Resolving a paradox about how vision is transformed into familiarity

Simon Bohn, Catrina M. Hacker, Barnes G. L. Jannuzi, Travis Meyer, Madison L. Hay, Nicole C. Rust
Department of Psychology, University of Pennsylvania, Philadelphia, USA

Summary: While humans and other primates are generally good at remembering the images they have seen, they systematically remember some images better than others ("image memorability"). Here, we leverage image memorability to resolve a puzzle around how the brain transforms seeing into familiarity. Namely, the neural signal driving familiarity reports is thought to be repetition suppression, an overall reduction in the vigor of the population response in brain regions, including inferotemporal cortex (ITC). However, within ITC, more memorable images evoke higher firing rates than less memorable ones, even when they are repeated. These two observations conflict: if *reduced* firing leads to stronger memory signaling, then why are the images that induce *greater* firing more memorable? To investigate this "familiarity-memorability paradox", we compared neural activity in ITC and the hippocampus (HC) as two macaque monkeys performed a single-exposure image familiarity task. We found a resolution to the paradox, reflected in two pieces of evidence. First, memory and memorability were not completely overlapping in ITC, suggesting that familiarity signals could be extracted independently of memorability modulation. Second, responses in HC reflected a transformation of ITC responses that extracted memory, leaving behind stimulus-evoked firing modulation. This transformation could be accounted for by a linear decoder that selectively extracts memory and preserves larger ITC memory signals for more memorable images. These results resolve the familiarity-memorability paradox by demonstrating that ITC familiarity signals are a component (but not the entirety) of repetition suppression, and they establish the existence of a previously undescribed medial temporal lobe computation.

Significance: Neural representations of image memorability create a paradox: familiarity reports are thought to derive from *reduced* firing in inferotemporal cortex (ITC), but more memorable images evoke more vigorous responses there. To resolve this contradiction, we compared ITC with an area downstream, the hippocampus (HC), as macaque monkeys performed a single-exposure familiarity task. We found evidence for a computation between ITC and HC that selectively extracts ITC memory signals. This transformation could be explained by a linear decoder that preserves larger memory signals for more memorable images. This work resolves the familiarity-memorability paradox by showing that ITC memory signals have been misunderstood, and it reveals a previously undescribed role for the medial temporal lobe in familiarity behavior.

Introduction:

Understanding how the brain stores information about experiences and remembers that information later is central to understanding how memory works. In the case of visual familiarity ("Have you seen this before?", also called recognition memory, Fig. 1a), one well-established behavioral signature is that some images are systematically more memorable or forgettable than others, a property called "image memorability" (Fig. 1b). Image memorability is consistent across human observers (1–6) and between humans and rhesus monkeys (7). Likewise, image memorability is reflected in deep artificial neural networks that are trained to categorize objects with no explicit notion of memorability (7, 8), suggesting that it is a natural byproduct of a system wired up to see. Here, we use image memorability variation as a lever to understand how the brain transforms vision into familiarity.

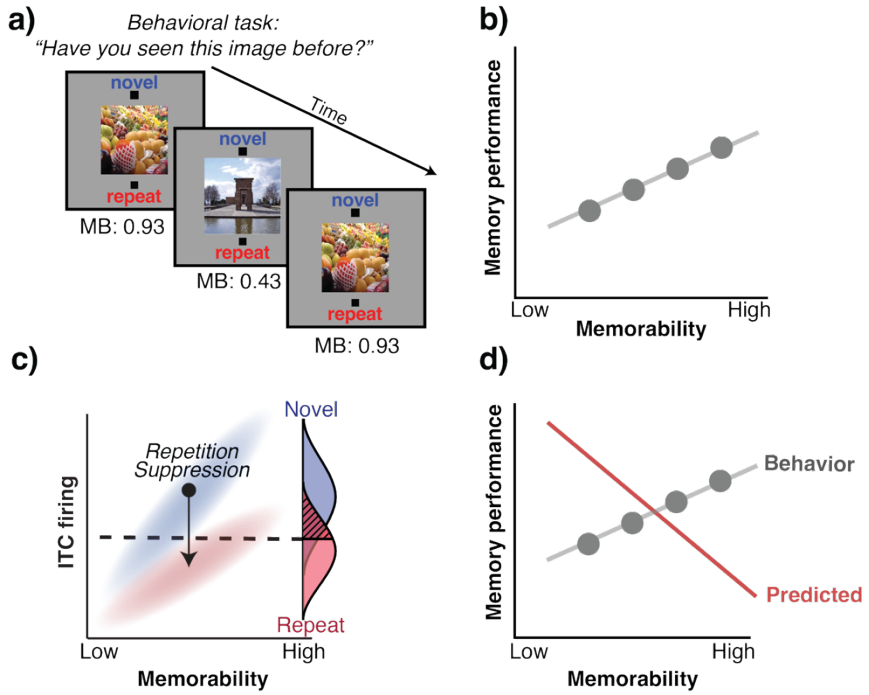

Figure 1: *The familiarity-memorability paradox.* Plots are schematics, idealized here for clarity. **a)** Summary of the behavioral task. Two rhesus monkeys were shown one image per trial and indicated whether they judged the image to be novel (never seen before) or repeated (seen exactly once before). Images naturally varied in memorability (MB), a scalar (range 0-1) that corresponds to the likelihood of a human correctly remembering that image. **b)** Sketch of familiarity behavior (solid gray dots) on this task as a function of image memorability, showing that the memory performance of one subject increases with image memorability. In question is the underlying neural signals that drive this pattern of behavior. **c)** Grand mean ITC population firing; blue and red clouds reflect idealized distributions of population firing rates to novel and repeated images, respectively. Repetition suppression (RS) is the difference between a novel and repeated firing rate, with an example image indicated with the solid black arrow, where the filled black circle represents ITC firing to an exemplar image's novel presentation, the arrowhead reflects the repeat presentation, and the length of the line between them is RS. The implementation of a simple firing rate classifier is shown as a decision boundary applied to the overall mean firing rate (black dashed line). This classifier would classify any image (regardless of the ground truth) below the boundary as repeated, and anything above it as novel. Also shown: 1-D projections of the firing rate clouds onto the classifier's axis. **d)** Predicted behavioral performance by the classifier shown in c. The familiarity-memorability paradox occurs because more memorable images produce higher ITC firing rates. Consequently, the most memorable familiar images (panel c, right side of red 2D distribution) are predicted to be the least memorable (because they have the highest firing rates and bleed into the novel side of the classifier, represented by the crosshatched region of the 1-D projections), at odds with behavior.

Several lines of evidence suggest that the signal the brain relies on for visual familiarity is related to repetition suppression: the average reduction in the overall vigor of a population response when a stimulus is repeated compared to when it is novel (9–15). The most straightforward version of this scheme proposes that the brain linearly decodes the population response to a given image to guide behavior, where images that evoke higher firing rates are classified as novel and images that evoke lower firing rates are classified as repeated. In ITC, this account has been shown to align with behavioral signatures of familiarity, such as the rate of forgetting as time passes after an image is first seen (14), and confusion between similar images (15).

Incorporating image memorability into this scheme leads to a paradox around the proposal that repetition suppression drives familiarity. The paradox follows from the fact that more memorable images produce more vigorous responses. This observation calls claims about the nature of the signal driving familiarity into question: if *reduced* population firing signals stronger memory, then how can more memorable images induce *greater* population firing (Fig. 1c-d)? We call this the "familiarity-memorability" paradox. While it may be tempting to propose that the repetition suppression signal by itself (the firing rate difference between novel and repeated images) drives familiarity, such explanations are circular, as they require memory to explain memory ("what was the firing rate of that image last time it was seen?") As such, it's unclear how the brain solves this paradox in the transformation from visual representations into familiarity signals.

This paradox is not limited to memorability. Other visual properties of images, like contrast, also modulate ITC population magnitude in ways that cannot be reconciled with the claim that familiarity is signaled exclusively by ITC population response vigor. One study attempted to reconcile a similar contradiction for image contrast by suggesting the existence of downstream computations that selectively extract memory information from ITC while discarding fluctuations attributed to image contrast (16). While that study proposed how ITC may be read out to resolve a similar question, it did not measure downstream of ITC to test that proposal. In addition, it's unclear whether and how the mechanism proposed for contrast would work for memorability, given that familiarity behavior is unaffected by contrast but impacted by memorability (and both impact stimulus-evoked firing rates in ITC). It is thus unclear whether information about familiarity is reflected in ITC in a manner that can be reconciled with image memorability behavior.

We wondered if the familiarity-memorability paradox might be resolved downstream. ITC richly innervates the perirhinal and parahippocampal cortices (17), linking ITC to the hippocampus via entorhinal cortex (18, 19). Repetition suppression to repeated images has been observed throughout the medial temporal lobe (MTL), including perirhinal cortex (9, 12, 20), entorhinal cortex (9, 21) , and HC (22–24). In this study, we recorded neural activity in HC with the goal of capturing any contributions made by MTL structures between ITC and HC. Here, we show that the ITC signal that best aligns with memorability behavior has been misunderstood — it is not the entirety of RS, but a component that does not overlap with memorability. We also describe

an MTL computation that can selectively extract ITC memory information such that larger ITC memory signals for more memorable images are reflected in HC. The computational transformation cannot be accounted for by simple mechanisms (like thresholding ITC responses) but can be described by a selective linear read-out that preserves larger memory signals for more memorable images. We thus 1) resolve the familiarity-memorability paradox by demonstrating that the ITC familiarity signal has been misunderstood, 2) demonstrate for the first time that the paradox is not reflected in HC, and 3) show evidence for a previously undescribed MTL computation.

Results

Single-exposure visual memory task

Two rhesus macaque monkeys performed a single-exposure visual recognition memory task (14) (Fig. 1a). The monkeys sequentially viewed natural images varying in memorability and reported after each trial whether the image was novel (never seen before in this or any previous sessions) or repeated (seen exactly once before in this session) for a juice reward. Monkeys initiated a trial by fixating on a central fixation point. After fixating for 500ms, the monkeys maintained fixation while an image appeared for 500ms, subtending 4 degrees of the central visual field. After the image disappeared, the monkeys indicated whether the image was novel or repeated by making a saccade to one of two response targets. Correct responses were rewarded with juice. All images were shown exactly twice, and the time between the novel and the repeated presentation of an image varied by the number of intervening trials (n-back) from seconds (“1-back”) to minutes (“64-back”) (see Methods). After training, the monkeys performed this task proficiently (on repeated trials across all n-backs: Monkey 1 at 91% correct and Monkey 2 at 84% correct – Fig. S1). Images depicted a diverse range of object categories and scenes and spanned a broad range of memorability (Fig. S2). Image memorability scores (a scalar ranging from 0 to 1 that reflects how likely an image is to be remembered) were obtained from a model trained to predict memorability, MemNet (5). This model was trained on human behavior and possessed accuracy near the ceiling imposed by inter-subject consistency. As previously reported (7), human-based memorability scores predicted memory performance for monkeys, a finding replicated in these data (Fig. 2a and Fig. S1 for individual monkeys).

Memory was reflected as repetition suppression in ITC and HC

As the monkeys performed this task, we recorded neural activity in ITC or HC with a 24-channel recording probe lowered acutely for each session. The dataset included 7 sessions (290 units) and 16 sessions (312 units) for monkey 1 and 8 sessions (333 units) and 21 sessions (503 units) for monkey 2, recorded in ITC and HC, respectively. We included units with activity that reflected task-responsiveness (see Methods), and images for which the monkey completed both the novel and repeated presentation trials. Because the responses of individual ITC and HC units are often tuned for specific visual features, estimating the overall population response requires many hundreds of units. Therefore, we concatenated units across sessions into a pseudopopulation and aligned images of similar memorability into “pseudoimages” as in refs (7,

14, 16, 25). All neural analyses were performed on these pseudoimages, to which we refer interchangeably with “images” when describing analyses of these data. When creating the pseudopopulations, we aligned trials in a manner that preserved novelty, n-back, and approximate memorability. Data were pooled between the two monkeys for each brain area and matched between brain areas for overall memory information (see Methods for details).

We began by comparing how visual memory information was represented in each brain area, focusing on whether it was reflected as repetition suppression, repetition enhancement, or some mixture across different units. To do so, we computed sensitivity to memory (d’) for each unit and compared the distributions of these units in ITC and HC. Histograms of d' in both brain areas were roughly normal, but shifted towards positive values (repetition suppression), indicating that both populations were dominated by repetition suppression (Fig. 2b). To determine whether repetition enhanced units were carrying reliable memory information in ITC or HC (as opposed to apparent repetition enhancement driven by noise), we applied a ranked population decoding analysis (Fig. S3) and found that the small fraction of repetition enhanced units did not meaningfully contribute to memory representations in either brain region.

We next wondered whether memory signals arose earlier in one region or the other, and so we examined how memory information changed as a function of time from stimulus onset in each brain area. To do so, we measured the performance of a cross-validated linear decoder trained to classify each image presentation as “novel” or “repeated” based on the firing rates of each unit (an FLD decoder; see Methods for details) as a function of time. We found that memory information increased with approximately the same dynamics in ITC and HC, reaching peak performance at the end of the 500ms viewing window (Fig. 2c).

Comparing memorability and memory in ITC and HC

Having established that repetition suppression reflected memory in both ITC and HC, we next turned to investigate how memorability modulates stimulus-evoked firing in both brain areas. To do so, we examined the relationship between image memorability and the overall vigor of the ITC and HC population response by averaging firing rates across all units to a given image (the “grand mean firing rate”) and plotting the results in each brain area as a function of memorability (Fig. 3a-b). Examining the data in this way revealed insights related to the nature of memorability, memory, and the evidence for a computational transformation between ITC and HC, and each is explained below, in turn. To determine the appropriate spike count window for these analyses, we selected windows that best capture the transformation that happens between ITC and HC (see Figure 4 for the details of these dynamics). This resulted in a spike count window of 100-500ms in ITC and 300-500ms in HC. The results described in Fig. 3 were not qualitatively different when matched spike count windows of 300-500ms were used for both areas (Fig. S4).

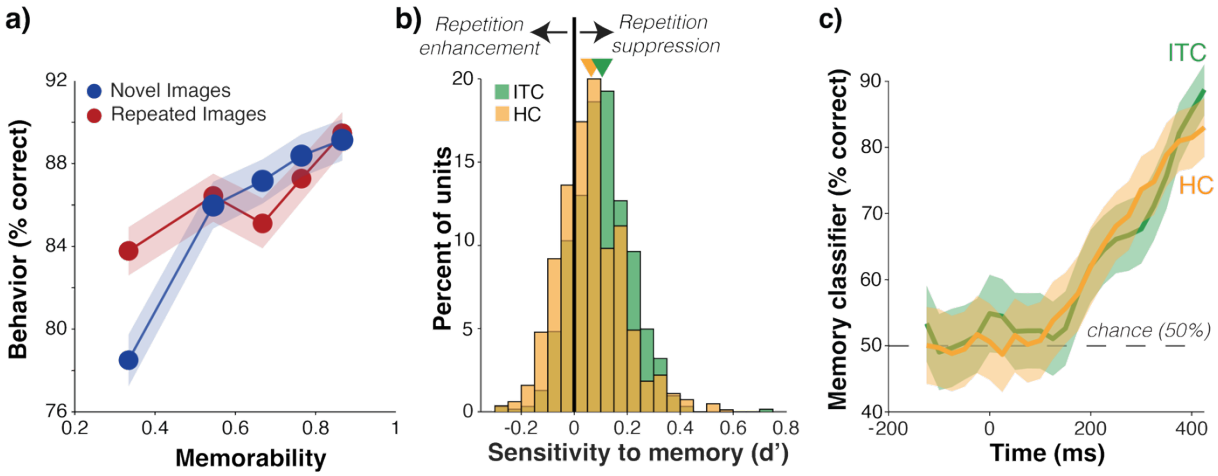

Figure 2. *Monkey behavior and memory signaling in ITC and HC.* **a)** The pooled results of the two macaque monkeys on this task for novel (blue) and repeated (red) trials (n=18,465 trials of each type). As predicted from the human-derived memorability scores, higher memorability images were more likely to be correctly recognized as repeated. Error shadow represents a 95% confidence interval computed by bootstrapping across trials (10,000 iterations). **b)** The distribution of sensitivity to memory (d') across units in the ITC and HC populations (602 units in ITC, 836 in HC) in the 300-500ms spike count window. On the x-axis, a positive value for repetition suppression corresponds to units that fire less to a repeated presentation than to a novel one. The solid black vertical line separates units that were enhanced by repetition (left) from those that were suppressed (right). Triangles point to the mean of each distribution. **c)** Time course of memory signal in ITC and HC measured by the performance of a weighted linear decoder as a function of time (150ms spike count window, 20ms sliding intervals). Shaded regions represent one standard deviation across cross validations.

Memorability affected the overall firing rate in ITC, but not HC

Memorability was strongly correlated with firing rate in ITC for both novel and repeated images (Fig. 3a; for repeated images: Pearson's r(206) = 0.70, p=6.7e-32, novel images r(206) = 0.74, p=4.9e-38). This was a large effect: the difference in firing rates between two novel images of low versus high memorability was approximately 4x greater than the magnitude of the difference between a novel and repeated viewing of the same image. Any scheme attempting to read out memory from ITC would thus need to somehow extract changes in firing due to memory in the presence of the much stronger confounding factor of stimulus-evoked memorability effects.

In comparison, in HC, the correlation between memorability and firing rate for repeated images was statistically indistinguishable from no relationship (Fig. 3b; r(206) = -0.01, p=0.935), and it was only modestly positive for novel images (r(206) = 0.20, p = 0.004). To statistically compare memorability modulation in ITC and HC, we fit linear regressions individually to both novel and repeated images in each brain area (solid lines on Fig. 3a-b) and tested whether the slopes differed between ITC and HC via a permutation test of the null hypothesis that the slope of the MB vs GMFR relationship was the same in ITC and HC (see Methods). We found that

memorability modulation was significantly attenuated in HC as compared to ITC for both repeated (ITC β=6.05 and HC β=0.49, H0: $\Delta\ \beta = 0$, $p = 0.0047$) and novel (ITC β=5.15, HC β=-0.01, H0: $\Delta\ \beta = 0$, $p = 0.0042$) images. The significant attenuation of memorability modulation in HC relative to ITC was observed in both monkeys individually (Fig. S5), as well as in different analysis windows (Fig. 4, black arrows).

Memory signals were larger for more memorable images in both ITC and HC

We next explored whether and how the magnitude of repetition suppression (computed as the difference in grand mean firing rate to a novel versus a repeated image) changed with memorability, with the rationale that if it is related to image memorability behavior, repetition suppression should increase with memorability. Indeed, the two were significantly correlated in ITC ($r(206) = 0.23$, $p = 0.001$; black arrows on Fig. 3a quantified in Fig. S6). This could follow in a straightforward way from repetition suppression that is approximately proportional to novel firing rate (e.g., 10% of 10 leads to a reduction of 1; 10% of 50 leads to a reduction of 5).

We also observed greater repetition suppression for more memorable images in HC ($r(206) = 0.17$, $p = 0.014$; Fig. S6). It is more difficult to envision how this might follow from processing within HC alone: if repetition suppression arose independently in HC, it's unclear what would lead to more of it for more memorable images, given the near absence of HC memorability modulation. This suggests that HC repetition suppression might be inherited from earlier structures, including ITC.

Evidence for a computational transformation between ITC and HC

To conceptualize differences between ITC and HC in relation to the monkeys' behavior, we visualized in a schematic how different from the observed slopes the MB-GMFR relationship (solid lines in Fig. 3a-b) would need to be to decode memorability behavior from the firing rates in each brain area (Fig. 3c). Attempting to read out the raw ITC firing rates as a magnitude code based on RS would produce highly misaligned predictions, requiring a large rotation of the space to match predictions to actual behavior (Fig. 3c left, cyan arrow). In contrast, in HC, the firing rates were nearly aligned with behavior (Fig. 3c right inset) and only a subtle adjustment would be needed (Fig. 3c, cyan arrow). This implies that the signals in HC are closer to those used to drive behavior, and it suggests that a computational transformation applied to ITC produces the responses observed in HC.

To summarize, the results shown in Figs. 2 and 3 suggest that a previously undescribed computation occurs between ITC and HC to extract memory information while leaving behind stimulus-evoked memorability modulation. That memorability affected overall firing rates in ITC (Fig. 3a), confirms that the ‘familiarity-memorability paradox' is reflected there. In comparison, this effect was attenuated in HC (Fig. 3b). At the same time, the magnitude of repetition suppression was larger for more memorable images in both ITC and HC (Fig. S6). Together, these results suggest that repetition suppression may follow from computations within (and perhaps before) ITC, and downstream computations retain *memory* while discarding modulation due to *memorability* to create the responses reflected in HC.

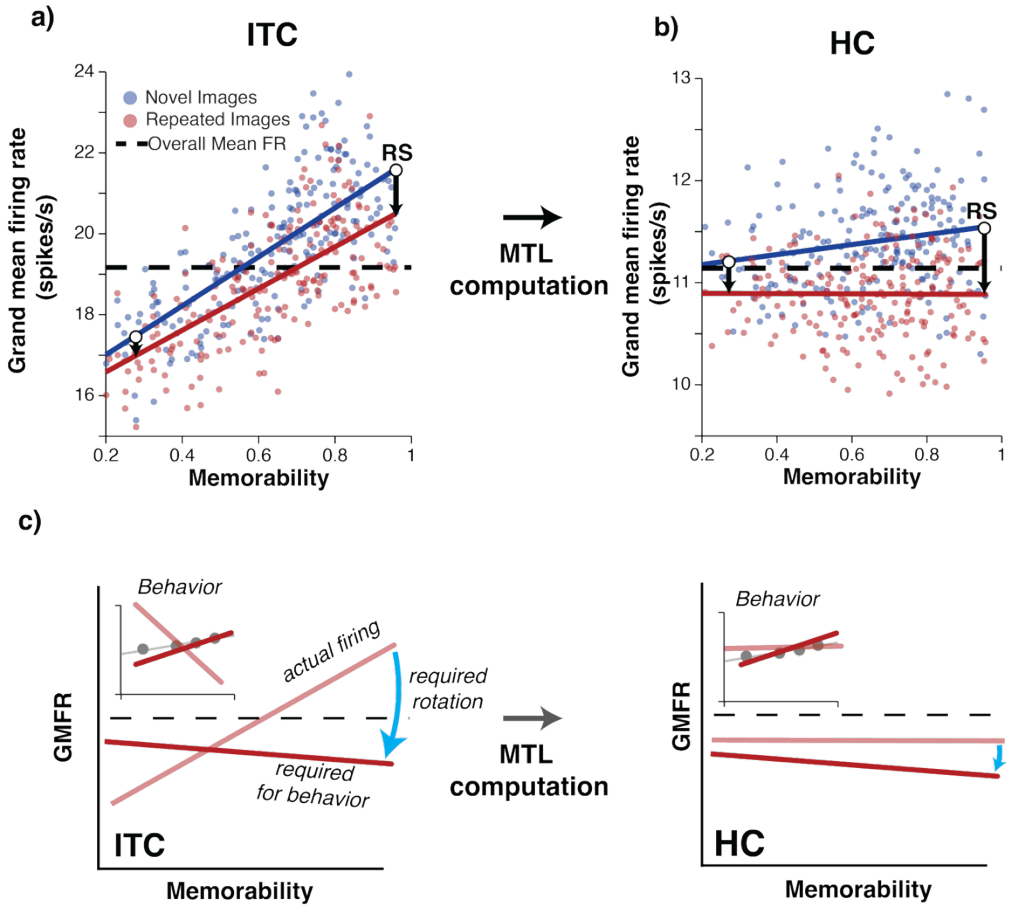

Figure 3. *Memory, memorability, and evidence for a computational transformation between ITC and HC.* **a-b)** Neural data recorded in ITC (panel a) and HC (panel b) with spikes counted in the 100-500ms window (ITC) and 300-500ms window (HC) following image presentation. Each dot represents the grand mean firing rate across all units (ITC: n=602 units, HC n=836 units) to a novel (blue) or repeated (red) image of a given memorability. Average firing rate across all units and images is shown as a dashed line; this line would be the decision boundary in the most straightforward version of the repetition suppression hypothesis (any image above the line would be decoded as "novel", and below the line would be decoded as repeated). In panel a (ITC), this decoding scheme would lead to poor predictions of observed memorability behavior (panel c left inset, compare pink line versus gray dots) because higher memorability repeated images land above the overall mean firing line and would thus be classified as novel; however, these images were the most likely to be actually remembered (Fig. 2a). In comparison, this decoding scheme applied to panel b (HC) would align much better with memorability behavior (panel d left inset). Gray arrows indicate repetition suppression, which increased with memorability in both ITC and HC. The arrow between a) and b) indicates the suggestion that this computational transformation takes place in the medial temporal lobe (MTL). **c)** Schematic highlighting the approximate amount of rotation required to map ITC and HC to behavior, suggesting a computational transformation between them. In each panel, the firing rate relationship between memorability and grand mean firing rate (GMFR) for repeated images is indicated with the light red line. To reach a relationship that would be plausible to read out to reach the observed memorability behavior (inset, gray dots) from a magnitude-based classifier (where more memorable images are farther away from the decision boundary) a large rotation (left, cyan arrow) of firing rates would have to occur for ITC (approximate predictions of behavior to each line are shown overlaid on the inset). In contrast, for HC, a smaller rotation (right, cyan arrow) would be required. For clarity, only the 'repeated' firing trend is shown, but the same holds for novel images.

The transformation between ITC and HC is non-trivial

Next, we turned to understanding how the brain might implement a transformation between ITC and HC to eliminate memorability while preserving memory. We first wondered if this computation was reflected within ITC itself across time, where memorability might be reflected early in the stimulus-evoked response but absent later. This was not the case (Fig. 4). In ITC, firing rate remained highly correlated with memorability across the entire 500 ms viewing period. In comparison, memorability was only transiently reflected in HC and dropped to levels indistinguishable from no correlation by approximately 300ms following stimulus onset. Notably, the late window, when memorability was most attenuated in HC, was *also* the time when memory information was the strongest (Fig. 2c). This result suggests that a "late readout" of ITC alone cannot plausibly explain the transformation from ITC to HC.

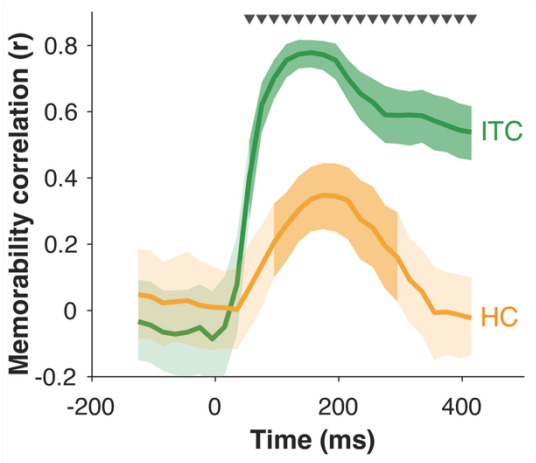

Figure 4: *Comparison of the time course of memorability's influence on firing rate between ITC and HC.* Pearson's correlation between grand mean firing rate for repeated images and image memorability, plotted as a function of time from image onset (150ms spike count window, 20ms sliding intervals). Shaded regions reflect 95% confidence intervals, computed by bootstrapping with 10,000 resamples. Darker shadows indicate times when the correlation is significantly different from zero ($p<0.05$). Black triangles indicate timepoints when the correlations in ITC and HC were significantly different (bootstrapping, H0: $\Delta r = 0$; $p<0.05$).

Following on evidence that visual representations in HC are more sparse than those in ITC (26, 27), a feature that may result from the feedforward circuitry of the MTL and support decorrelation of signals in the form of pattern separation (28, 29), we next wondered if the computational transformation of memorability from ITC to HC could be caused by simple thresholding. Specifically, we tested whether HC units decoding ITC units with a threshold that chopped off the flanks of ITC tuning curves could account for the transformation observed in the neural data. For a mechanism like this to be plausible, memory information (and increased memory for more memorable images) must be retained, but the modulation by memorability of the memory signal should be attenuated.

To evaluate this proposal, we performed a simulation by fitting a model to each ITC unit that describes its visual tuning and memory sensitivity so we could simulate the effects of sparsifying

ITC representations (Fig. 5a, see Methods). We confirmed that the model ITC population recapitulated salient aspects of our data, including the correlation between population response vigor and memorability (Fig. S7). Next, we applied the thresholding operation described in Fig. 5a, where *N*% of the flank of each unit's tuning function was set to a threshold, explored for the full range of *N* (0-100%). The resulting model populations were analyzed for both their memorability correlations (Fig. 5b) and memory classifier performance (Fig. 5c).

We found that matching the observed HC memorability correlation required thresholding a remarkable 93% of each unit's ITC tuning function (Fig. 5b). The consequence of this amount of thresholding was the elimination of nearly all memory information (Fig. 5c), at odds with the actual HC population whose memory performance is matched to ITC. As such, simple thresholding (leading to sparsification) cannot account for the transformation from ITC to HC.

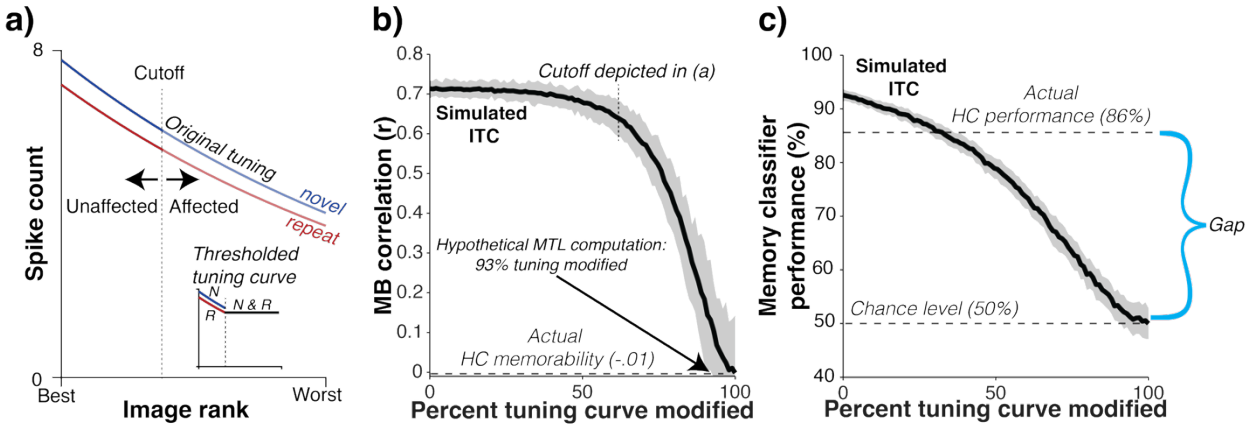

Figure 5*: Thresholding ITC could not explain the transformation to HC.* **a)** To evaluate the plausibility of the thresholding proposal, we performed a simulation based on tuning curves fit to each unit. Shown is an example unit. The red and blue lines correspond to the firing of this unit to novel and repeated images, arranged from best (highest evoked firing) to worst (lowest evoked firing) images on the x-axis. The vertical dashed line denotes the threshold at which the tuning curve is modified, and the inset shows the tuning curve after the modification. To assess the impact of thresholding, the same threshold (e.g., 60% as shown in panel a) was applied to all ITC units, and the modified tuning curves were used to generate a simulated population; this was then repeated for a range of thresholds. **b)** The correlation between memorability and GMFR for the simulated populations, plotted as a function of the percent of the tuning curves that were modified on all units. To match the observed memorability correlation in HC, 93% of the tuning curve would have to be thresholded away. **c)** Memory information remaining in each simulated population after thresholding, assessed by a weighted linear decoder. The threshold required to match memorability (highlighted by a cyan bracket) destroyed nearly all memory information, at odds with the reflection of robust memory information in the HC data, implying that this proposal is implausible. Across the figure, the error shadow depicts the 95% confidence interval. The same analysis windows were used as in Fig. 3a-b.

A linear decoder is capable of selectively extracting memory signals from ITC

Given the failure of simple mechanisms to account for behavior, we next asked if a more advanced but still relatively simple linear decoder could account for the transformation by selectively extracting memory information from ITC while leaving behind stimulus-evoked memorability modulation. Specifically, we built on the observation in Fig. 3c that HC's representation of firing rates as a function of memorability reflects a 'rotated' version of ITC. Broadly, we evaluated this proposal by projecting the data onto a plane defined by repetition suppression and memorability decoding axes, and sweeping through linear decoders at different orientations within this plane to identify whether behavior could be predicted by one of them using this scheme. If so, it follows that the transformation observed in Fig. 3 could result from a linear decoding of ITC. A previous study (16) demonstrated that a similar decoding scheme could disambiguate firing rate modulations due to contrast and memory in ITC. However, unlike contrast, where higher firing rates do not lead to stronger memory, the higher firing rates associated with memorability lead to better memory performance. Thus, it is possible, but not given, that such a scheme would work for memorability.

The decoding scheme we evaluate here proposes that as long as representations of memorability and repetition suppression are not perfectly overlapping, each could be decoded separately from the other (if the two completely overlapped, it would not be possible for a linear scheme to separate them). To assess the plausibility of this proposal, we defined two classifier axes in the space defined by the population neural spike counts: a repetition suppression-based classifier (RS) that weighted each unit equally (equivalently, the grand mean firing rate vector), and a memorability classifier (MB) trained to separate images with high versus low memorability (see Methods). We found that two decoding axes were 51.1 degrees apart in ITC (Fig. 6a), confirming that they were overlapping, but only partially so (where fully non-overlapping would correspond to orthogonality or equivalently 90 degrees).

To test if ITC neural representations could be linearly decoded to predict memorability behavior, we projected the data onto the plane defined by the MB and RS classifiers (Fig. 6a). The data projected onto this plane reflected the same properties highlighted in Fig. 3: The ovals (representing one standard deviation contours) span a longer length along the MB axis than the RS axis because there is more modulation by memorability than by repetition suppression in ITC. The familiarity-memorability paradox is illustrated here by the highest-memorability repeated images landing on the 'novel' side of the RS decoder (the data here are mean-centered, so the decision boundary is the origin), highest memorability images (darkest red) are to the right of the boundary, meaning those high memorability images would be decoded as 'novel'; see also Fig. S8).

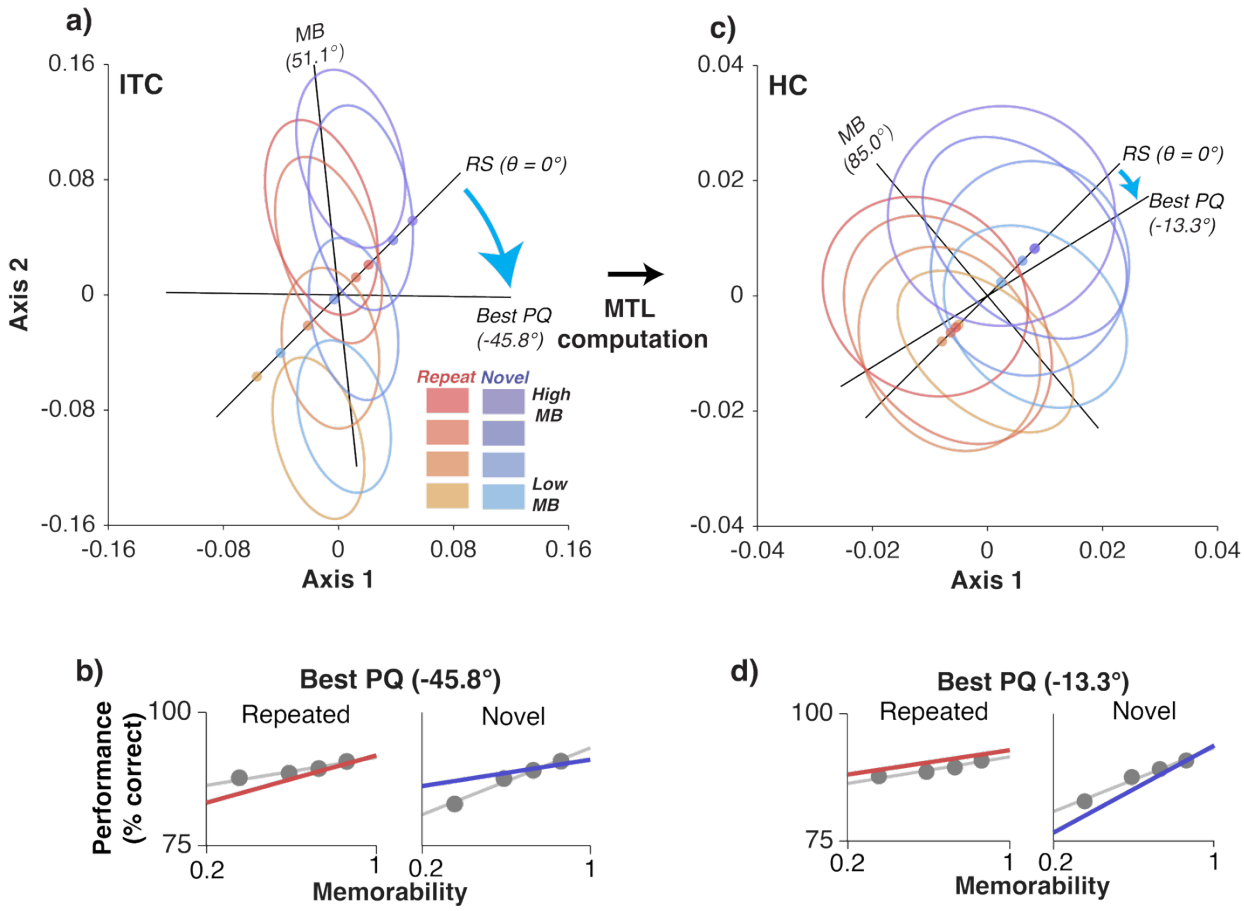

Figure 6. *A memorability-correcting decoder applied to ITC and HC predicts behavior.* **a)** Projections of ITC neural data onto a plane defined by repetition suppression (RS, where every unit is weighted equally or equivalently the grand mean firing decoder) and a memorability decoder (MB, optimized to decode high versus low memorability images). Ellipses indicate the one-standard-deviation contour for 2D histograms of the projections of ITC population responses onto this plane. The eight ellipses correspond to novel (blues) and repeated (reds) images grouped by memorability scores into quadrants (hue). Translucent dots are projections of the means of each MB bin onto the RS classifier. Also shown is the decoder on this plane with the best ability to map neural data to behavior (Prediction Quality, PQ). **b)** Comparison of behavior (gray) with neural predictions (red for repeat, blue for novel) for the best PQ decoder. Here, the neural predictions are rescaled to the same range as behavior with a multiplicative factor, loosely consistent with an adjustment in population size (see Methods). **c-d).** Same as a-b, applied to HC. Compared to ITC, a much smaller rotation (cyan arrow) from the RS decoder is required to achieve alignment with behavior. Across this figure, the same analysis windows as Fig. 3 are used. The arrow between a) and b) indicates the suggestion that a computational transformation takes place in the medial temporal lobe (MTL).

We then quantified the ability of a suite of linear decoders that populated the MB-RS plane to account for behavior by sweeping through the full set of possible decoding axes, asking if the geometry in this space could support a readout that aligned well with memorability behavior. At each rotation, the behavioral predictions of the linear decoder were compared with behavior for repeated images (In essence: was decoding performance higher for more memorable images?).

The alignment was quantified with a prediction quality metric (PQ, with higher PQ reflecting better alignment; see Fig. S9a and Methods).

The behavioral predictions for the decoder with the best PQ for repeated images are plotted in Figs. 6b, and S8, along with the decoder's prediction of performance for novel images (whose alignment did not determine it). The best PQ for ITC fell nearly orthogonal (90 degrees away from) to the MB classifier, at 96.7 degrees. Intuitively, this makes sense because 90 degrees corresponds to memory modulations with no memorability influence, and the goal is to extract memory information largely independent of memorability. The fact that the best PQ vector falls a bit further than 90 degrees away from the MB classifier aligns with positively sloped behavioral performance as a function of memorability. The best PQ also fell 45.8 degrees away from the RS classifier, suggesting that it was partially overlapping with RS but not identical to it.

Above, we ruled out thresholding as a mechanism with the rationale that the amount of thresholding required would destroy nearly all memory information (Fig. 5). We applied that same logic to this proposal by computing average memory performance across the four bins for both RS and the best PQ vector. Memory information was nearly perfectly maintained at the point of best PQ as compared to RS (Fig. S9b). That this decoding scheme produces good predictions of behavior and leaves memory information intact suggests that it is a plausible candidate for the MTL computation that transforms ITC representations into a form that can drive familiarity behavior.

Next, we applied the same logic to the HC representations. In HC, the MB and RS axes were closer to fully orthogonal than in ITC (85.0 as compared to 51.1 degrees), and novel and repeated images largely fell on opposite sides of the decision boundary when projected onto the RS axis, reflecting the intuition from Fig. 3b that the familiarity-memorability paradox was no longer present in HC. The hypothesis that a computation occurs between ITC and HC to perform a partial rotation predicts that the HC representation should require less rotation than ITC to align with behavior (like Fig. 3c). Consistent with that proposal, a much smaller rotation (13.3 degrees; Fig. 6c cyan arrow) was required to align HC representations with behavior (Fig. 6d). The difference in the amount of rotation required for ITC (45.8 degrees versus 13.3 degrees) suggests that a computation has taken place between them. Like ITC, there was no loss in memory information when decoding at the best PQ as compared to RS (Fig. S9b).

In sum, these results suggest that while memory and memorability information both overlap with repetition suppression in ITC, that overlap is partial. As such, a linear decoding scheme can be designed to selectively extract memory while leaving behind stimulus-evoked memorability modulation. A decoder of this type provided a good account of the transformation from ITC neural responses to memorability behavior. Furthermore, when applied to HC representations, less rotation was required to align the representations to behavior. This suggests that structures in the MTL perform this type of computation between ITC inputs and HC responses to selectively extract memory signals while retaining larger magnitude signals for more memorable images.

Discussion:

Here, we present evidence for a previously undescribed computation between ITC and HC that extracts memory information while leaving behind stimulus-evoked memorability modulation (Fig. 3). While simple thresholding cannot account for this computational transformation (Fig. 5), a selective linear readout that preserves larger memory signals for more memorable images can (Fig. 6). These results resolve the familiarity-memorability paradox about how memory and memorability are reflected in ITC — where *suppressed* firing rates are thought to signal memory, but more memorable images evoke *higher* firing (Fig. 1) — by demonstrating that both are reflected in the ITC population response, but only one (memory) is extracted downstream. Consequently, in HC, more memorable images evoke only negligibly higher firing, but larger memory signals for more memorable images remain (Fig. 3). In other words, the resolution to the familiarity-memorability paradox is that descriptions asserting familiarity are reflected in ITC population response vigor are not entirely accurate. Instead, familiarity is reflected in ITC in a *component* of population response vigor that lies orthogonal to stimulus-induced memorability modulation, and it is this component that is extracted downstream and reflected in HC.

These results lead to a simple, highly plausible account of how seeing an image is transformed into a memory that it has been seen. The process begins with more memorable images producing more vigorous responses in ITC (Fig. 7, left). This, in turn, causes more adaptation-like repetition suppression. Under the assumption that repetition suppression is a proportional reduction in the stimulus-evoked response, the more vigorous responses associated with more memorable images lead to larger reductions (e.g., 10% of 10 leads to a reduction of 1; 10% of 50 leads to a reduction of 5). At this stage, memory cannot be unambiguously decoded from population response vigor, leading to the familiarity-memorability paradox. Next, computations downstream of ITC selectively extract memory while leaving behind memorability such that a largely isolated memory representation is reflected in HC (Fig. 7, right). Crucially, this transformation preserves larger memory signals for more memorable images, thereby making it a plausible proposal of the signal that drives memorability behavior. This computational transformation between ITC and HC likely occurs in one or more MTL structures and may arise gradually across the MTL or singularly within HC.

Four key pieces of evidence lead to this account: (1) Downstream of ITC, in HC, memorability modulation is largely attenuated but memory modulation remains, leaving familiarity in the form of RS in HC (Fig. 3a-b & Fig. S6); (2) More processing is required to align ITC versus HC representations with behavior (Fig. 6b vs 6d; a 46 versus 13 degree rotation) suggesting the existence of an ITC-to-HC transformation; (3) The representational geometry in both ITC and HC support reading out memory in this way at a minimal cost to memory information (Fig. S9b) (4) Memory information in ITC can be decoded in this way to account for memorability behavior, suggesting all the information necessary to support memorability behavior is present in ITC (Fig. 6a-b).

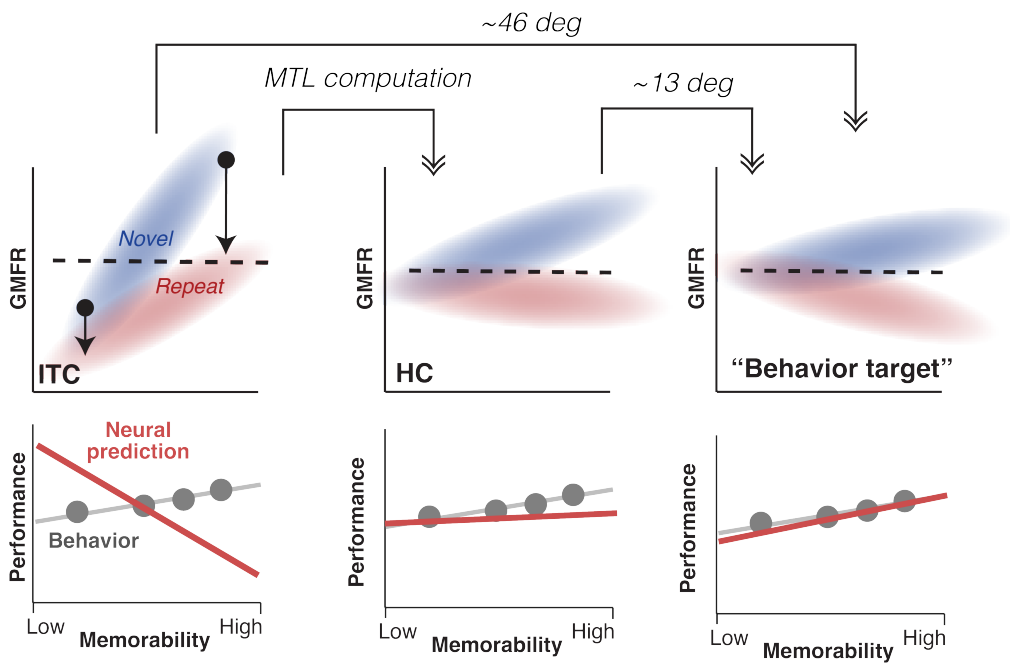

Figure 7*: Schematic of how seeing is transformed into memorability behavior.* First, in ITC, more memorable images evoke more vigorous firing rates, leading to more repetition suppression (Top, Left black arrows). However, decoding firing rate creates a prediction misaligned with actual behavior (Bottom, Left). All the information required to drive behavior is reflected at this stage, but a large rotation (~46 degrees) is required to extract it using a total spike count coding scheme. Next, MTL computations produce a rotated representation in HC that better aligns with behavior (Middle). Finally, the same computations applied to HC produce the final rotation (~13 degrees) required to drive behavioral responses. Top: Blue clouds are the distribution of firing rates to novel presentations of images, while red clouds are the distribution to repeated presentations. Black arrows with filled circles indicate repetition suppression (RS), and the difference in size as a function of memorability highlights the relationship between memorability and RS (where more memorable images lead to stronger memory signals). The dashed line indicates the overall mean firing rate, which would be used as the classifier threshold for the "RS hypothesis" of visual memory. Bottom: The approximate prediction of behavior based on the neural data sketched in the top row. Gray dots indicate the real behavior, while the red line shows the prediction.

An alternative class of models one might consider is one in which the neural signals that drive familiarity behavior are created in HC (not ITC), but this seems unlikely. In a first scenario, stimulus-evoked memorability modulation is eliminated before signals reach HC. Here, it is unclear what stimulus-evoked properties would lead to larger HC memory signals for more memorable images. In a second scenario, stimulus-evoked memorability information could be reflected in the input to HC, and HC could perform computations that simultaneously store memories while also removing memorability modulation in a manner that the former but not the latter is reflected in HC activity. However, evidence from lesion studies in humans (30) and animals (31) shows that damage linked to HC leaves overall familiarity behavioral performance intact, suggesting that HC is not the sole locus of familiarity storage. As such, this scheme seems unlikely as well. Instead, it may be the case that HC participates by performing the computation we have described, extracting memory information from the combined familiarity-memorability population response that arrives as its input.

One striking property of memorability behavior is that typical humans (5) (and monkeys (7)) tend to find the same images more and less memorable. The proposal we present in Fig. 7 predicts that damage to the MTL, where we propose that the memory extraction computation happens, should lead to different patterns of memorability behavior in MCI patients than in healthy controls. Relevant to that prediction, one study (32) investigated individuals with mild cognitive impairment (MCI, which is associated with MTL atrophy (33)). This study found that there was some overlap in images that both healthy controls and patients with MCI found memorable, but memorability scores were less predictive of memory performance in patients with MCI than for healthy controls. However, it's unclear whether this resulted from the specific memorability changes our results predict, or from overall reduced memory performance and inter-subject consistency in these patients — further investigation of this subject is thus warranted.

While the answer could have, in principle, been different, the familiarity-memorability paradox turned out to be more of a misunderstanding by scientists than a computational challenge faced by the brain. Many previous reports have suggested that familiarity is reflected in ITC and elsewhere as RS (9–15). It may be tempting to propose that RS could be decoded directly, but such explanations are ill-defined absent suggestions of how the brain would "remember" the response to a novel image when viewing a repeated one, such that a reference could be made to it. The simplest, first-order explanation is that population response vigor is decoded relative to a threshold (7), but that leads to the familiarity-memorability paradox (Fig. 1). Here we demonstrate that descriptions asserting that familiarity is reflected as ITC population response vigor are not complete either. Rather, familiarity is reflected as a component of population vigor that lies approximately orthogonal to stimulus-induced population-vigor responses such as memorability (Fig. 6).

Given the finding that familiarity behavioral patterns are linearly decodable from ITC, why might the brain bother to remove “nuisance” variation downstream? While unclear, we can speculate. Whereas "pattern-of-spikes" coding schemes (where some neurons fire more than others to convey information) are associated with variables such as the identities of the objects in the world, "magnitude" coding schemes (where variables are reflected in the vigor of a population or subpopulation) have been associated with the degree to which that information should be prioritized (34). Once familiarity information is encoded, eliminating memorability modulation could free up magnitude coding bandwidth for higher-order priority-related variables, such as ones associated with reward, attention, and prediction. That we observe a difference in familiarity representations between ITC and HC is notable in light of recent work showing that in the case of the high specificity of visual familiarity (tied to the hippocampus via fMRI, modeling, and natural lesion studies (28, 29, 35–37)), ITC and HC representations are largely aligned (15).

Future work will be required to determine how the MTL might learn to decode ITC in this way. The memorability classifier used in Fig. 6 was trained with outside knowledge of which images had high or low memorability, and in question is how the brain might have access to that variable or a proxy for it. The content of images can be predictive of memorability (38), creating a defined statistical landscape (39) that the brain could, in principle, learn for this purpose. Underscoring this point, neural networks trained to predict memorability are able to do so for

unseen images (5, 40), and the brain could learn these statistics as well. Likewise, for this scheme to work, there must be meaningful heterogeneity in the sensitivities of individual ITC units to memory and memorability (or their representations would completely overlap and the angle difference between their decoding axes would be zero). That the decoding scheme implemented in Fig. 6 finds these axes are separated in ITC and faithfully maps responses to behavior reveals that this heterogeneity exists in ITC, and it suggests that the MTL could use these differences to selectively extract memory information.

While the simulations shown in Fig. 5 show that a high-thresholded-readout (something that might generate the sparsity that supports decorrelation and pattern separation in HC) cannot by itself lead to attenuated memorability representations, our results do not rule out other schemes. For example, hippocampal place cell-like representations can arise via an autoencoder-based scheme designed to compress memories (41), and the architecture of hippocampus has been linked to learning statistical regularities in the environment (42). The attenuation of memorability observed in HC relative to ITC could conceivably be linked to these or other computations.

Understanding how the healthy brain transforms experiences into memories is foundational for understanding memory-related disorders. Here, we resolve a paradox related to how that transformation happens. By focusing on the nature of the neural code that the brain uses to drive familiarity reports, we uncover a previously undescribed computation that may help make sense of the behavioral patterns of individuals with MTL damage. In addition, we present a model of how visual memories are stored and signaled that can be tested and refined in the broader pursuit of understanding how memory and its underlying signals come to be, both in health and disease.

Methods:

All data were collected from two adult rhesus macaque monkeys (*Macaca mulatta*). Monkey 1 was male and monkey 2 was female. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

Anatomy and recording: The activity of neurons in ITC and HC were recorded via a single recording chamber (Crist Instruments, Hagerstown, MD) in each monkey. In monkey 1, we recorded from left ITC and HC. In monkey 2, we recorded from right ITC and HC. Chamber placement was guided based on anatomical magnetic resonance imaging. ITC and HC were recorded in separate sessions. Neural signals were recorded with 24-channel U-probes (Plexon Inc., Dallas, TX) with recording sites spaced at 100 μm intervals. Wideband signals were amplified and digitized at 30 kHz using a Grapevine Data Acquisition System (Ripple, Inc., Salt Lake City, UT). Spikes were hand-sorted offline (Plexon Offline Sorter).

The region of ITC recorded was located on the ventral surface of the brain, located 17 mm lateral to the sagittal plane defined by the midline and 13-16mm anterior to the coronal plane defined by the ear canals. The recorded units were in anterior ITC, where repetition suppression in response to familiarity has been previously observed (10); this regions is similar to that reported in ref (14).

The region of HC recorded was 11 mm lateral to the sagittal plane defined by the central sulcus and 11-16 mm anterior to the coronal plane defined by the ear canals, and only includes areas roughly in the dorsal half of HC (this targeting was guided by ref (24)). This means that our sample predominately includes hippocampal subfields CA3, DG and CA4 (approximately in that order of relative prevalence) but because of the organization of the primate hippocampus and the inherent ~+/- 1 mm level of precision in our recording preparation, we do not attempt to distinguish between the subfields. However, because subfield CA1 is located ventrally in primate hippocampus and we targeted our recordings to the dorsal half of HC, we are reasonably confident it is not represented in these populations.

We did not observe any obvious functional organization within the regions recorded in ITC or HC.

Inclusion Criteria: A multi-channel recording session was included in the analysis if (1) there were no technical problems with the recording hardware, (2), it fell within the anatomical boundaries described above, (3) the monkey completed at least 200 novel-repeat image-pairs during the session, (4) the session was stable, defined as the overall baseline firing rates across all channels did not change by more than 2-fold across the session, and there was no significant difference between baseline firing rates preceding novel and repeated images (t-test, $p<.01$, in the 500ms prior to image presentation) and (5) the session was composed of over 50% visually responsive neurons (in ITC) or contained at least a single visually responsive neuron (HC), determined by an inspection of the raster plots. We chose this relatively strict criterion for ITC to ensure high-quality data based on our previous experience recording this highly visual brain

area (based on extensive previous experience, a lack of visually responsive units in a nominally ITC session suggested the probe was, actually located in white matter dorsal to ITC), and a much looser criterion for HC because we did not wish to make any *a priori* assumptions about its visual organization. Approximate memory information was matched between populations by truncating a handful of sessions (criterion 6). Within a session, units were included in the population if they evoked a response in the post-stimulus window [50-350ms] versus the pre-stimulus [-300,0] window (two-sided t-test, $p<0.10$) across all trials in the session. In monkey 1, this filter passed 96% and 52% of units for ITC and HC respectively, and for monkey 2 this filter passed 93% and 57% of units for ITC and HC respectively.

Monkey 1's ITC population was collected over the course of approximately five weeks. We collected 20 sessions and 13 were removed (1, 1, 1, 6, and 4 from criteria 1, 3, 4, 5 and 6 respectively).

Monkey 1's HC population was collected intermittently over the course of approximately seven months. We collected 36 sessions and 20 were removed (3, 8, 3, and 6 from criteria 1, 3, 4 and 5 respectively)

Monkey 2's ITC population was collected intermittently over the course of approximately four months. We collected 9 sessions and 1 was removed (from criteria 1)

Monkey 2's HC population was collected over the course of approximately two months. We collected 30 sessions and 9 were removed (2, 3, 2, 1, and 1 from criteria 1, 2, 3, 4, and 5 respectively)

Behavioral task: A custom MWorks script (The MWorks Project, Cambridge, MA) was used to display images on an LCD monitor and collect behavioral responses. The experiment was performed in a darkened testing chamber with the monkeys' heads fixed while their gaze was tracked with an Eyelink 1000 (SR Research, Ottawa, Canada). Trials were initiated by the monkey fixating on a red square fixation point (0.35 degrees) in the center of a gray screen within an invisible square window (4 degrees monkey 1, 3.5 degrees monkey 2), followed by a 500ms delay before a 4-degree image appeared at the center of the screen. The monkeys had to maintain fixation on the fixation point for 500ms, after which the image disappeared from the screen and the fixation point changed to a green go cue, and two square white .5 degree response targets appeared 8 degrees above and below the stimulus. The monkey indicated a response by making a saccade to one of the two targets. If the monkey was correct, they received a liquid reward. For Monkey 1, the target associated with "novel" images was the top target; this was reversed for Monkey 2. If the monkey broke fixation, the image immediately disappeared, and a new trial was offered after a short delay. In cases where the stimulus was novel but the monkey broke fixation (after an image had appeared on the screen), the subsequent presentation of that image was still rewarded as a "repeat," although incomplete image pairs were excluded from further analysis.

A custom MATLAB (The MathWorks, Inc., Natick, MA) script was used to pseudo-randomly generate a sequence of images, where each image appeared exactly twice (the same algorithm as ref(14)). The script aimed to match a target distribution of n-back trials. To fill the sequence, some occasional off-target n-backs were used but were not included in analysis. Due to behavioral differences between the two monkeys, the exact n-back distributions varied between Monkey 1 and Monkey 2.

Monkey 1 was presented with n-back trials of 1, 2, 4, 8, 32, 48, 64, and 192. The distribution was uniform for n-backs between 2 and 64, while 1 and 192 were half as frequent (e.g., in a sequence, there could be 15 1-back trials, 30 2-back trials, 30 4-back trials, and so on).

Monkey 2's sequence included n-backs of 1, 2, 4, 8, 16, 32, 48, and 64, with a uniform distribution between 2 and 48 and half-frequencies for 1 and 64. N-backs differed between monkeys in order to approximately match performance between animals.

The images used were drawn from the same set used in Jaegle et al. (2019)(7), and images used during training a monkey were never reused during recording.

Pseudopopulation: In the pseudopopulation, a response to a given unit always contained the novel/repeated pairing of the same real images, and pairs were aligned so that a pseudoimage always included a similar n-back separation and images with similar memorability scores. We also used these alignments to pool the data between monkeys. This process of pooling and aligning limits the number of images included in the pseudopopulation to the number of eligible images in the shortest recording session, and therefore some trials from other sessions were discarded. This was done pseudo-randomly and not biased to any experimental conditions.

Permutation test for difference in slopes (Fig. 2)**:** To perform this test, a null hypothesis distribution was generated by randomly assigning an “IT” or “HC” label to each data point, without replacement, then fitting lines to these shuffled “populations” and calculating the difference in slopes between the two “brain areas”. This procedure was repeated 10,000 times to generate a distribution of the null hypothesis that there was no difference between the brain areas. P-values were calculated by computing the proportion of null-distribution differences in slopes that differed from the actual slope.

Unitwise sensitivity to memory (Fig. 3a): d’ for a unit was calculated by:

$$d' = \frac{\mu_n - \mu_r}{\sqrt{\frac{\sigma_n^2 + \sigma_r^2}{2}}} \tag{1}$$

Where μ_n and $\mu_r$are the mean to novel and repeated images, respectively, for that unit, and σ_n^2 and σ_r^2 are the variance within novel and repeated images, respectively, for that unit.

RS and FLD classifiers: These classifiers were cross-validated and took the form of linear discriminators in which the class of a population response vector to a pseudoimage was determined by the sign of:

$$f(\boldsymbol{x}) = \boldsymbol{w}.\boldsymbol{x} - b \tag{2}$$

Where **x** is the population response vector, **w** is an n-dimensional weights vector in the n-dimensional ITC or HC neural space (n being the number of units), and b is a decision boundary given by:

$$b = \frac{1}{2} * \boldsymbol{w}.(\mu_1 + u_2) \tag{3}$$

Where μ_1 and u_2 are the means to the two classes (novel and repeated).

For the FLD decoder, we used the same implementation as ref (14, 16):

$$\boldsymbol{W}_{FLD} = \Sigma^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_r) \tag{4}$$

Where Σ^{-1} is the inverse of the covariance matrix averaged across both conditions. Because the dimensionality of the neural data is very high relative to our sample size, we cannot make good estimates of the off-diagonal terms in the covariance matrix so they are set to zero. This results in the classifier effectively weighting each unit by its d'.

Because memorability confuses classifiers of this type (as described in Fig. 1), for the implementation of the FLD classifier in Fig. 3 and S3, the pseudopopulation alignment by memorability was broken by randomly shuffling the images that compose pseudoimages (within each unit). This means that every pseudoimage was composed of random images, ameliorating the influence of memorability.

Ranked FLD (Fig. S3): The same implementation of the FLD classifier described above was used. Weights for all units were calculated within each cross-validation iteration, then units were ranked by those weights, subsets of units were removed, and the weights (and threshold) were re-calculated before testing the classifier.

Population fits/synthetic population generation and modification (Fig. 4): We fit units following an adaptation of the method described by ref (16). In brief, this involved fitting a three-parameter exponential model for each unit that maximized the likelihood of observing the spike count data from 100-500ms:

$$y(x; M) = A_M e^{-\alpha x} \tag{5}$$

Where *x* is the stimulus rank along the tuning curve, M is the memory condition (novel or repeated), A is amplitude (A_N and A_R, for novel and repeated images, respectively, fit separately), and α controls the sharpness of the curve (to capture image selectivity).

To generate synthetic populations based on our observed data, we ranked the observed spike counts for each image seen (averaging across novel and repeated presentations) to get an *x* value (image rank) for each image for each unit. These values were then used to solve for a predicted spike count for a given image and unit using the fit constants for A_N, A_R and α. We then used a Poisson process to generate a synthetic spike count. Memorability in these simulations derive from the memorability scores of the originally shown images

To generate the 'modified populations in Fig. 4b-c, the predicted firing rate for both novel and repeated was changed to that of the repeated firing for any image ranked lower on the tuning curve than the 'cutoff point' rank.

Memorability-corrected linear decoder: To train and test a decoder that might account for the transformation between ITC and HC (Fig. 6), we used similar methods to ref (16) , but with RS and MB classifiers. Across each cross-validation split (1000 splits, with 70% of the data used for training), both classifiers were trained and tested on the same set of training data.

The RS decoder took the form:

$$f(\boldsymbol{x}) = \boldsymbol{W}_{RS} \cdot \boldsymbol{X} - b \tag{6}$$

Where $\boldsymbol{X}$ is the mean-centered matrix of spike counts to each unit and pseudoimage, and a value of f(x) > 0 classifies an image as novel, and f(x) < 0 as repeated. **W** evenly weights each unit so that:

$$\boldsymbol{W_{RS}} = [1,1,1 \dots] \tag{7}$$

And:

$$b = \frac{\boldsymbol{\mu}_{novel} - \boldsymbol{\mu}_{repeated}}{2} \tag{8}$$

The MB decoder was similar to the RS classifier, but the weights were defined by:

$$\boldsymbol{W}_{MB} = \boldsymbol{\mu}_{Hmb} - \boldsymbol{\mu}_{Lmb} \tag{9}$$

Where $\boldsymbol{\mu}_{Hmb}$is the vector of unit's mean response to images in the top 50% of memorability scores and $\boldsymbol{\mu}_{Lmb}$ is the mean responses to images in the bottom 50% of memorability scores.

The classifiers rotated within the plane took the form of:

$$f(\boldsymbol{x}) = \boldsymbol{W_\theta} \cdot \boldsymbol{X} - b \tag{10}$$

Where X is the mean-centered matrix of spike counts to each unit and pseudoimage projected onto the MB/RS plane and where W for an angle θ on the MB/RS plane is defined by:

$$\boldsymbol{W_\theta} = (\cos\theta - cot\gamma \sin\theta)\widehat{\mathbf{1}} + (\csc\gamma \sin\theta)\ \widehat{\boldsymbol{w}}_{\boldsymbol{MB}} \tag{11}$$

Where $\widehat{\mathbf{1}}$ is the unit vector representing the RS axis, and $\widehat{\boldsymbol{w}}_{\boldsymbol{MB}}$ is the unit vector representing the MB classifier and γ is the angle between the MB and RS axes. And

$$b = \frac{\mu_{novel} - \mu_{repeated}}{2} \tag{12}$$

Projections onto plane defined by $\widehat{\mathbf{1}}$ and $\widehat{\boldsymbol{w}}_{\boldsymbol{MB}}$ and error ellipses (Fig. 6a and 6c): The data used to visualize the contour ellipses were the same held-out test sets used to train the MB and RS decoders, split into four divisions by memorability quartile boundaries. The ellipses were computed by projecting the data onto the non-orthogonal $\widehat{\mathbf{1}}$ and $\widehat{\boldsymbol{w}}_{\boldsymbol{MB}}$ axes. This coordinate system was then rotated by 45 deg to visualize the RS classifier at this angle. For each condition, the covariance matrix of transformed data was computed, and eigenvectors of this matrix provided the major and minor axes of the associated ellipse. The ellipses themselves represent one standard deviation contours of the 2-dimensional histograms projected onto this plane.

Prediction quality (Fig. S9): To determine the quality of the prediction between actual behavior and that predicted by ITC decoded performance, we divided the testing set for each round of cross-validation into four memorability bins with the quartile boundaries defined by the quartiles of the overall memorability distribution (these were also used to plot the contours, see above). They were further divided by novel and repeated images, and novel and repeated prediction qualities were calculated separately. These test sets were used to create ITC decoded predictions of behavior. To quantify prediction quality (PQ), we fit a linear regression to both the behavioral and neural predicted data and compared the slopes of the regressions such that:

$$PQ = 1 - \frac{\Delta\theta}{90°} \tag{13}$$

Where θ is the angle between the behavioral and neural-predicted slopes. Because slopes separated by 180 degrees are equivalent, any angular difference greater than $90°$ was reflected across the $90°$ deg axis by replacing any difference greater than $90°$ with $180 - \Delta\theta$. This metric yields a value of 1 when the slopes are perfectly aligned and 0 when they are perpendicular. The final PQ plotted in Fig. S9 (and used to determine 'best PQ' in Fig. 6) is the repeated PQ. We used the repeated PQ because we were specifically interested in interrogating the ability of the scheme to read out correctly remembered images.

Behavioral predictions based on classifiers (Fig. 6b & 6d): Any comparison between population decoded predictions and actual behavior requires implicitly or explicitly specifying the

number of neurons included in the decoding analysis. Following on ref (16), we confirmed that performance using all recorded units in our dataset fell below saturation, and then simulated increases in population size by fitting a single rescaling parameter by solving for a single multiplicative scaling factor (i.e. one that does not change the slope of the lines) that minimizes the squared error between the neural predictions and behavior, and then multiplied the predictions by this rescaling factor. Any rescaled predictions over 100% were clamped to 100.

Acknowledgements:

This work was supported by the National Eye Institute of the NIH (award R01EY020851 to N.C.R.) and the Simons Foundation (Simons Collaboration on the Global Brain award 543033 to N.C.R.), and the National Science Foundation (award 2043255 to N.C.R.).

References:

1. P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable? in *CVPR 2011*, (IEEE, 2011), pp. 145–152.
2. W. A. Bainbridge, P. Isola, A. Oliva, The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.* **142**, 1323–1334 (2013).
3. P. Isola, Jianxiong Xiao, D. Parikh, A. Torralba, A. Oliva, What makes a photograph memorable? *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1469–1482 (2014).
4. Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, A. Oliva, Intrinsic and extrinsic effects on image memorability. *Vision Res.* **116**, 165–178 (2015).
5. A. Khosla, A. S. Raju, A. Torralba, A. Oliva, Understanding and predicting image memorability at a large scale in *2015 IEEE International Conference on Computer Vision (ICCV)*, (IEEE, 2015), pp. 2390–2398.
6. N. C. Rust, V. Mehrpour, Understanding image memorability. *Trends Cogn. Sci.* **24**, 557–568 (2020).
7. A. Jaegle, *et al.*, Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife* **8** (2019).

8. G. E. Koch, E. Akpan, M. N. Coutanche, Image memorability is predicted by discriminability and similarity in different stages of a convolutional neural network. *Learn. Mem.* **27**, 503–509 (2020).
9. F. L. Fahy, I. P. Riches, M. W. Brown, Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Exp. Brain Res.* **96**, 457–472 (1993).
10. L. Li, E. K. Miller, R. Desimone, The representation of stimulus familiarity in anterior inferior temporal cortex. *J. Neurophysiol.* **69**, 1918–1929 (1993).
11. E. K. Miller, L. Li, R. Desimone, A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* **254**, 1377–1379 (1991).
12. J. Z. Xiang, M. W. Brown, Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology* **37**, 657–676 (1998).
13. I. P. Riches, F. A. Wilson, M. W. Brown, The effects of visual stimulation and memory on neurons of the hippocampal formation and the neighboring parahippocampal gyrus and inferior temporal cortex of the primate. *J. Neurosci.* **11**, 1763–1779 (1991).
14. T. Meyer, N. C. Rust, Single-exposure visual memory judgments are reflected in inferotemporal cortex. *Elife* **7** (2018).
15. B. G. L. Jannuzi, *et al.*, Sharpened visual memory representations are reflected in inferotemporal cortex. *J. Neurosci.* e0833252025 (2025).
16. V. Mehrpour, T. Meyer, E. P. Simoncelli, N. C. Rust, Pinpointing the neural signatures of single-exposure visual recognition memory. *Proceedings of the National Academy of Sciences* **118**, e2021660118 (2021).
17. W. A. Suzuki, D. G. Amaral, Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *J. Comp. Neurol.* **350**, 497–533 (1994).
18. R. Insausti, D. G. Amaral, W. M. Cowan, The entorhinal cortex of the monkey: II. Cortical

afferents. *J. Comp. Neurol.* **264**, 356–395 (1987).

19. A. D. Garcia, E. A. Buffalo, Anatomy and Function of the Primate Entorhinal Cortex. *Annu Rev Vis Sci* **6**, 411–432 (2020).
20. J. L. Ringo, Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey. *Behav. Brain Res.* **76**, 191–197 (1996).
21. W. A. Suzuki, E. K. Miller, R. Desimone, Object and place memory in the macaque entorhinal cortex. *J. Neurophysiol.* **78**, 1062–1081 (1997).
22. E. T. Rolls, P. M. Cahusac, J. D. Feigenbaum, Y. Miyashita, Responses of single neurons in the hippocampus of the macaque related to recognition memory. *Exp. Brain Res.* **93**, 299–306 (1993).
23. M. J. Jutras, E. A. Buffalo, Recognition memory signals in the macaque hippocampus. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 401–406 (2010).
24. J. J. Sakon, W. A. Suzuki, A neural signature of pattern separation in the monkey hippocampus. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9634–9643 (2019).
25. B. G. L. Jannuzi, *et al.*, Sharpened visual memory representations are reflected in inferotemporal cortex. *bioRxiv* 2025.04.28.651105 (2025).
26. R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, I. Fried, Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
27. E. T. Rolls, M. J. Tovee, Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.* **73**, 713–726 (1995).
28. M. A. Yassa, C. E. L. Stark, Pattern separation in the hippocampus. *Trends Neurosci.* **34**, 515–525 (2011).
29. E. T. Rolls, Pattern separation, completion, and categorisation in the hippocampus and neocortex. *Neurobiol. Learn. Mem.* **129**, 4–28 (2016).
30. S. M. Stark, M. A. Yassa, J. W. Lacy, C. E. L. Stark, A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment.

Neuropsychologia **51**, 2442–2449 (2013).

31. E. A. Murray, M. Mishkin, Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and hippocampus. *J. Neurosci.* **18**, 6568–6582 (1998).
32. W. A. Bainbridge, *et al.*, Memorability of photographs in subjective cognitive decline and mild cognitive impairment: Implications for cognitive assessment. *Alzheimers Dement. (Amst.)* **11**, 610–618 (2019).
33. C. R. Jack Jr, *et al.*, Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology* **55**, 484–489 (2000).
34. N. C. Rust, M. R. Cohen, Priority coding in the visual system. *Nat. Rev. Neurosci.* **23**, 376–388 (2022).
35. A. Bakker, C. B. Kirwan, M. Miller, C. E. L. Stark, Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319**, 1640–1642 (2008).
36. C. Brock Kirwan, *et al.*, Pattern separation deficits following damage to the hippocampus. *Neuropsychologia* **50**, 2408–2414 (2012).
37. S. M. Stark, C. B. Kirwan, C. E. L. Stark, Mnemonic Similarity Task: A Tool for Assessing Hippocampal Integrity. *Trends Cogn. Sci.* **23**, 938–951 (2019).
38. M. A. Kramer, M. N. Hebart, C. I. Baker, W. A. Bainbridge, The features underlying the memorability of objects. *Sci. Adv.* **9**, eadd2981 (2023).
39. C. Revsine, W. A. Bainbridge, Memorability reflects statistical regularities of the environment. *Curr. Opin. Neurobiol.* **94**, 103095 (2025).
40. C. D. Needell, W. A. Bainbridge, Embracing new techniques in deep learning for estimating image memorability. *Comput. Brain Behav.* **5**, 168–184 (2022).
41. M. K. Benna, S. Fusi, Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).
42. A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, K. A. Norman, Complementary learning systems within the hippocampus: a neural network modelling approach to

reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017).

Supplementary Figures:

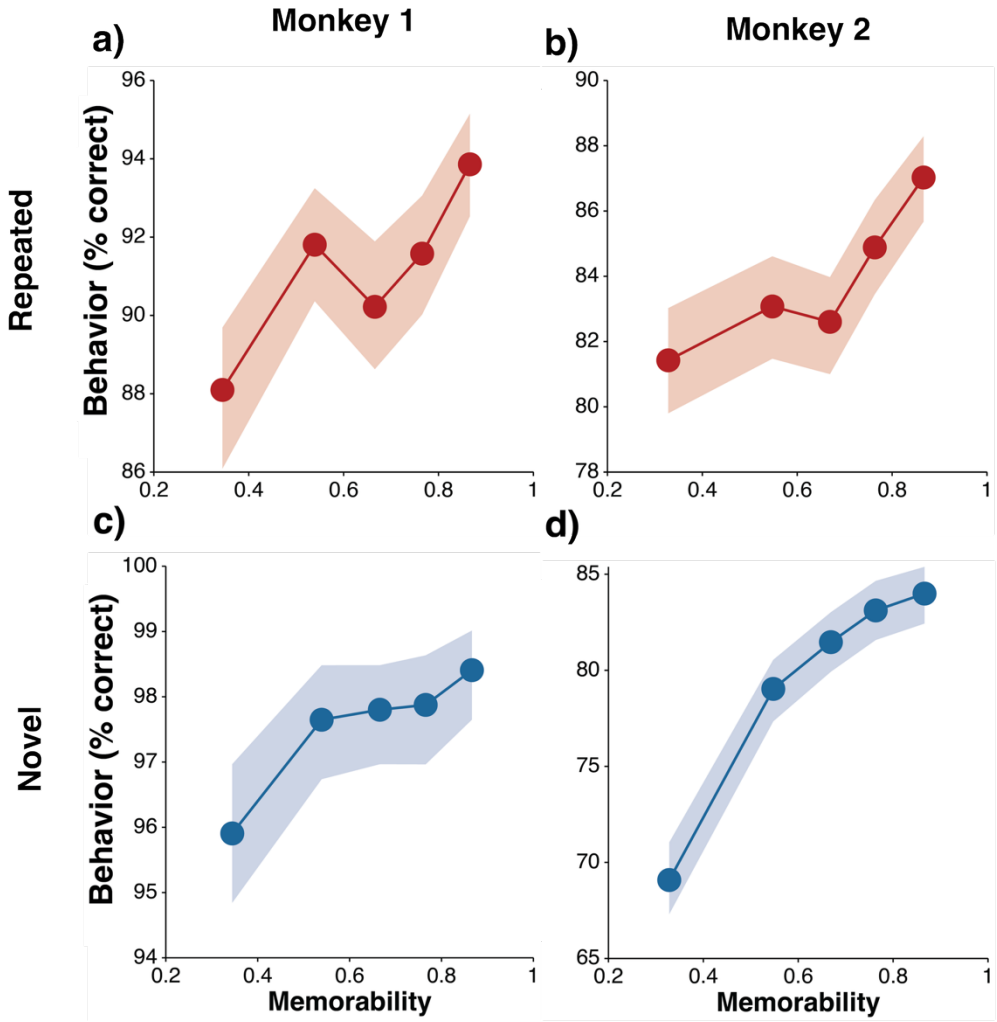

Figure S1*: Individual monkey behavior.* **(a-b)** Percent correct on repeated trials for monkey 1 (a; n=6,593 trials, 91.1% mean) and monkey 2 (b; n=11,872 trials, 83.8% mean). **(c-d)** Percent correct on novel trials for monkey 1 (c; n=6,593 trials, 97.5% mean) and monkey 2 (d; n=11,872 trials, 79.3% mean). Shaded regions represent 95% confidence intervals calculated by bootstrapping across trials with 10,000 resamples.

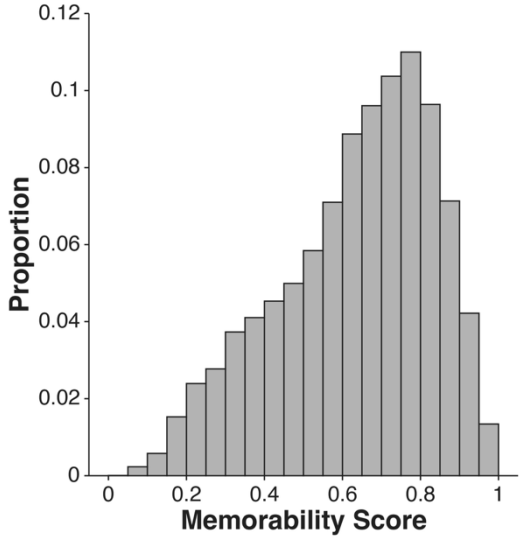

Figure S2: *Distribution of memorability scores across all images.* Shown for all images in which one or the other monkey completed both the novel and repeated trial (n=17,537 images). Images were scraped from the internet and covered a wide variety of objects and scenes. They were presented in full color. Memorability scores were acquired using the convolutional network MemNet(5), which is trained to predict human visual recognition memory on a similar visual recognition memory task.

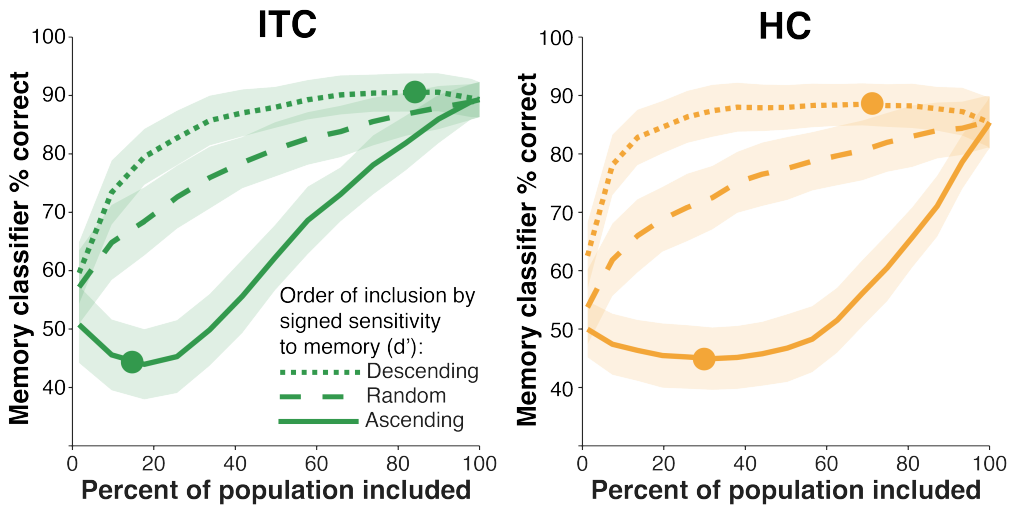

Figure S3: *Memory is reflected as repetition suppression in both ITC and HC.* To determine if the small fraction of repetition-enhanced units in Fig. 3a were contributing to memory, we performed a ranked d' analysis. Here, we trained a weighted linear classifier applied to neural data in each population (an FLD classifier, see Methods). Note that unlike the 'spike count classifier' sketched in Fig 1, this classifier is capable of negatively or positively weighting each unit to account for enhancement or suppression based on how each unit is affected by repetition). We ranked the units by their signed d' of memory (i.e., in the "descending" order of inclusion, units with the most repetition suppression received the highest ranks, with repetition enhanced units with negative d's receiving the lowest ranks) and cumulatively included portions of the population, testing for memory performance of each subpopulation. Broadly, we see a similar pattern between ITC and HC.

When including units starting with the lowest d' first (i.e., the repetition enhanced units; solid line), memory decoder performance in both brain areas did not exceed chance (50%) until well after repetition suppressed units were included in the population. Similarly, in both populations, a large amount of the classifier performance could be accounted for with a small number of the highly repetition suppressed units (dotted lines). Unit ranks were determined from their responses to the training set of images for each cross-validation split. As such, the below chance performance from the most repetition enhanced units (ascending, solid line) suggests that these units were not truly enhanced by repetition, but more likely to have appeared so due to noise (i.e. based on the training sets in each iteration, the units appeared repetition enhanced, but when tested against the images in the test set, they were actually, on average suppressed, reflected in classifier performance dipping slightly below chance level).

For this classifier, the memorability alignment of the pseudopopulation was shuffled to minimize the effects of the 'paradox of memorability' (see Methods). Error shadows show the standard deviation of classifier performance across 200 cross-validations at each population size. Analysis window for each brain area was 300ms-500ms following image presentation. Filled circles indicate where d'=0 and in the ranked analyses the population switches between repetition suppression and repetition enhancement (for ITC: 18% of the population was enhanced; for HC 31%).

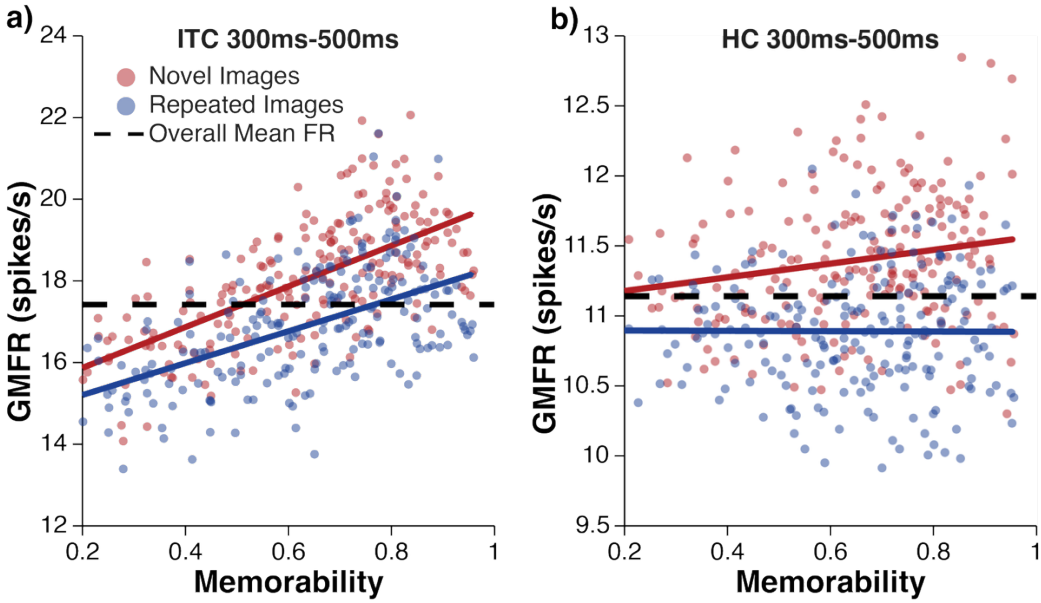

Figure S4: *Relationship between memorability and firing in 300-500ms window.* Same conventions as Fig. 3. **a)** For ITC: correlation between memorability and GMFR for novel r(206) = 0.67, p =8.24e-29 and repeated images r(206) = 0.56 p = 1.52e-18. **b)** For HC: correlation between memorability and GMFR for novel r(206) = 0.20 p = .004 and for repeated r(206) = -0.01 p = .93. Testing for attenuation of slope between brain areas, ITC β=4.983, HC β=0.489 (H0: Δβ = 0 p = 0.0034), for repeated ITC β=3.908, HC β=-0.012 (H0: Δβ = 0 p = .0047). The relationship between memorability and repetition suppression was r(206)=0.22 p = 0.001 in ITC and r(206) = 0.17, p= 0.014 in HC.

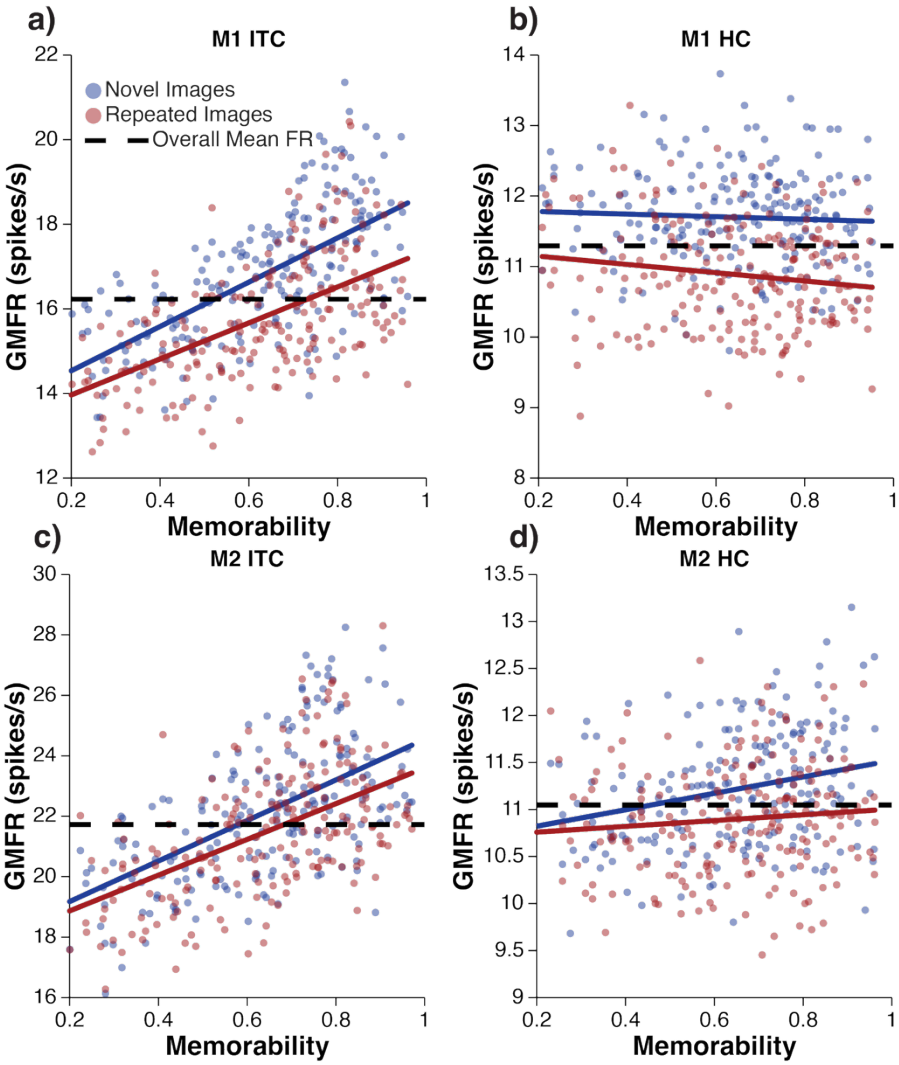

Figure S5: *Relationship between memorability and firing rate in each monkey analyzed individually.* Same convention and analysis windows as Fig. 3c-d. **a)** For monkey 1 ITC, the correlation between memorability and GMFR for novel images was r(206) = 0.67, p = 2.1e-28 and for repeated images r(206) = .58, p = 1.8e-20. **b)** For monkey 1's HC, the correlation between memorability and GMFR for novel images was r(206) = -0.05, p = 0.48 and for repeated images r(206) = -0.14, p = 0.05. **c)** For monkey 2's ITC, the correlation between memorability and GMFR for novel images was r(206) = 0.60, p = 2.1e-21, and for repeated images r(206) = 0.56, p = 1.0e-18. **d)** For monkey 2's HC, the correlation between memorability and GMFR for novel images was r(206) = 0.28, p = 3.43e-5 and for repeated images r(206) = 0.11, p = 0.13. Testing for attenuation of MB between ITC and HC in each monkey: For monkey 1 novel ITC β=5.234, HC β=-0.183 (H0: Δβ = 0 p = 0.0002), for repeated ITC β=4.254, HC β=-0.588 (H0: Δβ = 0 p = 0.0002). For monkey 2, novel ITC β=6.710, HC β=0.874 (H0: Δβ = 0 p = 0.0228), for repeated ITC β=5.916, HC β=0.308 (H0: Δβ = 0 p = 0.0219).

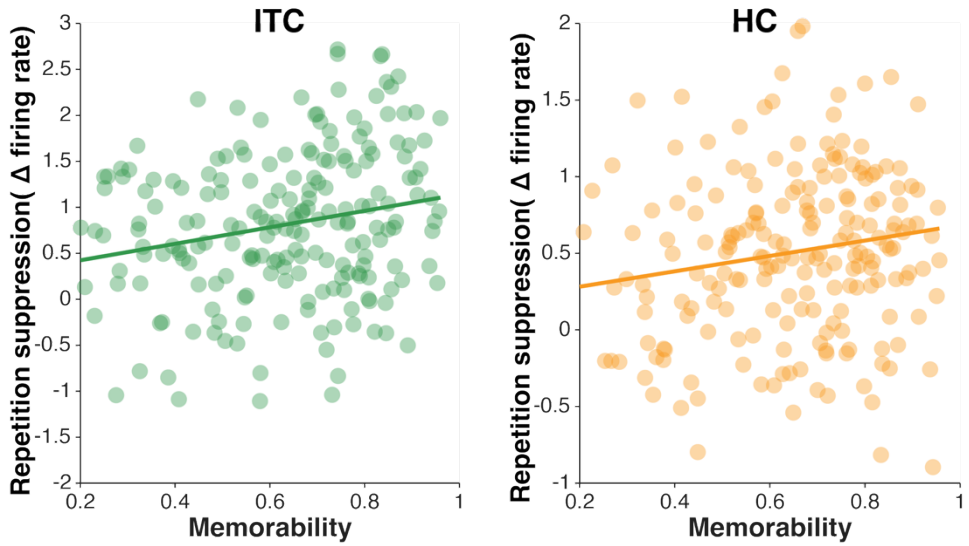

Figure S6: *Repetition suppression (RS) increases as a function of memorability in both ITC and HC*. Each dot represents RS, computed as the difference in grand mean firing rate across all units to a novel versus repeated pseudoimage with a given memorability; lines are linear regressions. For ITC, r(206) = 0.23, p = 0.001. For HC, r(206) = 0.17, p = 0.014. Analysis windows are the same as Fig. 3.

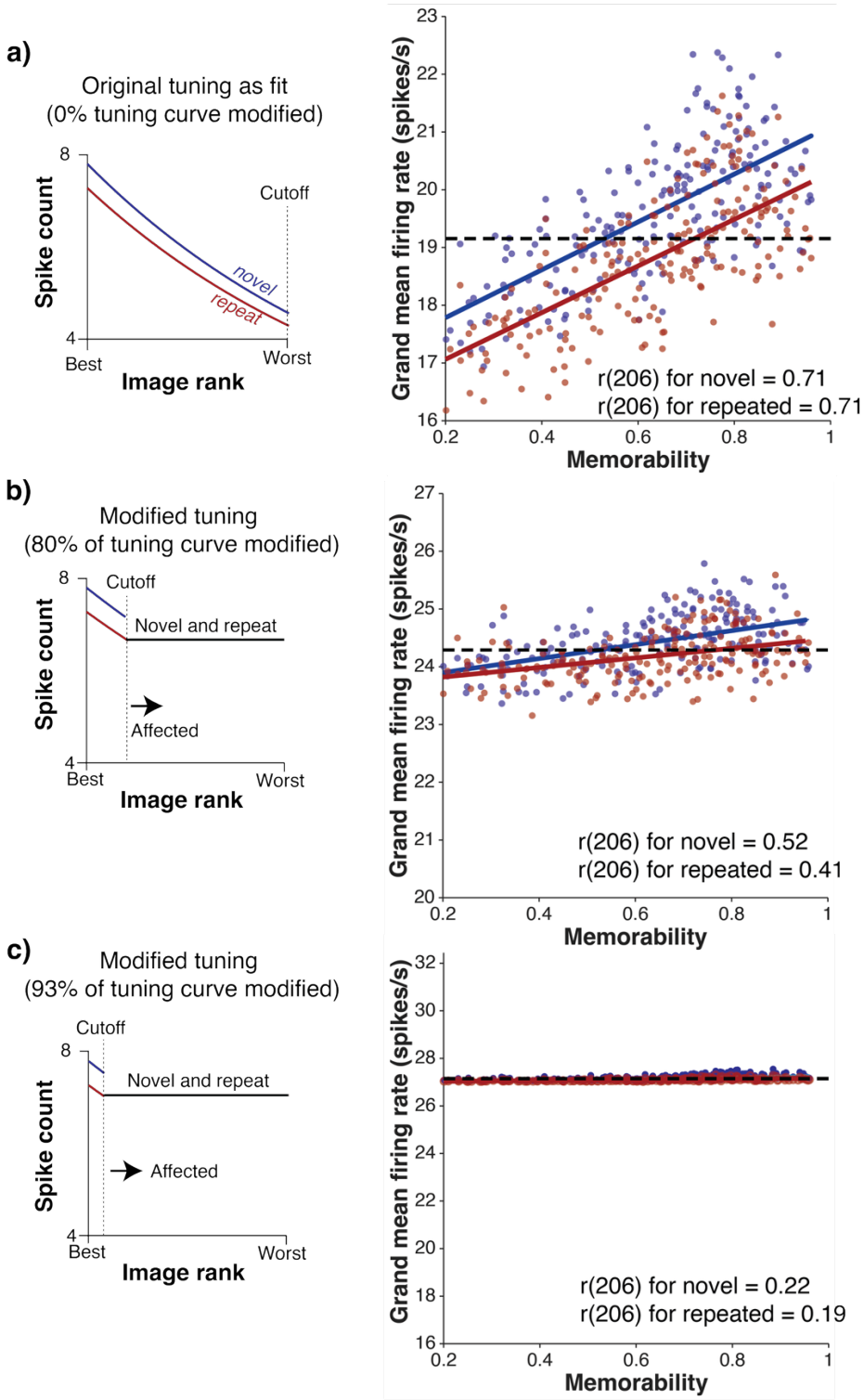

Figure S7 *Example synthetic ITC populations.* Example fit tuning curve (left) used to generate synthetic population. Equivalent analysis to Fig. 3a-b on synthesized data (right). The synthetic population was generated by fitting exponential tuning curves to each unit and using a Poisson process to simulate spikes (see Methods for details). b-c) Populations generated based on modified tuning curves at 80% tuning thresholded (b) and 93% of tuning thresholded (c).

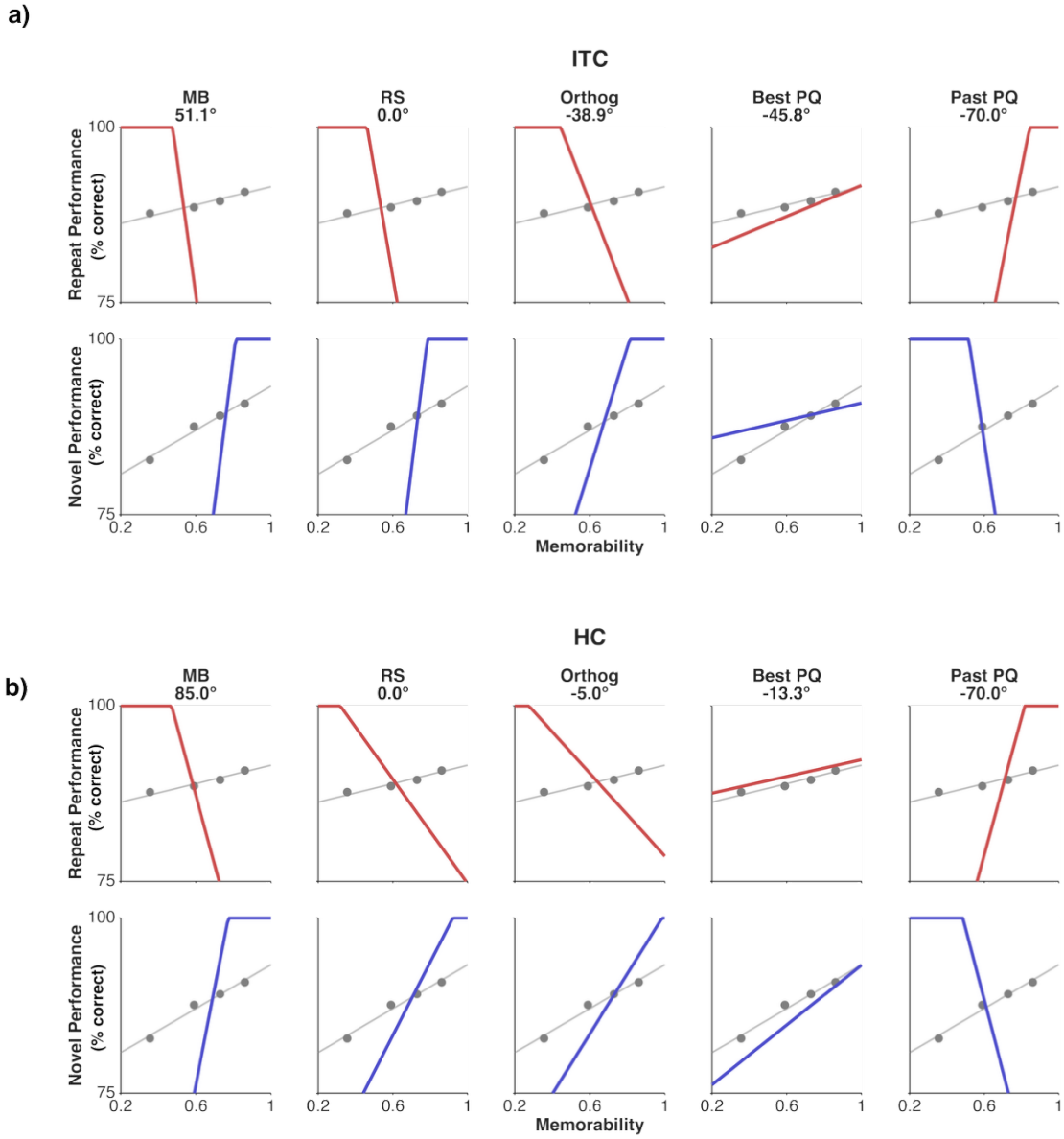

Figure S8. *Predictions of behavior at select rotations on the MB/RS plane.* **a-b)** Predictions of behavior from decoding ITC (a) and HC (b) with decoders at different rotations with respect to RS (0 degrees) on the MB/RS plane. As with Fig. 6, decoder predictions are rescaled to actual behavior with a single factor loosely based on an increase in population size (see Methods for details).

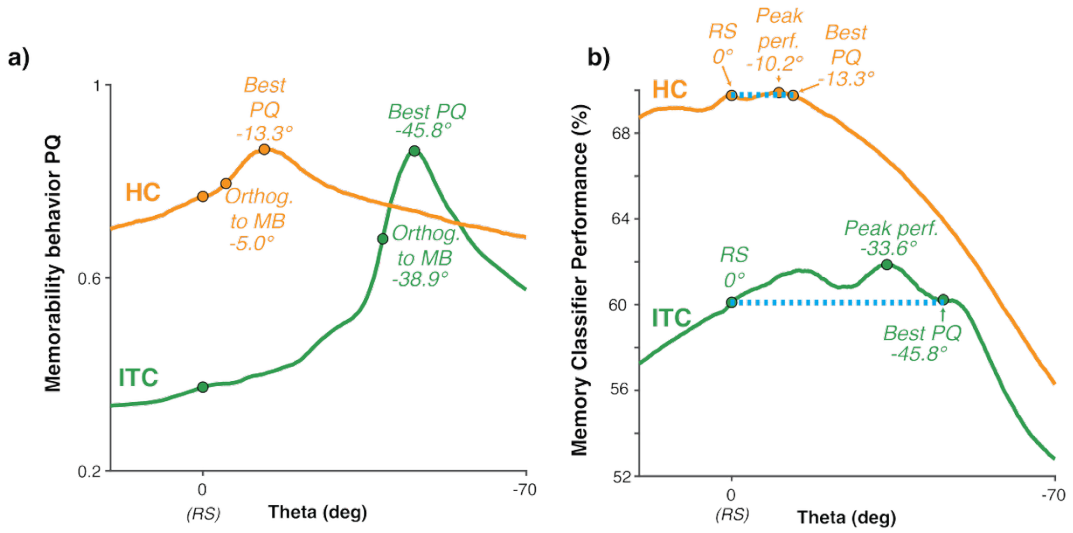

Figure S9*: Behavioral prediction quality (PQ) and memory information as a function of decoder rotation on the MB/RS plane.* **a)** PQ as a function of rotation away from the RS decoder. PQ ranges from 1 (a perfect prediction of behavior) to 0 (a prediction perpendicular to behavior; see Methods for details). **b)** Total memory information, quantified as the average of memory decoder performance across memorability bins, as a function of rotation away from the RS decoder. Here, the 'paradox of memorability' is kept intact by maintaining the alignment of image memorability in the pseudopopulation unlike Fig. 2 where this alignment is shuffled. As indicated by the dashed cyan lines, decoding memory performance at the point of best PQ does not cause any performance loss compared to decoding at RS in either ITC or HC.