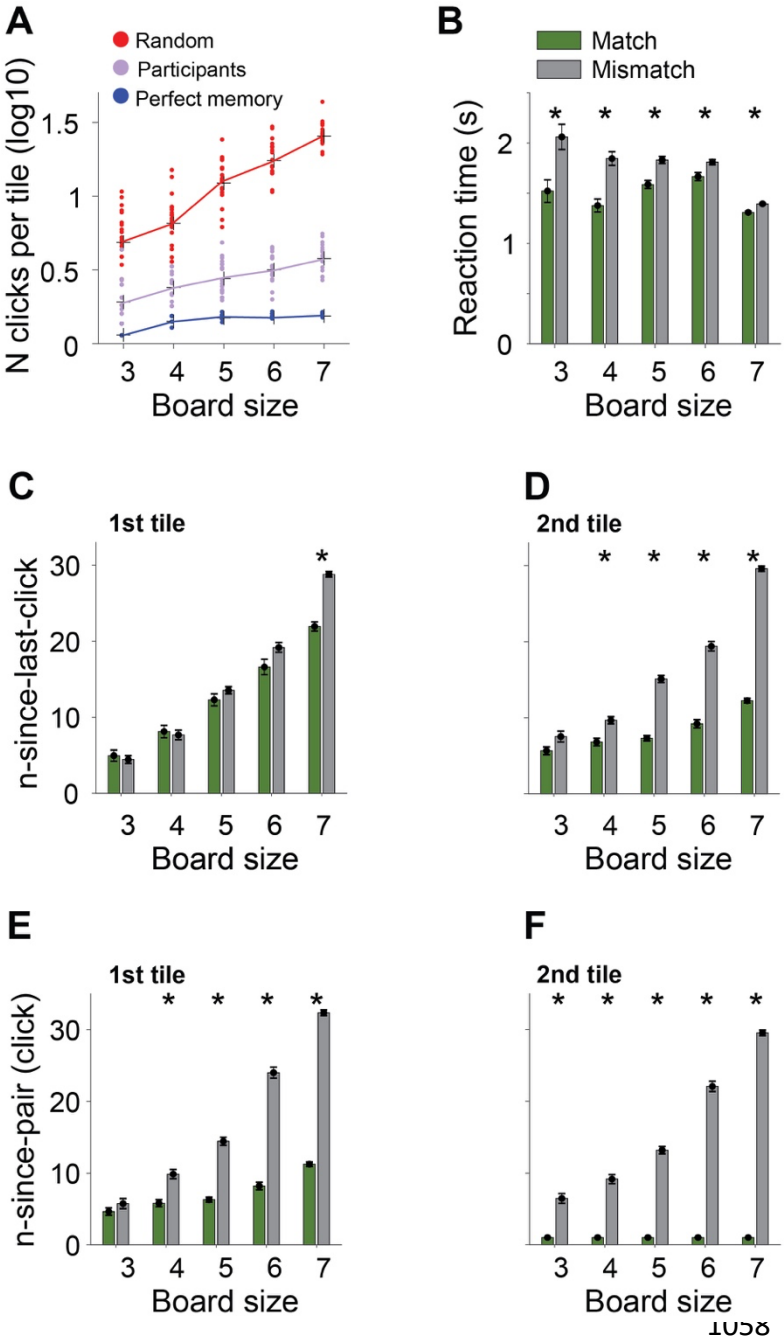

Figure 2. Behavioral measures show that participants thrived in the task and that finding matching pairs displayed classical memory effects

A. Number of clicks per tile (log scale) as a function of board size for random simulation model (red, n=20), perfect memory simulation model (blue, n=20), and epilepsy patient participants (purple, n=20) (**Methods**). Perfect memory simulation models may generate different number of clicks per tile because the click location for new tiles was randomized. The performance of epilepsy patients was better than the random model and worse than the perfect model. The number of clicks per tile increased as board size incremented. **B.** Reaction times for match (green) and mismatch (gray) trials for different board sizes. Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01). Reaction time of mismatch trials was longer than match trials. **C-F.** Average n-since-last-click (**C**, **D**) and n-since-pair (**E**, **F**) for the 1st tile (**C**, **E**) and the 2nd tile (**D**, **F**) for each board size. Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01). For n-since-last-click, trials in which a tile was clicked for the 1st time were excluded in this figure. For n-since-pair, trials in which any tile's matching pair had not been seen before were excluded in this figure. All error bars indicate s.e.m. (n=20 participants).

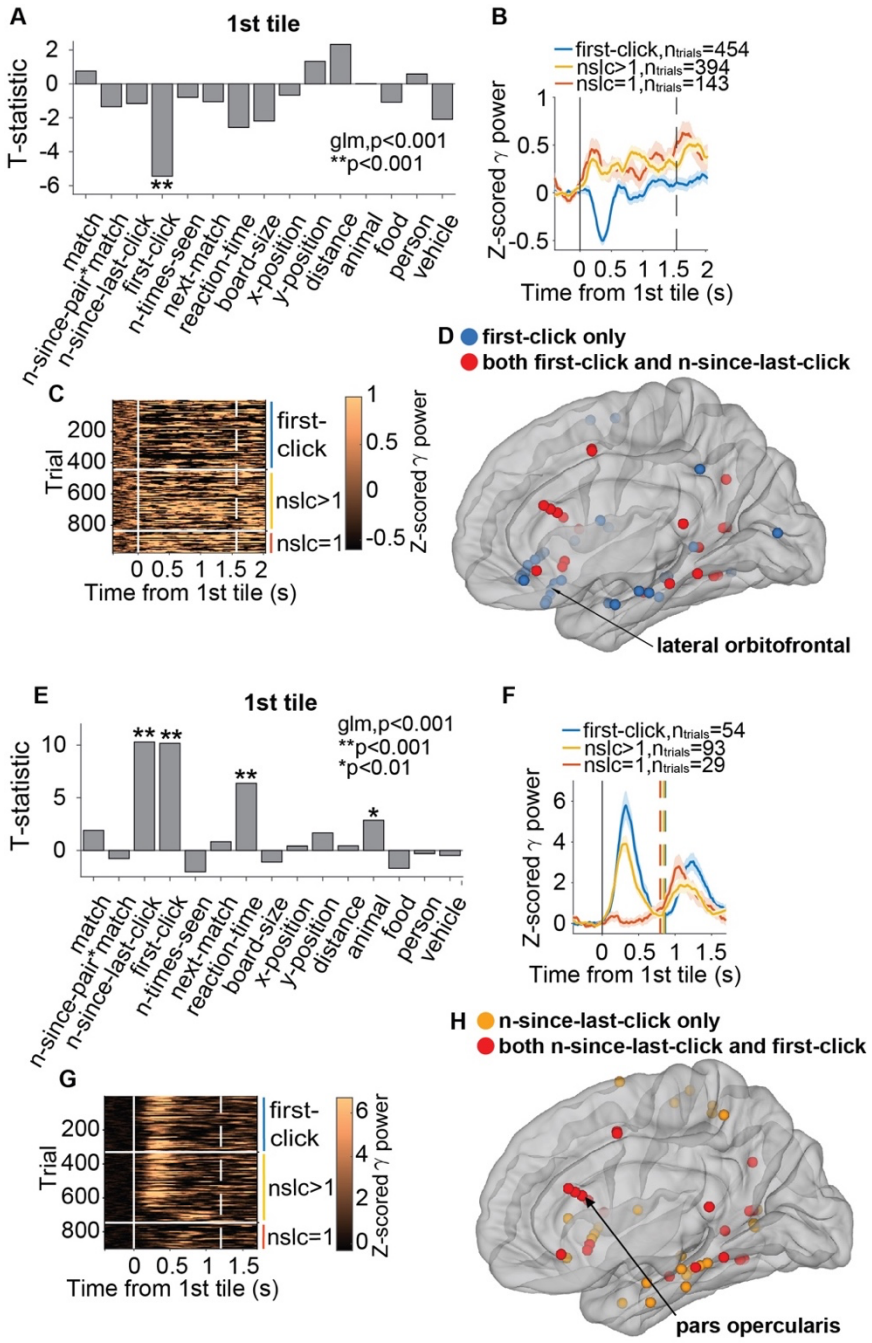

Figure 3. Neural signals reflect novelty and familiarity

Panels show two example electrodes, one in the right lateral orbitofrontal cortex (**A**-**D**), one in the left pars opercularis (**D**-**H**), and population locations in **D**, **H**. **A**, **E**. T-statistic of each predictor in the GLM analyses (**Methods**). Asterisks indicate statistically significant predictors for the neural signals.

B, **F**. Z-scored gamma band power aligned to the 1st tile onset (solid vertical line) for novel tiles (blue), unfamiliar tiles (n-since-last-click>1, yellow), and familiar tiles (n-since-last-click=1, red). The vertical dashed line indicates the mean reaction time. Multiple dashed lines in **F** indicate reaction time equalization (**Methods**). The time axis extends from 400 ms before the click to 500 ms after the average reaction time. **F** displays only trials after RT equalization (**Methods**). Shaded error bars indicate s.e.m.

C, G. Raster plots showing the z-scored gamma power in individual trials ordered by first-click and then larger to smaller n-since-last-click; division indicated by white horizontal lines/spaces and colored vertical bars.

D. Locations of all electrodes where first-click was a significant predictor during the 1st tile. Blue: first-click only; red: both first-click and n-since-last-click were significant predictors.

H. Locations of all electrodes where n-since-last-click was a significant predictor during the 1st tile. Orange: n-since-last-click only; red: both n-since-last-click and first-click were significant predictors. All electrodes were reflected on one hemisphere for display purposes.

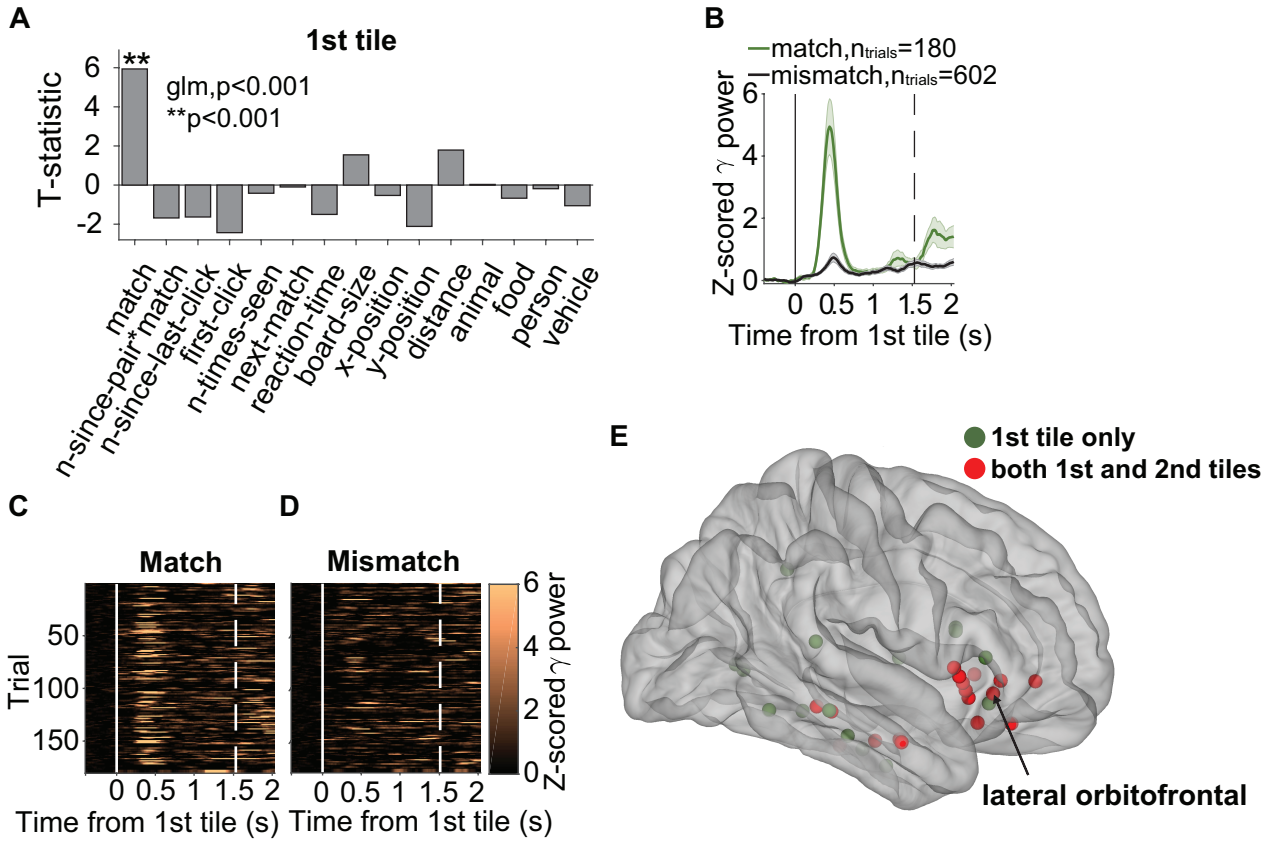

Figure 4. Neural signals predict correct retrieval

Panels show an example electrode in the right lateral orbitofrontal cortex (see arrow in part **E**) and population locations in **E**. The format follows **Figure 3**. **A.** T-statistic of each predictor in the GLM analysis. Asterisks indicate statistically significant predictors for the neural signals. **B**. Z-scored gamma band power aligned to the 1st tile onset (solid vertical line) for match trials (green) and mismatch trials (black). The vertical dashed line indicates the mean reaction time. Shaded error bars indicate s.e.m. **C**-**D**. Raster plots showing the gamma power in individual trials for match (left) and mismatch (right) trials. **E**. Locations of all electrodes where match was a significant predictor during the 1st tile only (green) and during both tiles (red). All electrodes were reflected on one hemisphere for display purposes.

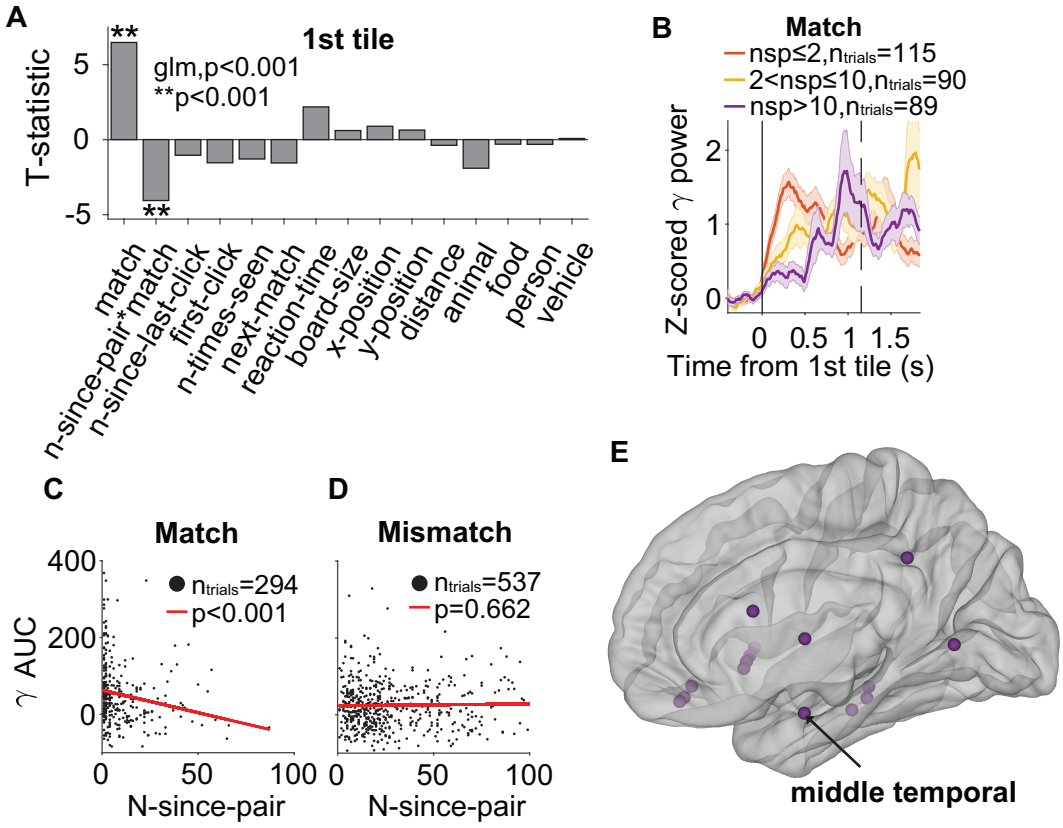

Figure 5. Neural signals reflect the strength of memory retrieval

Panels show an example electrode in the left middle temporal gyrus (see arrow in part **E**) and population locations in **E**. **A**. T-statistic of each predictor in the GLM analyses. Asterisks indicate significant predictors for the neural signals. **B**. Z-scored gamma band power aligned to the 1st file onset (solid vertical line) for match trials with small n-since-pair (nsp) (red, stronger memories), intermediate nsp (yellow), and large nsp (purple, weaker memories). The vertical dashed line indicates the mean reaction time. Shaded error bars indicate s.e.m. **C-D**. Scatter plots of the area under the curve (AUC) of the gamma band power as a function of nsp for match trials (**C**) and mismatch trials (**D**). Each dot represents one trial. Red lines show linear fits to the data. **E**. Locations of all electrodes where nsp was a significant predictor. All electrodes were reflected on one hemisphere for display purpose.

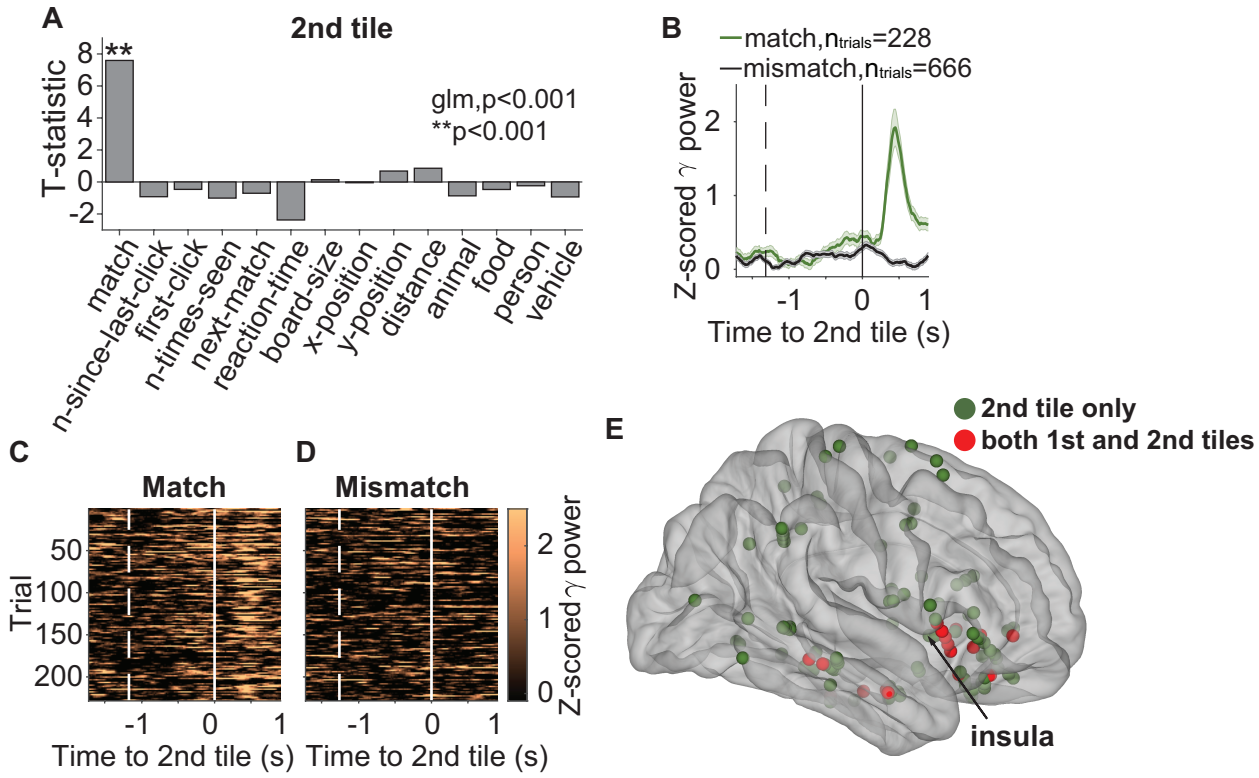

Figure 6. Neural signals after the second tile reflect correct retrieval

Panels show an example electrode located in the insula (see arrow in part **E**) and population locations. **A**. T-statistic of each predictor in the GLM analyses for the responses after the 2nd tile. Asterisks indicate significant predictors of neural signals. **B**. Z-scored gamma band power aligned to the onset of the 2nd tile (solid vertical line) for match (green) and mismatch (black) trials. The dashed line indicates the mean onset of the 1st tile. **C**-**D**. Raster plots showing the gamma band power in individual trials for match (left) and mismatch (right) trials. **E**. Locations of all electrodes where match was a significant predictor of neural responses after the 2nd tile only (green) or during both tiles (red). All electrodes were reflected on one hemisphere for display purpose.

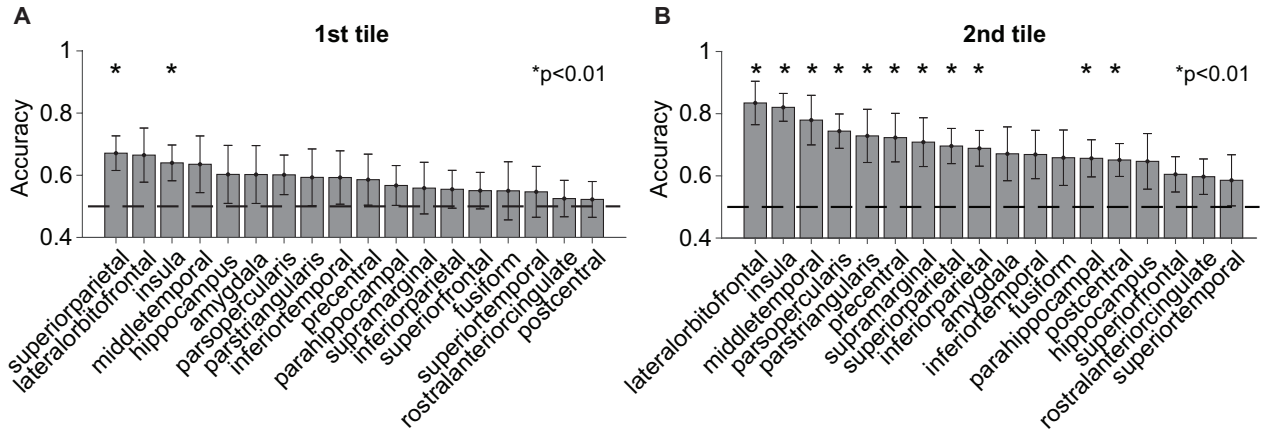

Figure 7. Machine learning decoding of match versus mismatch for all channels within each brain region

Average decoding accuracy for each brain region (**Methods**) using neural responses after the 1st tile (**A**) or after the 2nd tile (**B)**. Brain regions are ordered from higher to lower average decoding accuracy. The dashed horizontal line indicates chance accuracy. Asterisks denote significant decoding accuracy above chance (α=0.01). All error bars indicate SD (n=1,000 iterations).

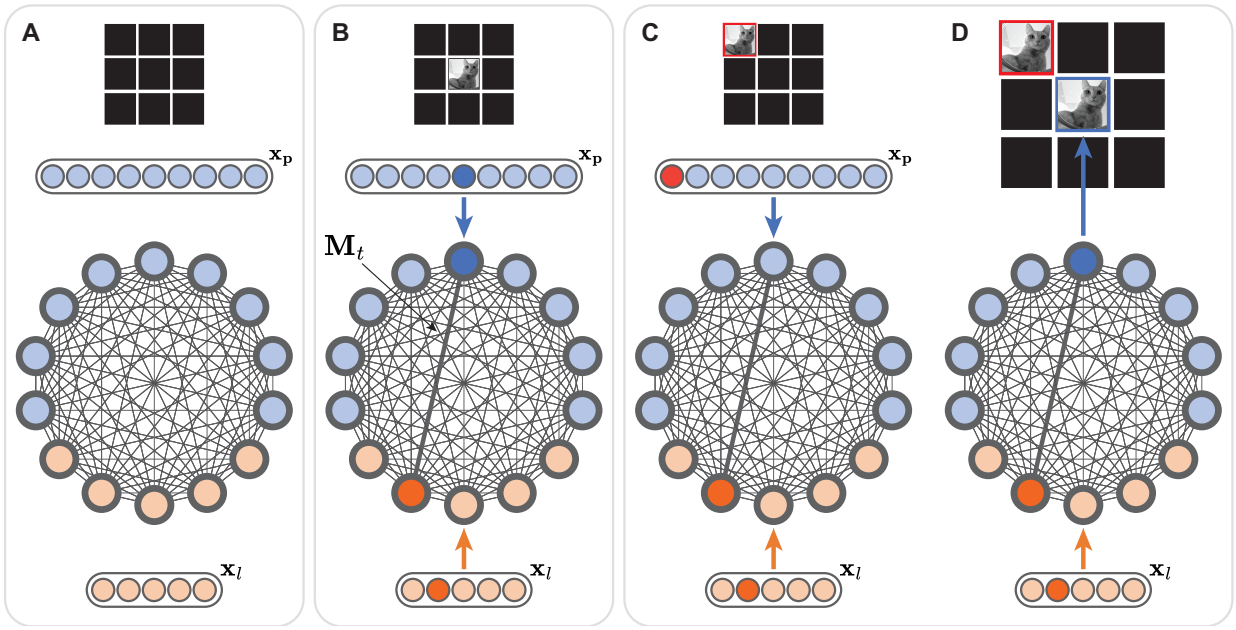

Figure 8: Hebbian attractor model architecture and operating regimes

A. Schematic representation of the model architecture used for the 3x3 grid. The 9 blue units encode position ($\boldsymbol{x}_p$), while the 5 orange units represent the image label ($\boldsymbol{x}_l$). The black lines between units illustrate the Hebbian weights $\boldsymbol{M}_t$ in the attractor network. **B.** Learning regime. In this example, the model represents a cat (label=2) at position=5. **C, D.** Inference regime. In this example, the model is tasked with matching the cat (label=2) observed at position=1. Only the label information is provided to the model in the inference regime. The model's updates (**Methods**) lead to the unit representing position=5 to exhibit the highest activity (**D**), thereby determining the corresponding tile to be clicked. The darker color indicates stronger activation of the corresponding units. The red color indicates the tile to match (which is unavailable for clicking, **Methods**) and its corresponding positional unit.

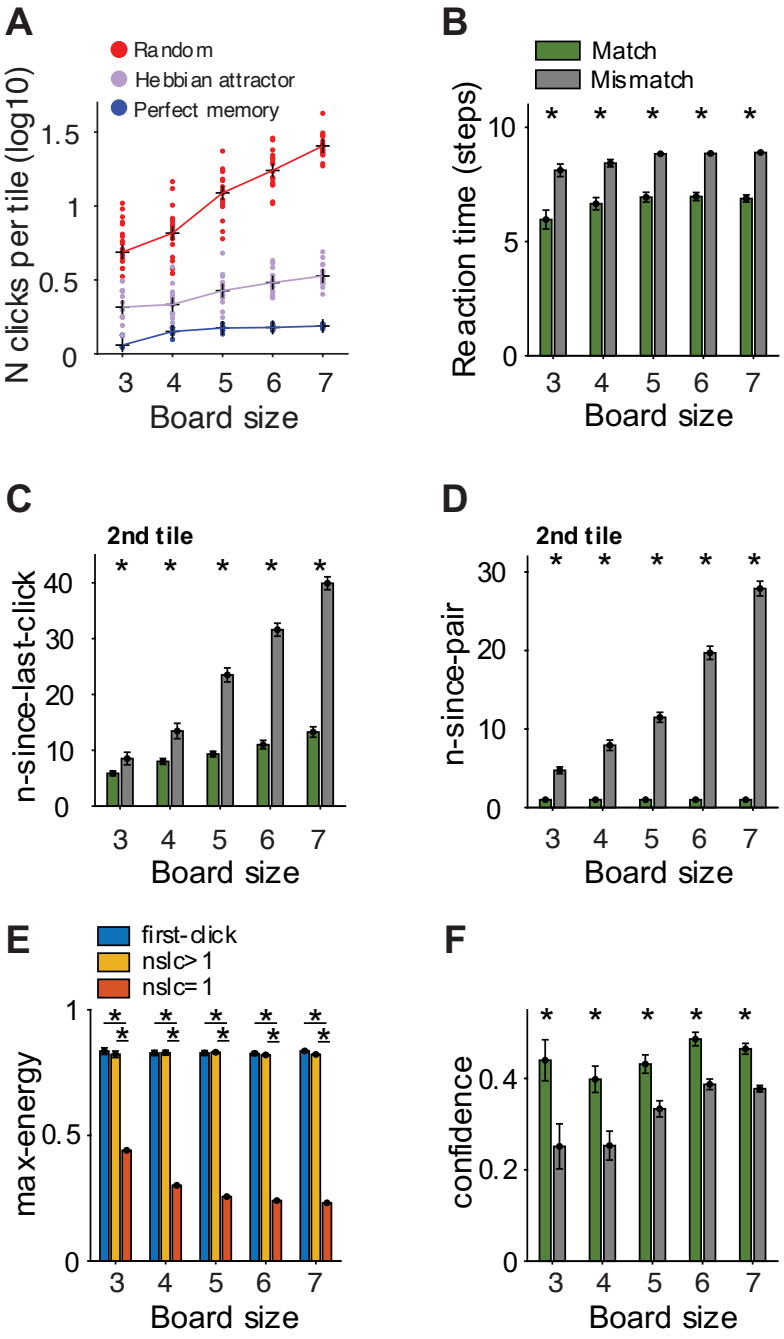

Figure 9: Hebbian attractor model captures behavior and neural measures

A. Number of clicks per tile (log scale) as a function of board size for random simulation model (red, n=20), perfect memory simulation model (blue, n=20), and Hebbian attractor (purple, n=20) (**Methods,** compare to **Figure 2A**). **B**. Reaction times of the model (Methods) for match (green) and mismatch (gray) trials for different board sizes (compare to **Figure 2B**). **C**, **D**. Average n-since-last-click (**C**) and n-since-pair (**D**) for the 2nd click for each board size (compare to **Figure 2D**, **F**. **E**). Max-energy for novel tiles (blue), unfamiliar tiles (n-since-last-click>1, yellow) and familiar tiles (n-since-last-click=1, red, compare to **Figure 3**). **F**. Model confidence for match (green) and mismatch (gray) trials for different board sizes (**Methods**). The model confidence in match trials was larger than in mismatch trials (compare to **Figure 5**). All asterisks denote significant

difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01). All error bars indicate s.e.m. (n=20).

Tables

Predictor	Description	Which tile
match	Whether the trial was a match or mismatch	both
n-since-pair*match	how many clicks ago the tile's pair was clicked (matched trials only)	1st
n-since-last-click	how many clicks ago the same tile was clicked	both
first-click	Whether a tile was clicked the very first time	both
n-times-seen	number of times the same image had been previously clicked	both
next-match	whether the next trial was a match or mismatch	both
reaction-time	time between the 1st and 2nd tile	both
board-size	Total number of tiles in the current block	both
x-position	x position in pixel	both
y-position	y position in pixel	both
distance	distance between the 2nd tile of the current trial and the 1st tile of the next trial in pixel	both
animal	image belonged to animal category	both
food	image belonged to food category	both
person	image belonged to person category	both
vehicle	image belonged to vehicle category	both

Table 1. Predictors in the generalized linear models, their definitions, and applicable tiles.

Supplementary Materials

The Supplementary Material (separate file) includes:

1 supplementary movie

9 supplementary tables

13 supplementary figures