Supplementary Materials for

Flexible recruitment of memory-based choice representations by human medial-frontal cortex

Juri Minxha, Ralph Adolphs, Stefano Fusi, Adam N. Mamelak, Ueli Rutishauser

correspondence to: ueli.rutishauser@cshs.org or jminxha@gmail.com

This PDF file includes:

Other Supplementary Materials for this manuscript includes the following:

Movie S1

Materials and Methods

Subjects

Subjects were 13 adult patients being evaluated for surgical treatment of drug-resistant epilepsy that provided informed consent and volunteered for this study (see **Table S1**). The institutional review boards of Cedars-Sinai Medical Center and the California Institute of Technology approved all protocols. We excluded potential subjects who did not have at least one depth electrode in medial frontal cortex.

Electrophysiology

We recorded bilaterally from the amygdala, hippocampus, dACC, and preSMA using microwires embedded in hybrid depth electrodes(*80*). From each micro-wire, we recorded the broadband 0.1-9000Hz continuous extracellular signals with a sampling rate of 32 kHz (ATLAS system, Neuralynx Inc.). Subjects from which not at least one well identified single-neuron could be recorded were excluded.

Spike sorting and single-neuron analysis

The raw signal was filtered with a zero-phase lag filter in the 300-3000Hz band and spikes were detected and sorted using a semi-automated template-matching algorithm (*81*, *82*). All PSTH diagrams were computed using a 500ms window with a step-size of 7.8ms. No smoothing was applied.

Electrode localization (relevant for Figure 1)

Electrode localization was performed based on post-operative MRI scans. These scans were registered to pre-operative MRI scans using Freesurfer's mri_robust_register (*83*) to allow accurate and subject-specific localization. To summarize electrode positions and to provide across-study comparability we in addition also aligned the pre-operative scan to the MNI152-aligned CIT168 template brain (*84*) using a concatenation of an affine transformation followed by a symmetric image normalization (SyN) diffeomorphic transform (*85*). This procedure provided the MNI coordinates that are reported here for every recording location. Note that the electrode locations shown on the Atlas Brain (Figure 1) are for illustration only. Apparent localization outside the target area or in white matter are due to usage of an Atlas brain alone.

Eye tracking (relevant for Figure S1)

Gaze position was monitored using an infrared-based eye tracker with a 500Hz-sampling rate (EyeLink 1000, SR Research)(*86*). Calibration was performed using the built-in 9-point calibration grid and was only used if validation resulted in a measurement error of <1 dva (average validation error was 0.7 dva). We used the default values for the thresholds in the Eyelink system that determine fixation and saccade onsets.

Task

Each session consisted of 8 blocks of 40 trials shown in randomized order. At the beginning of each block, an instruction screen told subjects verbally the task to be performed for the following 40 trials (categorization or recognition memory), the response modality to use (button presses or eye movements), and which visual category is the target (for categorization task only; either human faces, monkey faces, fruits, or cars;

order was pseudo-random so that each image type was selected as the target at least once) (see Figure 1). The task to solve was either "Have you seen this image before, yes or no?" or "Does this image belong to the target category, yes or no". Odd-numbered blocks (1,3,5,7) were categorization blocks; even numbered blocks were memory blocks (2,4,6,8). Button presses (yes or no) were recorded using a response box (RB-844, Cedrus Inc.). Eye movements to the left or right of the image served as responses in the eye movement modality (left=yes, right=no). The mapping between button and screen side and yes/no responses was fixed and did not change; "yes" was on the left and "no" was on the right. Subjects were reminded that left = yes, and right = no, at the beginning of each of the 8 blocks. In the first block, all images were novel (40 unique images). In all subsequent blocks, 20 new novel images were shown randomly intermixed with 20 repeated images (the "old set"). The 20 repeated images remained the same throughout a session. We used entirely non-overlapping image sets for patients that completed multiple sessions. The response modality (button presses or eye movements) was selected randomly initially and switched in the middle of each block (an instruction screen in the middle of each block showed the response modality to be used for the remainder of the block). In sessions where eye tracking was not possible due to problems with calibration (5 sessions in 3 patients; see **Table S1**), all trials used the button presses as the response modality. No trial-by-trial feedback was given. In between image presentations, subjects were instructed to look at the fixation cross in the center of the screen.

Control Task (relevant for Figure S6)

In 5 of the 13 subjects in this dataset (6/33 sessions), we ran an additional control task in order to help determine if neural responses reflected processing of stimuli, of decision variables, or of motor response plans. Unlike the standard task where the subjects could respond at any time after the stimulus onset (thus making it difficult to distinguish decision from choice), in this control task the subjects were instructed to wait until the response cue in order to register their answer, either with a button press or with a saccade. The stimulus was presented for a fixed amount of time (1s duration) and after a 0.5 – 1.5s delay period, the subjects were asked to respond to the question relevant for that block.

Mixed-effects modeling of behavior (relevant for Figures 1, S1)

For the group analysis of behavior, we used mixed-effects models of the form $y = X\beta + Zb + \varepsilon$, where y is the response, X is the fixed-effects design matrix, β is the fixed-effects coefficients, Z is the random-effects design matrix, b is the random-effects coefficients, and ε is the error vector. In all analysis, we used a random intercept model with a fixed slope. The grouping variable for the random-effects was the session ID. The reported p-values in the main text correspond to the fixed-intercept for the relevant variable. In the case of measuring the effect of number of expositions (i.e. number of times an image was seen) on the subject's accuracy during the memory trials, we used a mixed-effects logistic regression with the independent variable as an ordinal-valued whole number ranging from 1-7. The response was a logical value indicating success or failure on each memory question. Prior to running any analysis of reaction time data, we excluded outliers from the distribution using the following procedure: a sample was considered an outlier if it was outside the 99th percentile of the empirical distribution.

Reaction time matching procedure

As a control, we matched for RTs between the two tasks (categorization and memory) to exclude for potential differences due to difficulty. To achieve this, we first added noise to all reaction times (s.d. = 1ms), followed by locating pairs of trials with RTs that were equal to within a tolerance of 0.1s. Matching pairs were then removed and this procedure was repeated iteratively until no further matches could be found. Unmatched trials were excluded (resulting in reduced statistical power due to fewer trials available). We only used the resulting match if the RTs between the two groups were not significantly different. If not, the procedure above was repeated.

Selection of visually (VS) and memory-selective (MS) cells (relevant for Figure S3)

A cell was considered as a VS cell if it response co-varied significantly with visual category as assessed using a 1x4 ANOVA test at $p<0.05$. For each selected cell, the preferred image category was set to be the image category for which the mean firing rate of the cell was the greatest. All trials were used for this analysis. MS cells were selected using the following linear model:

$$fr_{cell} \sim 1 + \beta_1 \cdot category + \beta_2 \cdot new/old + \beta_3 \cdot rt$$

where *category* is a categorical (1x4) variable, *new/old* is a binary variable, and *rt* is a continuously valued variable. A cell was determined to be memory selective if the t-statistic for β_2 was significant with $p<0.05$. We excluded the first block of trials (40 images) from the analysis, in order to keep the number of new and old stimuli the same. Spikes were counted for every trial in a 1s window starting at 200ms after stimulus onset.

Selection of choice cells (relevant for Figure S5)

Choice cells were selected using a regression model applied to the firing rate in a 1s size window starting 200ms after stimulus onset. We fit the following regression model:

$$fr_{cell} \sim 1 + \beta_1 \cdot category + \beta_2 \cdot response + \beta_3 \cdot rt$$

where the response is binary (yes or no), category is a categorical variable with four levels, and RT is the reaction time. We fit this model separately to trials in the memory- and categorization condition, assuring independent selection of cells. RT was included as a nuisance regressor to control for reaction time differences between the two possible responses (see **Fig. S1A).** A cell qualified as a choice cell if the t-statistic of the β_1 term was significant at $p<0.05$ for at least one of the two task conditions. The response preference of significant cells for either yes or no was determined based on the sign of β_1 (positive = yes, negative = no). Notice that the selection process uses separate trials for memory choice cells and categorization choice cells. All trials regardless of whether the answer was correct or incorrect were used for selection. To estimate the significance of the number of selected cells, we generate a null distribution by repeating above selection process 1000 times after randomly re-shuffling the response label. We estimated this null distribution separately for choice cells in for the memory-and categorization condition and used each to estimate the significance of the number of selected cells of each type.

Chance levels for cell selection (relevant for Figure S2, S3, S5)

To estimate the chance levels for cell selection, we repeated above procedures for selection of visual category, memory selective, and choice cells after randomly scrambling the order of the labels determining the category membership being selected for (yes/no response, visual category, and new/old ground truth, respectively). We repeated this procedure 1000 times.

Single-cell decoding (relevant for Figure S5)

Single-cell decoding was done using a Poisson naïve-bayes decoder. The features used were spike counts in a 1-second window, in the interval [0.2 1.2s] relative to stimulus onset. The decoder returns the probability of a class label, given the observed spike count. The class label was binary ("yes" or "no"). The model assumes that the spike count is generated by a univariate Poisson distribution, and a separate mean rate parameter (λ) is fit to each feature-class pair. For a new observation, class membership is determined on the likelihood value. Notice that we used a single spike-count as a feature, so the naïve assumption of the decoded is no longer relevant in this case.

Population decoding (relevant for Figures 2, 3, 4, S2, S3, S4, S5, S7, S8)

Single-trial population decoding was performed on a pseudo-population assembled across sessions (*87*). We present decoding results for a variety of task variables: (1) image category, (2) new vs. old, (3) choice during memory trials, (4) choice during categorization, (5) task, and (6) response type. In order to estimate the variance of the decoding performance, on each iteration of the decoder (minimum of 250 iterations), we randomly selected 75% of the cell population that was being analyzed. For example, to measure choice decoding in MFC (as shown in **Fig. 4**), we would randomly select 575/767 cells on each iteration of the decoder. The total number of available cells depended on the variable that was being decoded. For example, for response type decoding, the number of cells in MFC was 593, since 28/33 sessions included both response types. We matched the number of trials per condition contributed by each cell that was selected to participate in the population decoding. For most task variables (image category, new/old, context, effector-type) the number of samples from each cell was equal since the task structure remained the same across all subjects and sessions. For choice decoding, however, the number of instances varied, since the subjects were free to respond with a "yes" or a "no" for each stimulus. We therefore matched the numbers for the smallest group across all subjects. Note that this matching procedure can further reduce the number of cells we included in the decoding that do not have the minimum number of trials needed per condition. For the population decoding and cross-condition generalization of familiarity presented in Figure 3, we used all image categories for which the subjects showed above chance recognition performance (see Fig. S1B). Therefore we used images of cars, fruits, and faces, excluding images of monkeys from the analysis. For all other analysis, we used all available trials.

A series of pre-processing steps were carried out before training the decoder. Firing rates for each cell were first de-trended (to account for any drift in the baseline-firing rate) and then normalized (z-scored) using the mean and standard deviation estimated from the training set. We then performed 10-fold cross validation using a linear support vector machine (SVM) decoder to estimate performance, as implemented by the '*fitcecoc'* function in MATLAB. We used an SVM with a linear kernel and a scale of 1. Decoding results are reported either as a function of time or in a fixed time window. Time-resolved decoding was done on spike counts measured in 500ms moving window, with a 16ms step size. For fixed-window decoding, we used spike counts in a 1-second window. The location of the window depended on the analysis. In **Fig. 2,** for example, we used a [-1, 0] relative to stimulus onset for task type and response type decoding. In **Fig. 3**, we used spike counts in [0.2, 1.2] relative to stimulus onset, for decoding image category and new/old.

Null models for testing significance of decoding performance

Throughout the manuscript, we compare the performance of our decoders against the 95th percentile of a null distribution. The way that this null distribution is generated depends on the variable being decoded. For variables such as image category, new vs. old, and response (i.e. yes vs. no), we used a simple shuffling procedure for the labels. For variables such as task-type, which had structure over time (memory blocks were always preceded by categorization block), small drifts in firing rate might lead to inflated decoding accuracy. Therefore, for such variables, the shuffling was done in such a way as to preserve their temporal relationship. Specifically, we offset (i.e. circular shift) the labels by a random integer value (sampled from the range ±10 – 20 trials). In the case of task decoding from the baseline firing rate, this is a very conservative measure of the null decoding performance since many trials retain their original label, thereby inflating the accuracy. This also means that the mean performance of the null distribution will not be the theoretical chancel level. In the case of task decoding, the theoretical chance level is 50% (binary classification). Using the circular shift method for scrambling labels, the mean of the null distribution was ~60%.

To compare the performance between different decoders, for example choice decoding from the HA vs. MFC population, we constructed an empirical null distribution from the pairwise differences in the performance of these two decoders trained using the shuffled labels. For example, if we get N estimates of the *null* performance (i.e. after shuffling the labels) of the HA decoder and N estimates of the *null* performance of the MFC decoder, we construct a distribution of the $N \bullet N = N^2$ pairwise differences. We can then compute the significance of the true difference in decoding performance between MFC and HA, $\boldsymbol{\Delta_{true}}$, relative to this distribution. Note that the variance of the null distribution is sensitive to the number of trials available for decoding because it changes the resolution (stepsize) by which decoding accuracy can change. For example, for 10 trials, the accuracy can take values from 0 to 1 in increments of 0.1. This results in different values for the 95th percentile of the null distribution and is the reason why in some cases a given difference in decoding accuracy is significant while it is not in others. Unless otherwise specified, all p-values for comparing decoding performance between conditions or brain areas are calculated using this approach. In the one case where the number of trials in a condition was too low to reliably estimate the null distribution (**Fig. S5I**) and for comparing the generalization index (**Fig. 3J**) we used a bootstrap test for equality of means (*88*) to compare the two conditions to assign a p-value to the true difference (repeated 1000 times to estimate the null distribution).

Multidimensional scaling (relevant for Figures 3, 4)

Multidimensional scaling (MDS) was used only for visualization. We computed MDS using Euclidean distances (MATLAB function *mdscale*) on z-scored spike count data in the [0.2 1.2s] window relative to image onset. In **Fig. 3E,** for example, MDS was computed on the activity across the entire population of HA and MFC cells, averaged across the 8 conditions plotted (new/old ⊗ task ⊗ image category, where ⊗ denotes the Cartesian product). Here the image category was restricted to images of human faces and fruits, for visualization purposes. For the cross-condition generalization performance, we use all four image categories. In **Fig. 4D** we compute MDS on the population of MFC cells, averaged across 8 conditions (response ⊗ task ⊗ effector, where ⊗ denotes the Cartesian product). In all cases, we use MDS to map the neural activity to a 3-dimensional space.

Normalized weight metric (relevant for Figures 4, 5, S5, S7, S8, S10)

The normalized weight metric is computed from the weight that a decoder assigns to a particular cell for a given classification. This weight is denoted as w_i^t, where the index $\boldsymbol{i}$ denotes the cell and the index $\boldsymbol{t}$, denotes the condition (for example, categorization or memory). The weight is converted into a normalized measure called an importance index, defined as:

$$\omega_i^t = \frac{|w_i^t|}{\sum_1^n |w_i^t|}$$

State-space analysis (relevant for Figure 4I)

We used Gaussian Process Factor Analysis (GPFA) (45) to analyze the dynamics of the average population activity for the 8 conditions arising from the combination of choice (yes, no), response modality (button press, saccade), and task (memory, categorization). The recovered latent space was 8 dimensional and all similarity measurements between trajectories were performed in this space (not in the 3D projections shown in the figure). The activity was binned using 20ms windows. All analysis was computed and visualized using the DataHigh (*89*) MATLAB toolbox. Similarity measurements between two conditions were computed and averaged over the first 500ms after stimulus onset as follows:

$$sim(\boldsymbol{r_1}(t), \boldsymbol{r_2}(t)) = \frac{\boldsymbol{r'_1}(t)}{\|\boldsymbol{r'_1}(t)\|} \cdot \frac{\boldsymbol{r'_2}(t)}{\|\boldsymbol{r'_2}(t)\|}$$

where $\boldsymbol{r_1}(t)$ and $\boldsymbol{r_2}(t)$ are the 8D state-space trajectories for condition 1 and 2 respectively.

ANOVA model (relevant for **Figure S4, S11)**

We used a single-cell ANOVA model to tease apart the contributions of choice, visual category, memory, and response time on the firing rate of a cell. The model was of the following form:

$$fr_{cell} \sim \beta_1 \bullet category + \beta_2 \bullet familiarity + \beta_3 \bullet choice + \beta_4 \bullet rt$$

fr_{cell} is the mean firing rate in a fixed window (0.2-1.2 s following stimulus onset) or a moving window of 500msto analyze the time course. The ANOVA model is fit independently at each point of time. We then compute the F-statistic for each of the regressors and report the average F-statistic across the entire population of recorded cells, fit twice to each cell for the memory and categorization task (**Fig. S4D-E, S11**). To compare the effects of task on the representation of individual variables, we compare the distribution of F-statistics estimated separately on each task, for each cell in the population. We use this approach as a measure of modulation in the strength of representation for a variable induced by task switching. Note that this comparison does not make predictions about generalizability from one task to the next since the model is fit independently.

Generalization index (relevant for **Figure 3, 4**)

To compare the within-condition decoding to the across condition generalization, we used a generalization index defined as following:

$$g = \frac{cross - chance}{within - chance}$$

Where "*within"* is the decoding performance within condition, "*cross*" is the decoding across condition, and "*chance*" is the chance decoding performance for the variable of interest (choice = 0.5, new/old = 0.5, familiarity = 0.5, image category = 0.25).

Spike-field coherence analysis (relevant for **Figure 5, S9**)

LFP preprocessing: The local-field potential recordings were highpass filtered at 1Hz. The raw recordings, sampled at 32kHz, were then downsampled to 500Hz. The downsampling procedure was done with the 'resample' command in MATLAB, which applies the appropriate antialiasing filter prior to reducing the sampling rate. For each session, we screened all MFC and HA electrodes in order to make sure that there were no artifacts that could contaminate the spike-field metrics. We excluded all electrodes with interictal discharges (IEDs) visible in the raw trace (by visual inspection). Specifically, in screening for IEDs, we looked for large stereotyped, recurring transients in the raw recording that did not correspond to cellular spiking activity. The presence of such transients would disqualify an electrode from further consideration.

Spike-field coherence (SFC): All spike-field coherence analysis was performed on snippets of the LFP extracted around the spike. We extract snippets for every cell-electrode pair. For example, to measure inter-area SFC between a single cell in preSMA and HA LFPs, we extracted n snippets each (n = number of spikes) from each of the 8 ipsilateral electrodes in hippocampus and 8 electrodes in the ipsilateral amygdala. For sessions where we used a local reference (i.e. bipolar referencing), we exclude the reference wire. For intra-area coherence (ex. HA spikes to HA field) we also exclude the wire on which the cell was recorded to avoid contamination by spike waveform. For each snippet and for each cell-electrode pair, we compute the spike-triggered spectrum using the FieldTrip 'mtmconvol' method, which computes the Fourier spectrum of the LFP around the spikes using convolution of the complete LFP traces. The spectrum was computed with a single 'hanning' taper, at 56 logarithmically spaced frequencies ranging from 2 Hz on the low end, to 125 Hz on the high end. The length of the snippet window was dynamic as a function of the frequency examined; the snipped length was set to equal to two cycles of the underlying frequency at which the spectrum was estimated (i.e. 2 Hz → 2 s snippet). We estimated the phase for each snippet and for each of the 56 frequencies from the complex-valued Fourier coefficients (i.e. phasor). We use the pair-wise phase consistency (PPC) metric as the measure of coherence. For the spike-triggered power, we compute the magnitude of the spectral coefficients returned by the Fourier transform (also computed for each cell-electrode pair) for each snippet and averaged the spectra. Unless otherwise stated, all SFC results in the paper are based on spikes recorded during the baseline period between trials (1s window preceding stimulus onset).

Group comparisons using the SFC metric: When comparing two or more groups using PPC (such as memory vs. categorization), we balanced the number of spikes between the two groups. To reduce bias involved in subsampling the larger group, we resampled the spikes from the two groups 200 times, and computed the PPC metric on each iteration. The final coherence measure for a given cell-electrode pair was an average across all 200 iterations.

To ensure that the underlying local field potential does no vary in a consistent way across conditions, we compare the distribution of average voltage values for each of the conditions in our spike-field coherence analysis. In the case of the task contrast during baseline (i.e. memory vs. categorization), we show the distribution of AUC values computed separately for each electrode in the amygdala and hippocampus (**Fig. S9D** shows that there was no significant difference). The AUC for each electrode is computed using the average baseline magnitude across memory and categorization trials. In the case of the spike-field coherence results during the stimulus onset (**Fig. 5H**), to reduce any potential confounds related to event-related potentials, we used only sessions with local referencing (bipolar). The local reference (set to one of the 8 microwires in the electrode cluster implanted in each brain area) significantly diminishes the magnitude of any event-related potentials after stimulus onset. To confirm this, we repeated the AUC analysis mentioned above, for the contrast in **Fig. 5H** (i.e. true positive (TP) vs. false negative (FN)). The results (shown in **Fig. S9E**) show that there is no significant difference between the two conditions of interest.

Supplementary Text

Pupillometry (relevant for Figure S1)

To test whether levels of engagement and arousal varied between tasks, we used pupillometry (pupil size; see **Fig. S1J** for an example session). We compared two metrics, the baseline pupil size (0-100ms after stimulus onset, **Fig. S1K**) and the slope of the pupil as it responds to the stimulus on the screen (measured from 350-600ms, **Fig. S1L**). Neither metric showed a significant modulation as a function of task (p = 0.12 and p = 0.11 for size and slope respectively, sign test), thereby indicating that levels of arousal were similar for the two tasks. This analysis is based on 25 of the 28 sessions where we measured eye movements. The remaining 3 sessions were not used because the measurement of the pupil was determined to be too noisy.

Relationship between decoding weight and single-cell response (relevant for Figures S5J, S11A)

What does the weight index measure? **In Fig. S5J**, we correlate the weight index for each cell, as assigned by a population-level choice decoder, with the d' measure computed for each cell individually. The sensitivity index is computed as follows:

$$d' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{2}(\sigma_1^2 + \sigma_2^2)}}$$

where μ_1 is the average firing rate for condition 1 (in this case this corresponds to one of the two possible choices) and μ_2 is the average firing rate for condition 2. In the denominator, σ_1 and σ_2 correspond to the variance of the firing rates for condition 1 and 2 respectively.

Upper limit of cross-task generalization performance)

The maximal possible across-task decoding performance is limited by the within-task decoding performance. For example, if training and testing within the same task is possible with 70% accuracy, the maximal possible cross-condition performance is 70%. Thus, to compare cross-condition performance between situations where within-task performance is similar the absolute values can be compared directly (i.e. **Fig. 4F, G**). In situations where within-task performance differs (i.e. **Fig. 3G**), cross-condition performance needs to be considered relative to the within-task condition performance. To operationalize this intuition, we quantified the generalization index g (see methods), which is a ratio (normalized to chance level) between the between relative to within-chance performance. A g value of 1 means the maximum possible generalization was achieved, i.e. the representation is maximally abstract given the underlying representation strength within each task.

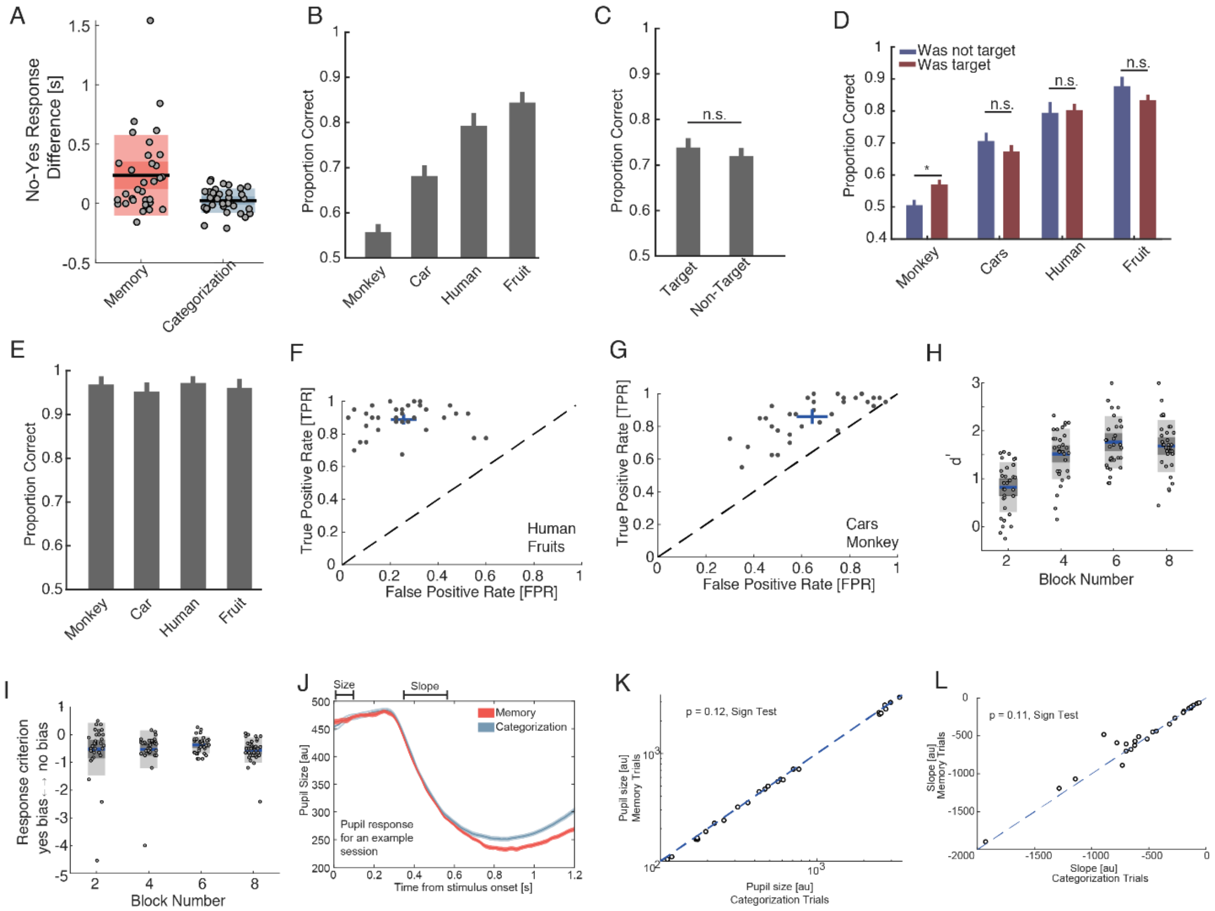

Fig. S1.

Additional behavioral and pupillometry results. (A) RT difference between "yes" and "no" responses for both tasks. RTs were significantly different for the memory task (p=3.4e-4; as expected from a declarative memory task but not for the categorization task (p=0.21) task (t-test, light shading indicates ± std, whereas darker shading indicates ± sem). **(B)** Performance on memory trials varied as a function of image category (1x4 ANOVA, p = 3.45e-19). **(C)** Making an image category the target on a categorization block did not significantly change recognition accuracy for that category on follow-up memory blocks (p = 0.97, t-test). **(D)** Same as (c) but shown separately for each category. Recognition performance increased significantly after being a target only for the monkey category (p = 0.02, t-test; uncorrected). (**E**) Performance on the categorization trials did not significantly dependent on image category (1x4 ANOVA, p = 0.74). (**F**) ROC analysis of the performance on memory trials for the two best image categories (human faces and fruits). (**G**) ROC analysis of the performance on memory trials for the two difficult image categories (cars and monkeys). While the true positive rate (TPR) is not different between these two image groups, (p = 0.24, t-test), the subjects produced significantly more false positives for the more difficult group (p=1.2e-13, t-test). **(H)** d' as a function of block number (4 memory blocks). d' increased significantly across blocks (1x4 ANOVA, p = 2.6e-11). **(I)** The response bias for each session as a function of block number. The response bias did not vary significantly across blocks (1x4 ANOVA, p = 0.6). **(J)** Example pupil response during a session. We focus on two measures: (1) the baseline size, measured shortly after image onset, and (2) the slope of the pupil response. The differences that emerge after 600ms are due to response time differences between the tasks (the image goes off the screen earlier for categorization trials). **(K)** Scatter of the pupil size, measured shortly after image onset, for all sessions with stable pupillometry (25/28 sessions with eye tracking). There was no significant difference between tasks (p = 0.12, sign test). **(L)** Scatter of the slope size across sessions with good pupillometry (25/28 sessions with eye tracking). There was no significant difference between the tasks (p = 0.11, sign test).

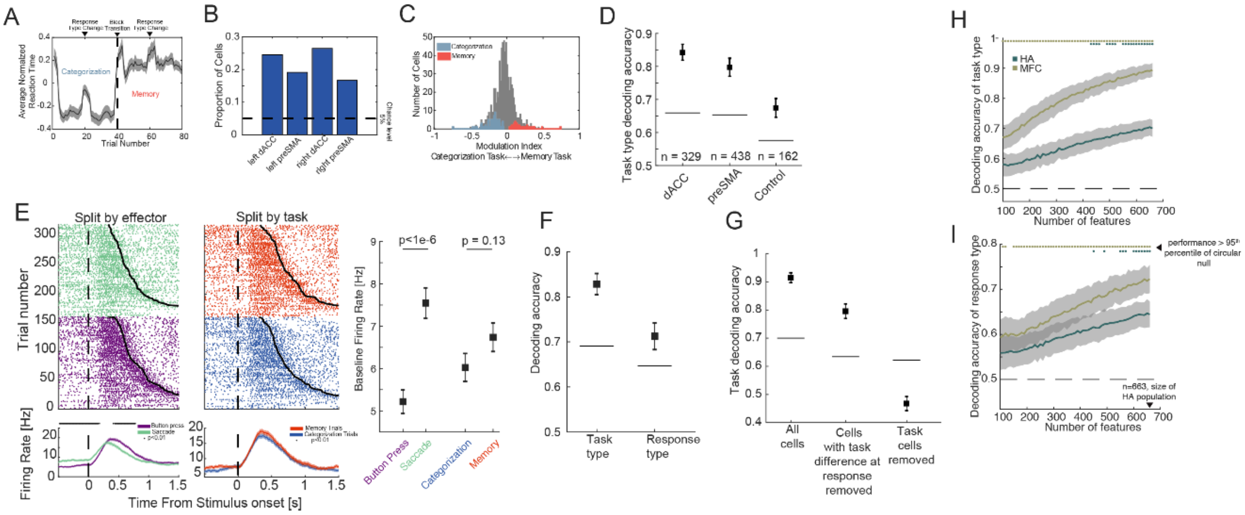

Fig. S2

Context information in the medial frontal cortex. (A) Average reaction time as a function of trial number, averaged across all subjects and block switches. Trial 41 marks the transition from a categorization task to a memory task. Halfway through each block, there is a change in response modality. Reaction time is smoothed with a 5-trial kernel. **(B)** The proportion of cells sensitive to either task type or response modality, as a function of hemisphere and area in the medial frontal cortex (L=left, R=right). The proportion of cells sensitive to context during the baseline period was significantly greater in the dACC than pre-SMA (χ^2 test of proportions, p = 0.02). **(C)** Selected cells are sorted into two groups, memory task-preferring or categorization task –preferring based on their firing rate during the baseline period. Shown here is a histogram of the modulation index, $mi = \frac{Fr_{mem}-Fr_{cat}}{Fr_{mem}+Fr_{cat}}$. **(D)** Context can be decoded from both the dACC and pre-SMA. Also shown is decoding accuracy for the control sessions (labeled as "control"), in which the response time does not differ between task types. The numbers indicate the number of cells that were included for each of the three decoders. Bars are standard deviation of decoding accuracy across all iterations of the fitting procedure (n = 250). The dotted lines mark the 95th percentile of the null distribution of decoding accuracy, computed by shuffling the labels, and performing the fit 250 times. (**E**) An example cell recorded in the pre-SMA that shows baseline modulation of firing rate with effector type (left side) but not task type (right side). **(F)** Decoding of task and effector from the MFC population after excluding choice cells (n = 207). **(G)** Task decoding is not due to post-stimulus processes from the previous trial. Shown is task decoding accuracy after removing all cells that show a significant difference in firing rate during the response period (middle bar). This ensures that any firing rate differences between tasks must arise after the response on the previous trial and during the baseline period of the next trial. Shown on the left for reference is the decoding accuracy using all MFC cells. Shown on the right is the decoding accuracy after removing the selected task cells (see Methods for selection model), reducing decoding to chance level. **(H)** Decoding of task type (left panel) and response type (right panel) in the MFC and HA with an increasing number of features. We sweep all population sizes from 100 to 663 (size of the HA population) in increments of 10. The MFC population consistently outperforms the HA population in decoding task. The dots at the top of the plot indicate if the decoding performance is better than the 95th percentile of the null distribution, where the null is estimated using a circular shifting (see **Methods**) and not just a random shuffle of the decoder labels. **(I)** Same as (H) but for decoding of response type (saccade or button press).

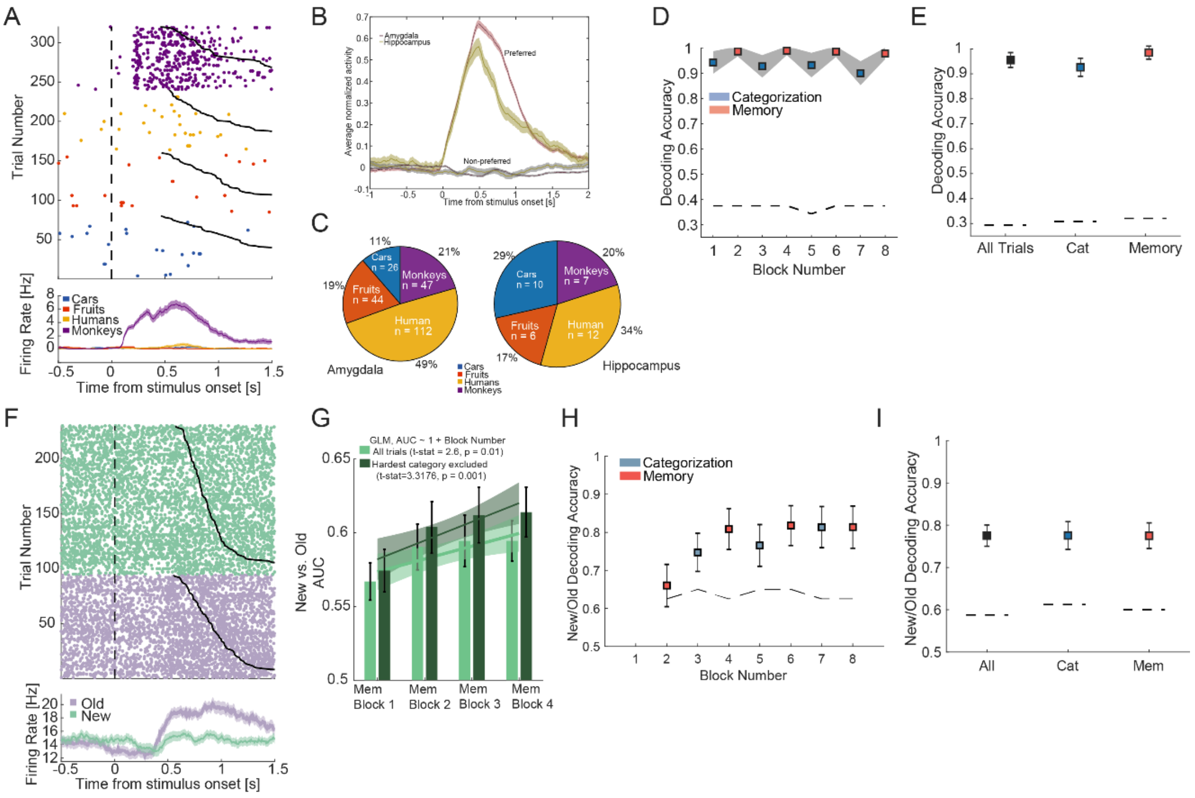

Fig. S3

Comparison of visually-and memory selective HA cells between tasks. (A) Example visually selective cell recorded in the amygdala. **(B)** Average normalized response to preferred vs. non-preferred images for all visually selective cells in the amygdala and hippocampus (n=264/663, see methods for selection criteria). **(C)** Breakdown of the preferred category of visually selective cells in the amygdala and hippocampus. As previously reported, most cells respond to faces of conspecifics. **(D)** Trial-by-trial decoding of image category over the 8 blocks within a session. The gray shading indicates the standard deviation across 200 iterations of the population decoder (see methods), using the [0.2 1.2s] time bin after stimulus onset. The decoder was trained using all trials. Shown here is the cross-validated accuracy of the decoder on each block separately, with categorization blocks in blue and memory blocks in red. The dotted black line shows the 95th percentile of the null distribution, computed by shuffling the labels. The chance level is 25%. **(E)** Same as in (d) but collapsed across task types. The dotted lines once again indicated the standard deviation across 200 iterations of the decoder, using different subsets of cells and trials (see methods of details). **(F)** Example memory selective cell in the HA. **(G)** Average AUC across all memory selective cells (n=73/663, see methods for model used to identify this cell type) for new vs. old stimuli, shown across all memory blocks. The number of new and old stimuli in each block is equal (20 of each). In light green, we show average AUC across all cells, for all the 4 image categories. New and old stimuli became more separable over the blocks (GLM, AUC ~ 1 + Block_Number, t-stat = 2.6, p = 0.01). If the category with weakest memory is removed (monkey images), the effect becomes more evident (t-stat = 3.32, p = 0.001), which is expected from a memory strength signal. **(H)** Trial-by-trial population decoding of new vs. old (using selected cells, n = 73) across all blocks in the session. The first block is excluded because it does not contain “old” stimuli. The dotted line shows the 95th percentile of the null distribution of decoding performance (chance level is 50%). **(I)** Same as (f) but collapsed across task type.

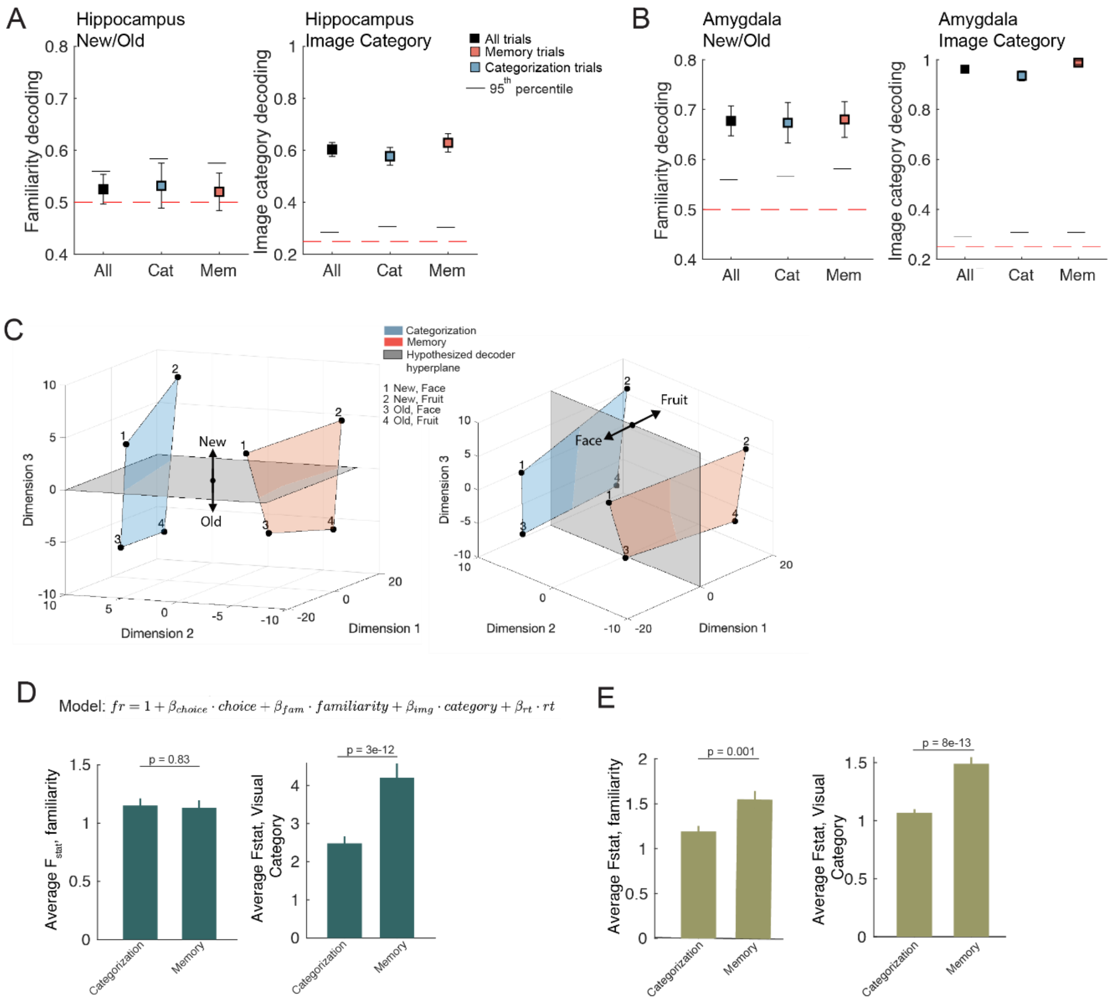

Fig. S4

Additional analysis of new/old and image decoding. (A-B) Single-trial population (all recorded cells) decoding accuracy in (A) hippocampus, (B) amygdala of new/old (left column) and image category (right column). Decoding performance is shown separately for all trials, categorization trials, and memory trials. **(C)** Rotated version of the MDS plots shown in **Fig. 3E**, with *example* decision boundaries for a new/old (left) and image category (right) decoder. The locations of the condition averages are computed from the population activity in the HA, whereas the decision boundary is schematized to show an example decoder that would generalize well across tasks. **(D)** Changes in the amount of information related to the familiarity and visual category of an image present in the population quantified using an ANOVA with regressors for familiarity, choice, image category and response time fit to each cell individually (identical to **Fig. S11C**), but here in the single time window ([0.2 1.2] seconds relative to stimulus onset). Average F-statistic for familiarity (left panel) and image category (right panel) for all cells in the HA. **(E)** Same, but for MFC**. (D, E)** F-values were significantly different for familiarity in the MFC (p = 0.001, parried t-test) but not the HA (p = 0.83, parried t-test). F-values were significantly different in both the HA and MFC (p = 3e-12 and p = 8e-13, in HA and MFC, respectively).

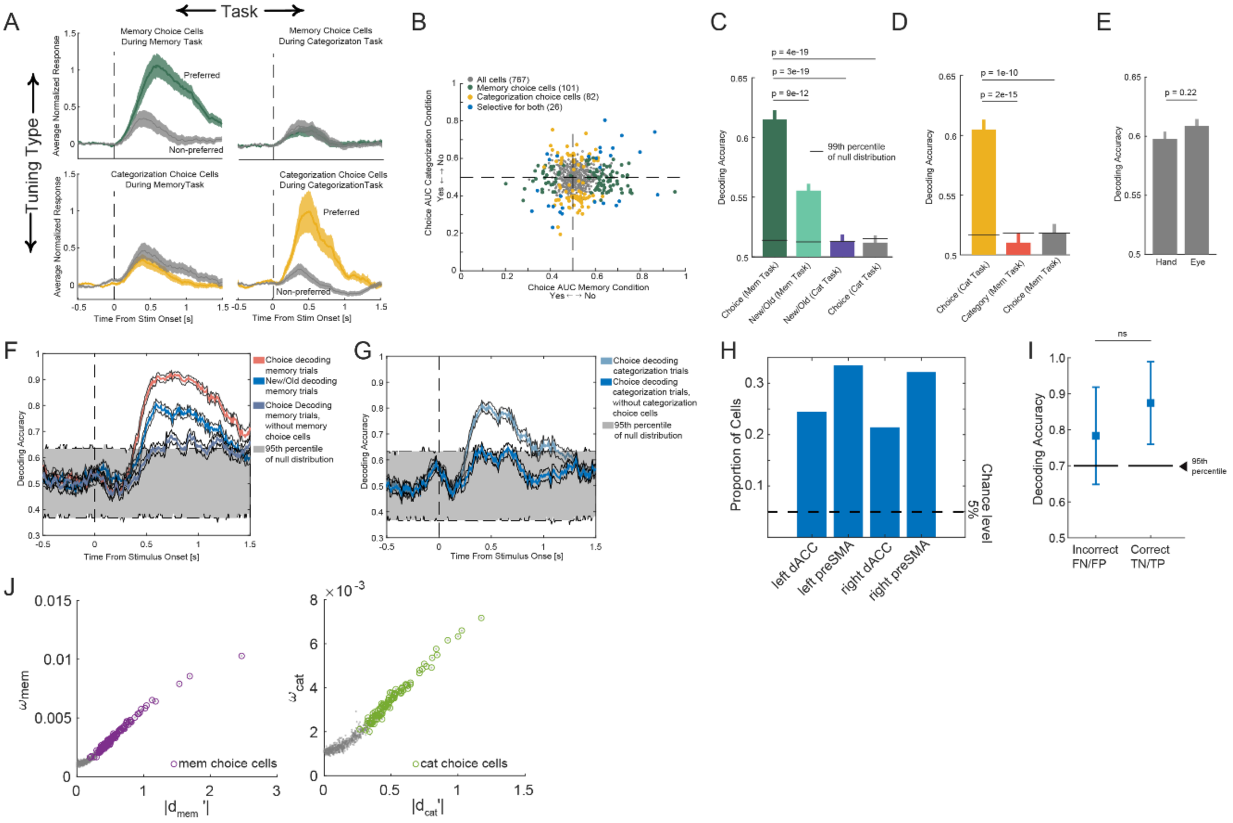

Fig. S5

Additional single-cell analysis of choice cells in MFC. (**A**) Average PSTHs for memory choice cells (green, n = 101/767) and categorization choice cells (yellow, n = 82/767), shown separately for the two tasks. Memory and categorization choice cells were selected independently using trials from the corresponding task (see methods for selection model). Omitted from this visualization are choice cells that were selected in both task conditions (n=26/767). (**B**) Population summary. AUC values were computed separately for response made during the memory and categorization condition. A negative AUC value indicates a preference for "yes" responses and a positive one indicates a preference for "no" responses. Yellow indicates categorization choice cells, green indicates memory choice cells, and purple indicates cells that signal choice in either task. **(C)** Single cell decoding across all memory choice cells (101/767). Decoding performance is shown for choice during the memory trials (green), new vs. old during the memory trials (cyan), new vs. old during the categorization trials (purple), and choice during the categorization trials (yellow). **(D)** Single cell decoding across all categorization choice cells (82/767). Decoding performance is shown for choice during the categorization trials (yellow), image category during the memory trials (orange), and choice during the memory trials (gray). (**E**) Comparison of choice decoding (collapsed across both tasks) performance between response modalities. There was no significant difference. (**F-G**) Population decoding performance as a function of time during the memory (g) and categorization (h) task. Performance was reduced significantly after choice cells were removed from the population. **(H)** Proportion of selected choice cells in medial frontal cortex, separated by area and hemisphere. The proportion of choice cells found is greater in the pre-SMA than dACC (χ^2 comparison of proportions, p = 0.004). (**I**) Trial-by-trial choice decoding at the population level was possible in both correct and incorrect trials. The decoder was trained on equal examples from the following memory trials: (1) yes-correct, (2) yes-incorrect, (3) no-correct, (4) no-incorrect. The decoder was then tested on two subsets of trials: incorrect (FN and FP) and correct (TN and TP) trials. Cells were included in the analysis only if there were at least 10 instances of each of the four trial types (n=347 cells). Decoding accuracy did not differ significantly between correct and correct trials, indicating that neurons signaled choices regardless of whether they were true or false (as expected from a choice signal; p = 0.3; Δ_{true} = 0.09, compared to the empirical null). Note that error bars in this figure are larger compared to main paper due to low number of trials used due to equating the number of trials in each of the four categories (i.e. FN, TN, FP, TP). **(J)** The weight index assigned to each cell by a population decoder (trained on choices) was strongly correlated to the **d'** (see **Methods** for calculation) estimated for each cell individually.

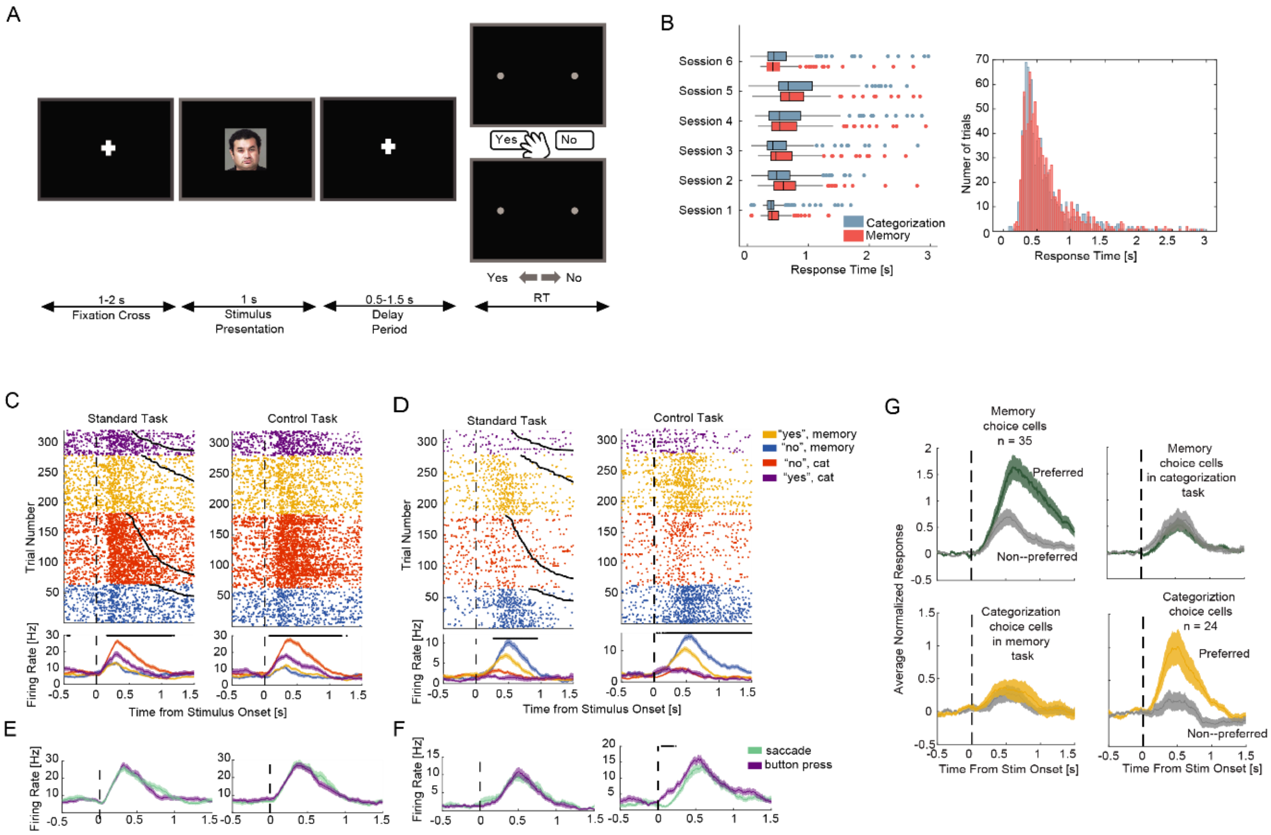

Fig. S6

Choice signals during a non-reaction time control task. **(A)** Task layout for the non-reaction time control task. Subjects are instructed to wait until the response screen comes up before registering their response with a button-press or a saccade. The stimulus length is fixed at 1 second, for both the categorization and memory trials. **(B)** The response times between the categorization and memory trials are no longer different (mean ± std, 0.67 ± 0.57s and 0.72 ± 0.77s for categorization and memory trials respectively, p = 0.1, 2-sample t-test). **(C-D)** Raster plots and PSTH of two example choice cells recorded in the dACC (C) and pre-SMA (D) during the standard task (left panel) and control task (right panel). Notice that there is no button press or saccade prior to 1.5 second during the control task**.** **(E-F)** The preferred response for the cell shown in C ("no" during categorization) and the cell shown in D ("no" during memory condition) split up by effector type, with saccade responses in green and button press in purple. **(G)** Average PSTH for preferred and non-preferred responses across all the choice cells identified in the control task. Top row shows the preferred vs. non-preferred response of memory choice cells during the memory task (left panel) and categorization task (right panel). The same is shown for categorization choice cells in the bottom row.

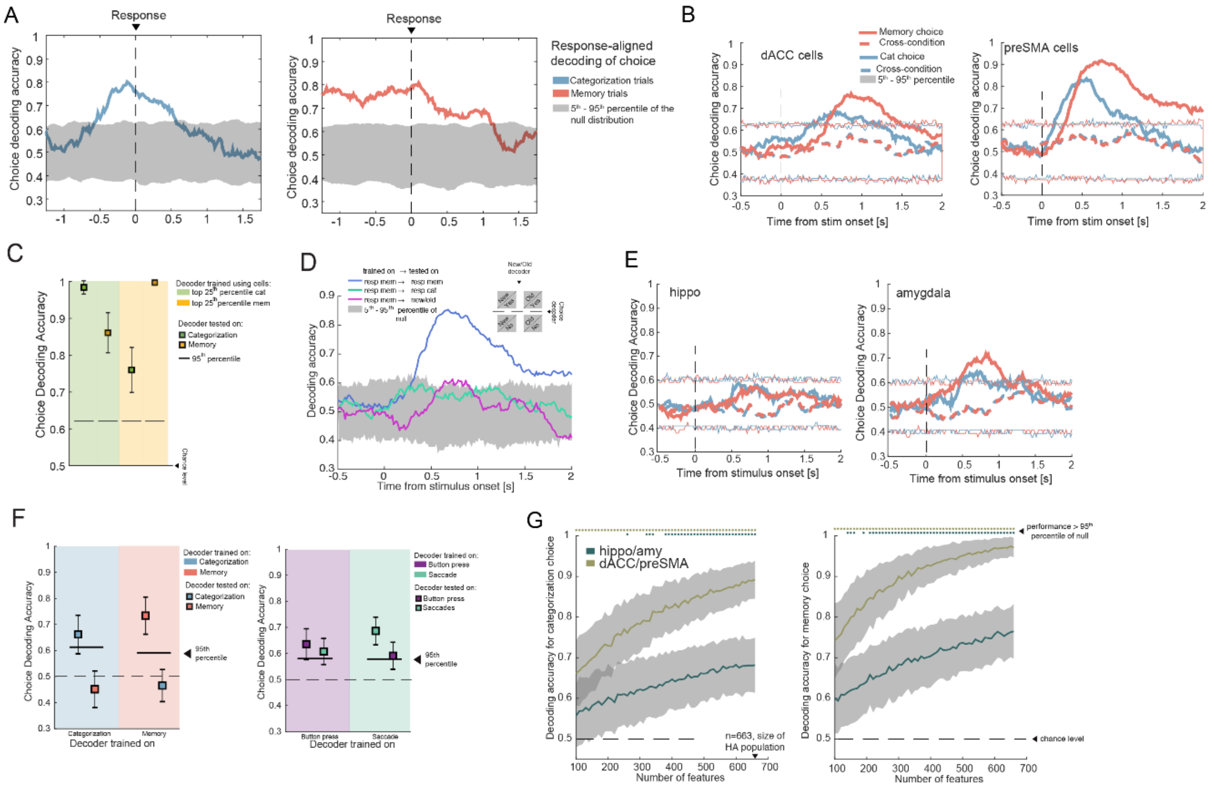

Fig. S7

Cross-task generalization of choice signals in MFC and HA. (A) Population-level decoding of choice during the categorization trials (left) and memory trials (right) using firing rates that are aligned to the response time instead of stimulus onset. Compare with **Fig. 4E**. **(B)** Cross-task generalization of choice decoding in the dACC (left) and pre-SMA (right) shown as a function of time. This is the same analysis as that in **Fig. 4E**, but shown separately for the two areas. **(C)** The cells that are in the top 25th percentile of the weight index distribution for either task (see **Fig. 4I**) can be used to train a *new* decoder that predicts choice in the other task, albeit with a significantly diminished performance. Note that this is not cross-condition generalizations since we are training a new decoder on a subset of the MFC cells. **(D)** Same as **Fig. 4C**, but as a function of time. The three traces show (1) strong decoding of pure choice during the memory task (blue), (2) this decoder cannot predict new/old (magenta), and (3) this decoder does not generalize to the choices during the categorization task (cyan, as expected). **(E)** Cross-task generalization of choice decoding in the hippocampus (left) and amygdala (right) shown as a function of time. **(F)** (Left) Summary of within and cross-task choice decoding performance in the HA in a fixed window after stimulus onset ([0.2 1.2] second interval). (Right) Within and cross-task decoding of response modality. (**G**) Decoding of choice in categorization trials (left panel) and choice in memory trials (right panel) in the MFC and HA with an increasing number of features. We sweep all population sizes from 100 to 663 (size of the HA population) in increments of 10. The MFC population consistently outperforms the HA population in decoding both variables. The dots at the top of the plot indicate if the decoding performance is better than the 95th percentile of the null distribution, where the null is estimated using a random shuffling of the labels (see **Methods**).

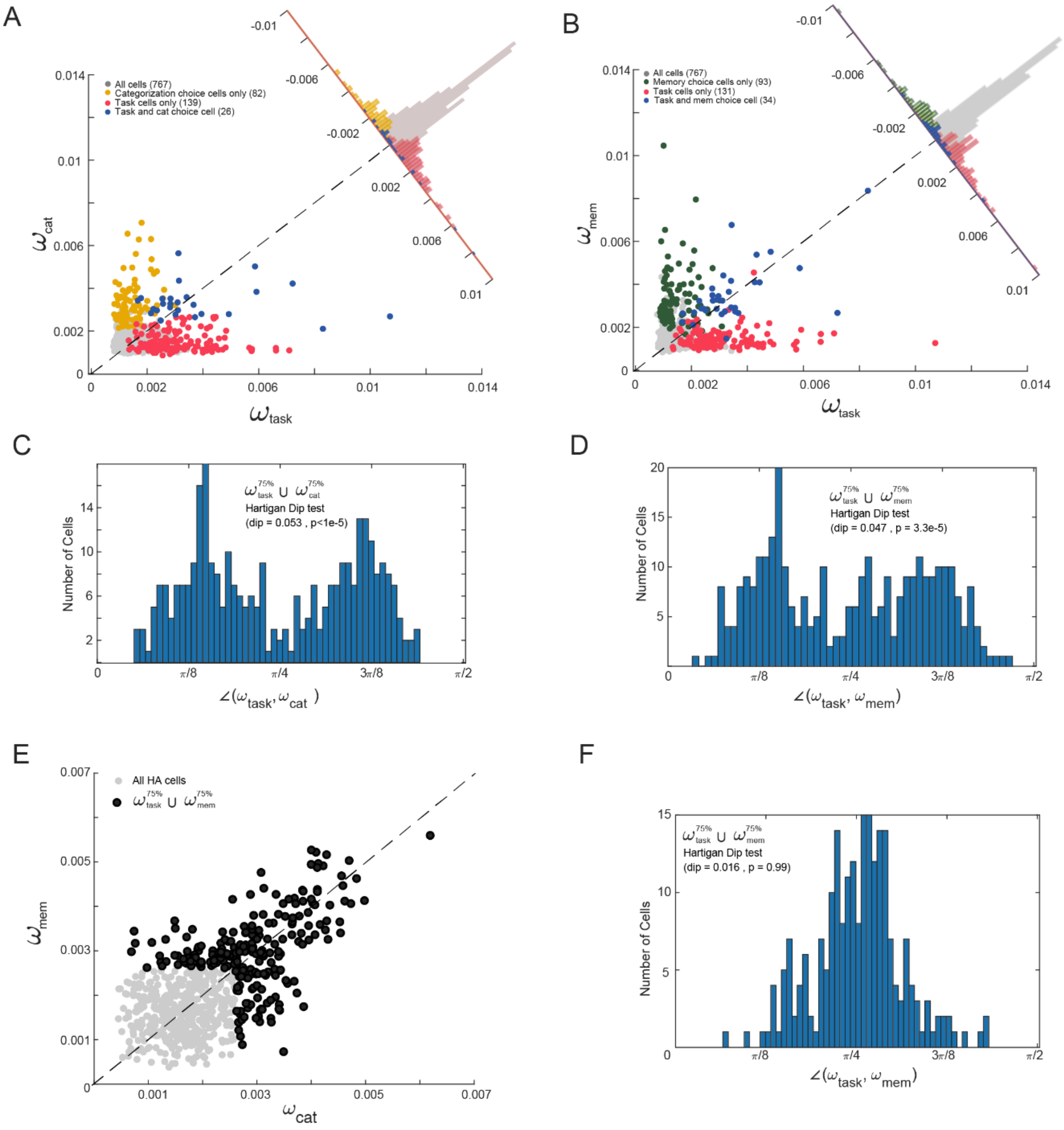

Fig. S8

Comparison of task and choice cells using assigned decoder weight. (A) Scatter plot of the weight assigned by a decoder to each cell in decoding categorization choice (y-axis), and task (x-axis). The features for the choice decoder are firing rates across the entire MFC population in the [0.2 1.2s] window after stimulus onset. The features for the task decoder are firing rates computed during the pre-stimulus baseline period, [-1 0s] with respect to image onset. As in Figure S6, the decoder weight is converted into a normalized measured (importance index). Superimposed are the populations of categorization choice cells and task cells, as identified by the choice and task selection models described in the Methods section. **(B)** Same as in (a), but shown for memory choice decoding and task decoding. Highlighted in green are the memory choice cells, and in pink are the task cells (see Methods for selection model). The cells that qualify as both are shown in blue. (**C**) Similar to Figure S6D, we look at the cells that have a high weight index for either categorization-choice or task decoding. Specifically, we take the *union* of the sets of cells whose weight index is in the top 25th percentile for either task or categorization-choice decoding. For these cells, we plot the angle created by the vector $[\omega_i^{task}, \omega_i^{cat}]$ with respect to the x-axis (i.e. the task axis). We test for bimodality with a Hartigan dip test (dip = 0.053, p<1e-5), the result of which suggests that these are largely different populations of cells. (**D**) Same as (c) but in this case we measure the overlap between memory choice cells and task cells. The histogram shows two modes, suggesting non-overlapping populations of cells (dip = 0.047, p = 3.3e-5). (**E**) Decoding of image category from the HA population is a good example of a case where the same cells are recruited for decoding in the memory and categorization task. Shown in light gray is the weight index for all HA cells, computed separately for the categorization and memory task. The dark dots indicate the union of the sets of cells that have a weight index top 25th percentile for either task. (**F**) Hartigan dip test for the weight index pair assigned to each cell in black from (e). The distribution is centered at π/4, which suggests that the same cells are recruited to decode image category during the memory and categorization tasks (dip = 0.016, p = 0.99).

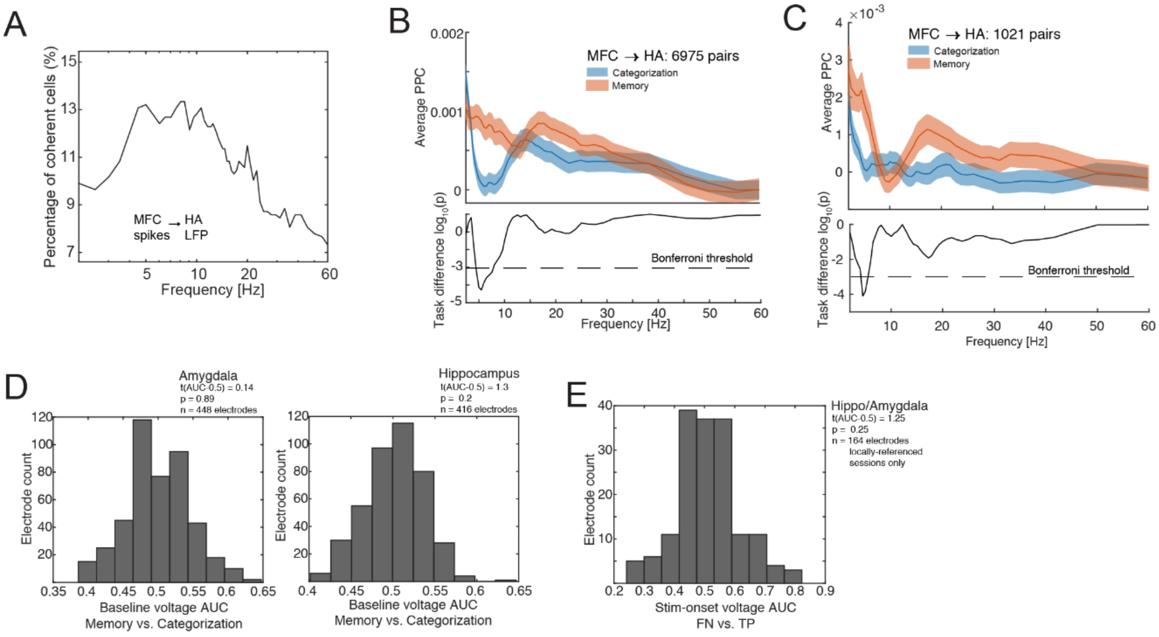

Fig. S9

Controls for inter-area spike filed coherence between MFC cells and HA local field potential. (A) Proportion of MFC cells that are coherent with hippocampal oscillations using spikes from the inter-trial period of all trials. Coherence was determined using the Rayleigh test for non-uniformity of a circular distribution. Since the comparison was done across many electrodes ($N_{channels}$ can be anywhere from 0 to 16, depending on the number of LFP recordings accepted after screening for artifacts, see **Methods**), the significance threshold was corrected appropriately for multiple comparisons using FDR (false discovery rate). **(B)** Same as **Fig. 5C**, with task cells removed (n=165/767). A cell was labeled as a task cell (see **Fig. S2** and **Methods** for selection) if it should significant modulation of firing rate as a function of task type. **(C)** Same as **Fig. 5C**, but only using HA electrodes that had spiking activity. **(D)** AUC of comparing the average magnitude of the LFP during baseline ([-1 0] seconds relative to stimulus onset) for each electrode in the amygdala (left) and hippocampus (right) between memory and categorization trials. There was no significant difference (p-values in figure). **(E)** AUC of comparing the average magnitude of the LFP following stimulus onset (0.2-1.2s following stimulus onset) for all HA electrodes used in the analysis shown in **Figure 5H** (FN vs. TP). This result shows that the ERPs do not differ between these two trial types. Note that we limited this analysis (and that in Fig. 5H) to locally referenced (bipolar) recording sessions, which is why the ERPs are not different. We found not significant difference between these two conditions (p-value in figure).

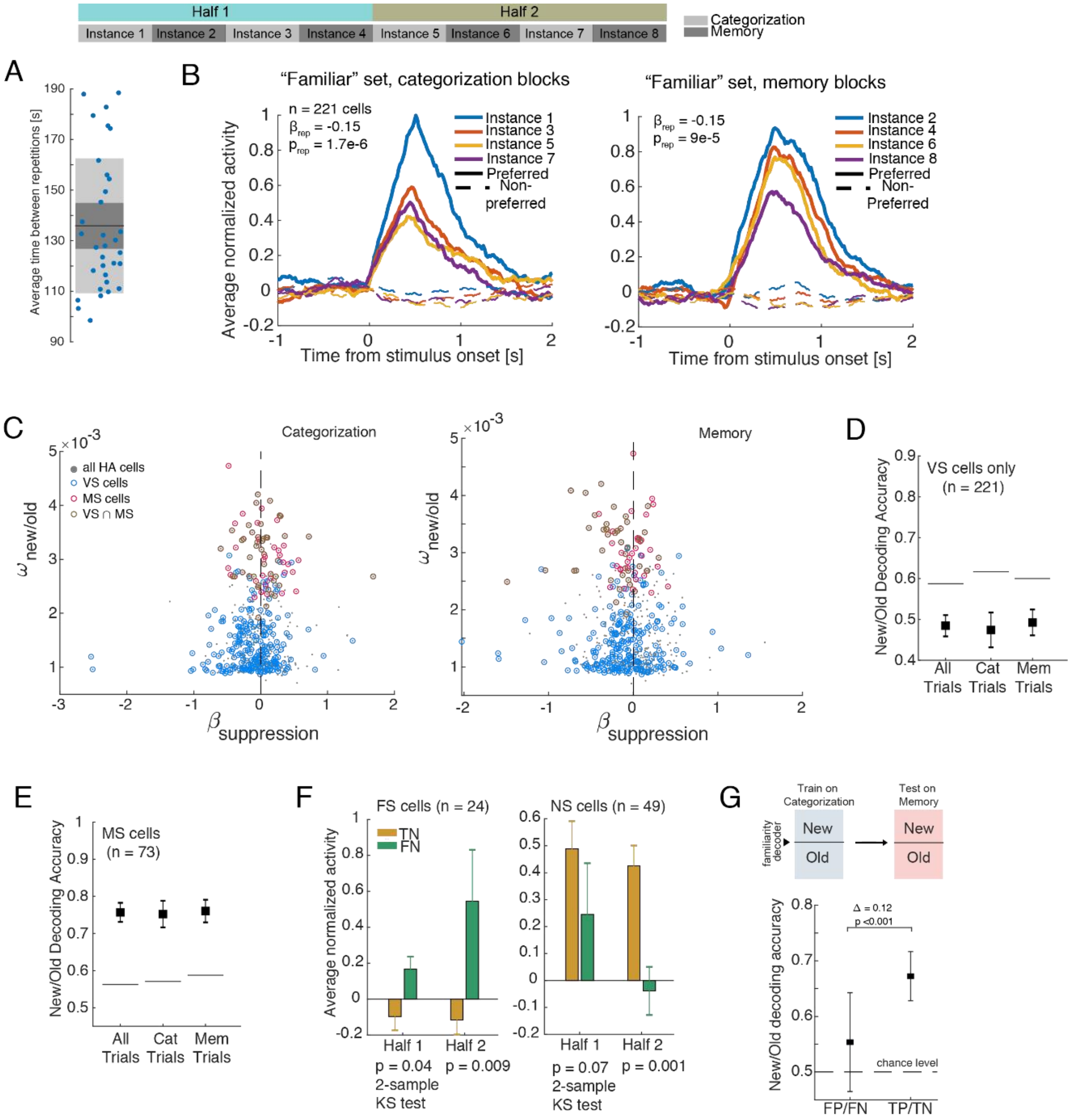

Fig. S10

Controls for memory signal in the HA. (A) The 20 images (5 images/category) that make up the set of "old" stimuli, are shown 8 times throughout the experiment. The average time between expositions of an "old" stimulus is greater than 130 seconds (approximately 40 trials). **(B)** Repetition suppression of VS cells. With each repetition, the response of the visually selective (VS) cells for their preferred stimulus is diminished. Shown is the average response for each cell's preferred category (solid lines) and non-preferred categories (dashed lines), separately for the categorization (left) and memory blocks (right). In both trial types, VS cells show strong modulation by repetition number ($p = 1.7e-6$ and $p = 9e-5$ in categorization and memory blocks respectively, p-value is estimated from a linear model with block number and reaction time as the predictor and normalized firing rate for preferred category as the response variable). (**C**) Despite the prevalence of repetition suppression in VS cells, there was no significant relationship between the degree of suppression exhibited by a ell (as measured by the β of the linear model regressing the block number on the firing rate of the cell) and the weight index (ω) assigned to the same cell by a population decoder trained on new/old labels. Result is shown separately for the categorization trials (left panel; $p_{ms} = 0.11$, $p_{ms\cap vs} = 0.62$) and memory trials (right panel; $p_{ms} = 0.6$, $p_{ms\cap vs} = 0.64$). (**D**) Decoding accuracy for stimulus familiarity (new/old) was not significantly different from chance for VS cells despite the presence of repetition suppression. (**E**) Decoding accuracy for all memory selective (MS) cells for new/old was significantly different from chance (compare to panel D). (**F**) The response of MS cells (both familiarity selective (FS) and novelty selective (NS)) differed between false negatives (FN, stimulus was "familiar" but the subject perceived it as "novel") and true negatives (TN, stimulus was "novel" and the subject perceived it as "novel"). The extent of this difference increased as the stimuli become more familiar as expected from a memory signal. Note that the response in both cases was the same, but the underlying memory signal was different. The statistics shown are 2-sample ks tests. **(G)** Comparison of new/old decoding performance between correct and incorrect trials (left) and trials with different memory strength (right). We fitted the new/old decoders to categorization trials (during which no new/old decisions are made, leaving only the memory strength signal) and tested them on subsets of memory trials. (Left) Decoding performance on correct and incorrect trials. As expected for a memory signal, decoding was weaker for incorrect vs. correct trials ($\Delta_{true} = 12\%$, $p < 0.001$, bootstrap equality of means test). Note that this plot shows that it is significantly easier to differentiate between TP vs TP trials than it is to differentiate between FN vs. FP trials (that is, [TP vs TP] > [FN vs. FP]). Note also that while weaker, decoding accuracy on incorrect trials was significantly different from chance ($p<0.005$, t-test, $t(perf-0.5) = 3.19$; $n = 21$ trials which is the smallest of incorrect trials across all sessions and the decoding procedure requires that we match correct/incorrect across all sessions; note that this is different than the typical comparison against the 95th percentile of the null decoding distribution).

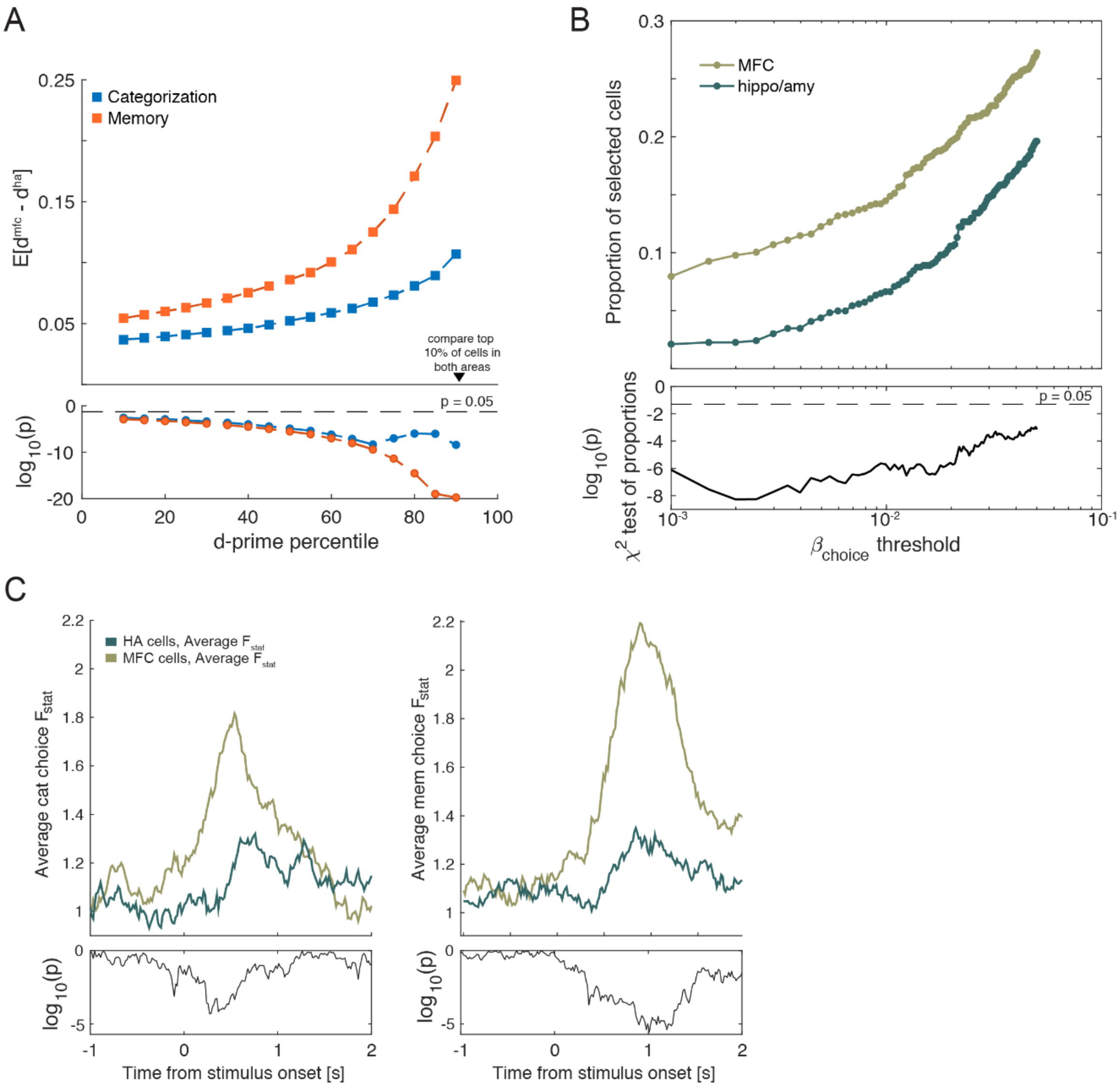

Fig. S11

Comparison of choice representation strength between HA and MFC. (A) (Top) Comparison of choice sensitivity (d’, yes vs. no) between MFC and HA across all cells in the population. Shown is the difference in mean d' values between MFC and HA, separately for choices in the memory (red) and categorization (blue) task. The comparison is shown for increasingly more selective subsets of cells within each area (from left to right). The first point on the left shows the difference between the mean d’ for all MFC and HA cells that are greater than the 10th percentile of all d’ values in the respective populations. This data shows that regardless of selection threshold and task, the strength of choice representations is significantly stronger in MFC compared to HA (bottom shows statistics; Two-sample Kolmogorov-Smirnov test of all MFC vs. all HA d’ values of all cells selected at that particular threshold, separately for both tasks). **(B)** (Top) Proportion of choice cells selected in HA and MFC as a function of selection threshold. Cells were selected using the GLM-based selection model (see Methods). The selection threshold used in the main paper is the rightmost point (threshold for the choice regressors β_{choice} p <= 0.05). The proportion of cells selected is significantly larger in MFC compared to HA for all thresholds tested (bottom; χ^2 – test of proportions). **(C)** Average single-neuron effect size across the entire population of recorded cells without selection. This analysis based on an ANOVA model with factors choice, familiarity, image category and response time. (Top) Average F - statistic for the choice dependent variable in the ANOVA model across all cells recorded from the MFC (light green) and HA (dark green) as a function of time (binsize = 500ms, stepsize =16ms; datapoints are plotted at the center of each bin). Stimulus onset is at =0. (Bottom) Significance of difference in average F values between HA and MFC.

Table S1.
List of recording sessions.

Patient ID	**Session ID**	**# HA cells**	**# MFC cells**	**Response modality used (1=button press only, 2 eye+hand)**
P41	1	1	5	2
P41	2	3	9	2
P41	3	2	1	2
P42	4	13	32	1
P42	5	20	42	1
P43	6	19	0	2
P43	7	25	1	1
P43	8	23	0	2
P44	9	11	59	1
P44	10	8	40	1
P47	11	28	8	2
P47	12	37	8	2
P47	13	33	5	2
P48	14	19	39	2
P49	15	2	3	2
P49	16	5	2	2
P51	17	20	38	2
P51	18	20	18	2
P51	19	18	21	2
P51	20	18	21	2
P51	21	11	14	2
P53	22	8	12	2
P53	23	16	21	2
P56	24	32	14	2
P56	25	15	11	2
P56	26	34	6	2
P57	27	31	23	2
P57	28	28	34	2
P57	29	28	34	2
P58	30	43	75	2
P58	31	43	75	2
P58	32	34	53	2
P61	33	15	43	2
Total		**663**	**767**	**28 with, 5 without eye tracking**

Movie S1

Dynamics of Neural activity in state space. The video shows trajectories of the average population activity for combinations of choice (yes vs. no) and task type (memory vs. categorization). The 3-dimensional space shown is a projection of an 8-dimensional latent space recovered using Gaussian process factor analysis. The gray dots denote the location in state-space of the population activity at the time of the stimulus onset. The trajectories evolve over a period of 750ms from the stimulus onset.

References

80. J. Minxha, A. N. Mamelak, U. Rutishauser, in *Extracellular recording approaches*. (Springer, 2018), pp. 267-293.

81. U. Rutishauser, E. M. Schuman, A. N. Mamelak, Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods* **154**, 204-224 (2006).

82. U. Rutishauser, M. Cerf, G. Kreiman, in *Single Neuron Studies of the Human Brain,* I. Fried, U. Rutishauser, M. Cerf, G. Kreiman, Eds. (MIT Press, Boston, 2014), pp. 59-98.

83. M. Reuter, H. D. Rosas, B. Fischl, Highly accurate inverse consistent registration: a robust approach. *Neuroimage* **53**, 1181-1196 (2010).

84. J. M. Tyszka, W. M. Pauli, In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. *Human brain mapping* **37**, 3979-3998 (2016).

85. B. Avants *et al.*, Multivariate analysis of structural and diffusion imaging in traumatic brain injury. *Academic radiology* **15**, 1360-1375 (2008).

86. S. Wang, N. Chandravadia, A. N. Mamelak, U. Rutishauser, Simultaneous Eye Tracking and Single-Neuron Recordings in Human Epilepsy Patients. *J Vis Exp*, (2019).

87. E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, T. Poggio, Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology* **100**, 1407-1419 (2008).

88. B. Efron, R. Tibshirani, *An introduction to the bootstrap*. Monographs on statistics and applied probability (Chapman & Hall, New York, 1993), pp. xvi, 436 p.

89. B. R. Cowley *et al.*, DataHigh: graphical user interface for visualizing and interacting with high-dimensional neural activity. *Journal of neural engineering* **10**, 066012 (2013).