Science
AAAS

science.sciencemag.org/content/368/6498/eaba3313/suppl/DC1

Supplementary Materials for

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

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

*Corresponding author. Email: ueli.rutishauser@csmc.edu

Published 26 June 2020, *Science* **368**, eaba3313 (2020)
DOI: 10.1126/science.aba3313

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/368/6498/eaba3313/suppl/DC1)

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. Note that **Fig. S1J** shows a difference in pupil size after 600ms. This is due to a difference in average trial-duration between tasks, with subjects taking a little longer to respond on memory trials than categorization trials (i.e. the stimulus, on average, disappears earlier on categorization trials). 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. What this shows is that the weight assigned to a cell by a decoder is largely a function of the firing rate distance between the conditions being decoded.

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 compute the generalization index, g (see **Methods**), which is a ratio (with chance-level subtracted) of across-condition performance to within-condition 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* condition).

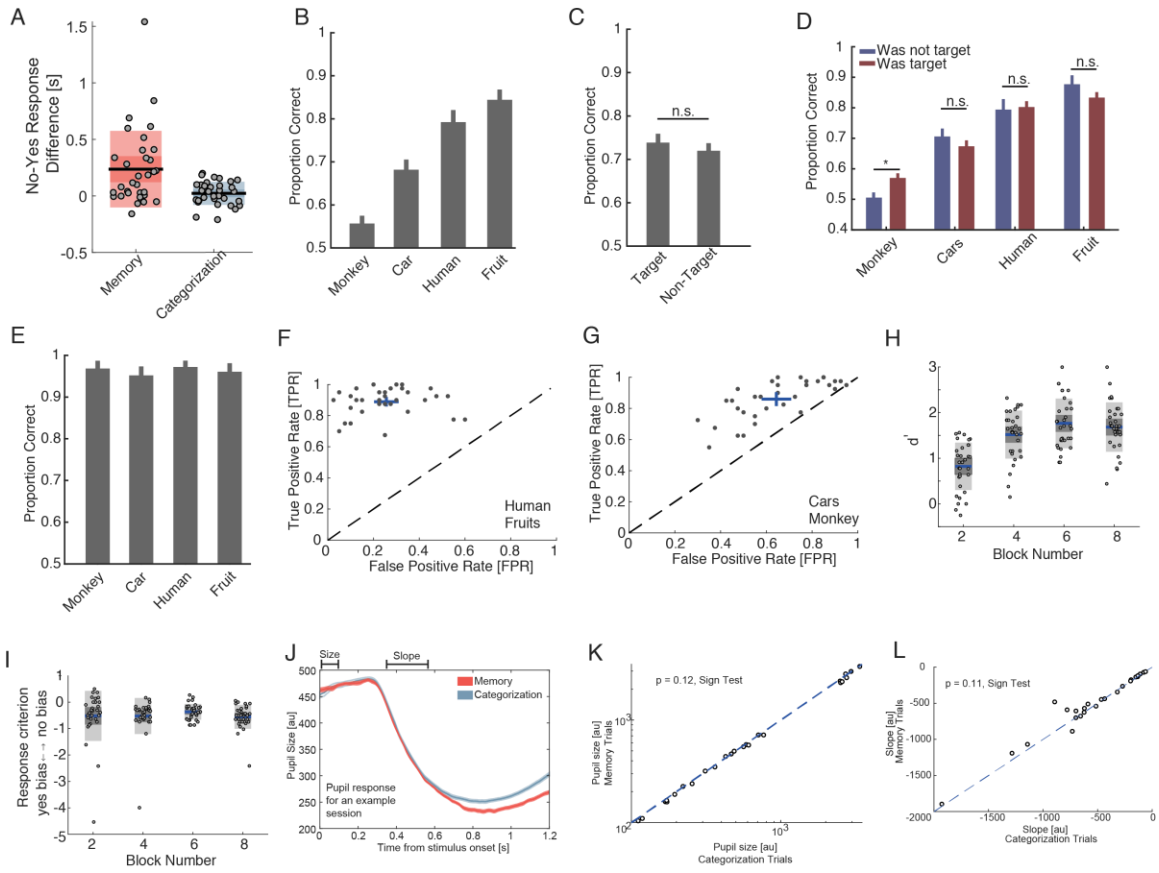

Fig. S1.

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, t-test; as expected from a declarative memory task) but not for the categorization task (p=0.21, t-test). Light shading indicates ± std, whereas darker shading indicates ± s.e.m. **(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 monkey faces (p = 0.02, t-test; uncorrected). (**E**) Performance on the categorization trials did not significantly depend 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, paired t-test), the subjects produced significantly more false positives for the more difficult group (p=1.2e-13, paired 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 differences in trial length between the tasks (subjects are faster during categorization). **(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).

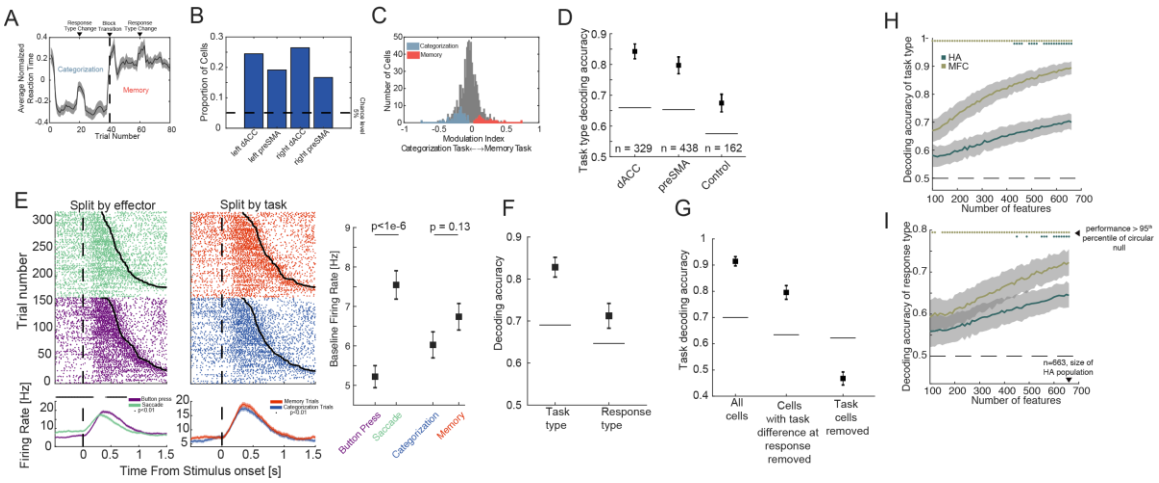

Fig. S2

Context information in the medial frontal cortex. (A) Average response 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. Response 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 task-type 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 response modality (left side) but not task type (right side). **(F)** Decoding of task and effector from the MFC population after excluding choice cells (n = 560). **(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-type. 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 shift (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).

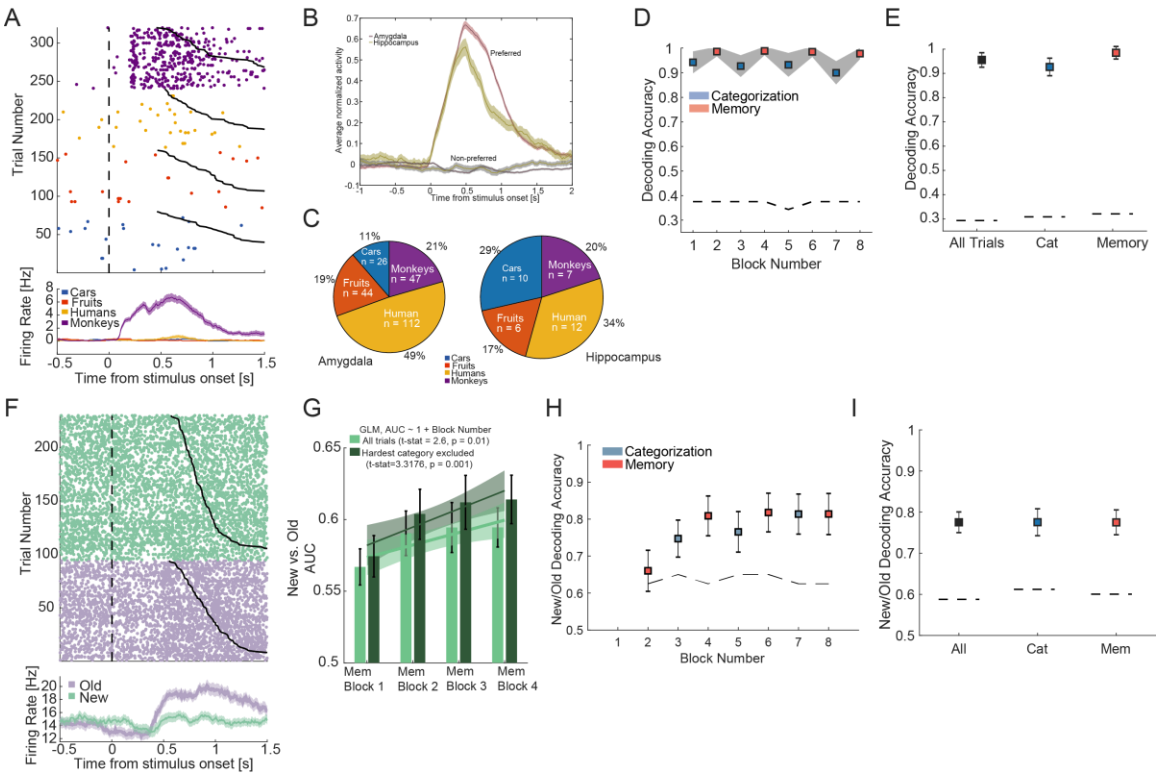

Fig. S3

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 250 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 blocks of the same task. The dotted lines once again indicate the standard deviation across 250 iterations of the decoder, using different subsets of cells and trials (see **Methods** for 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 blocks of the same task.

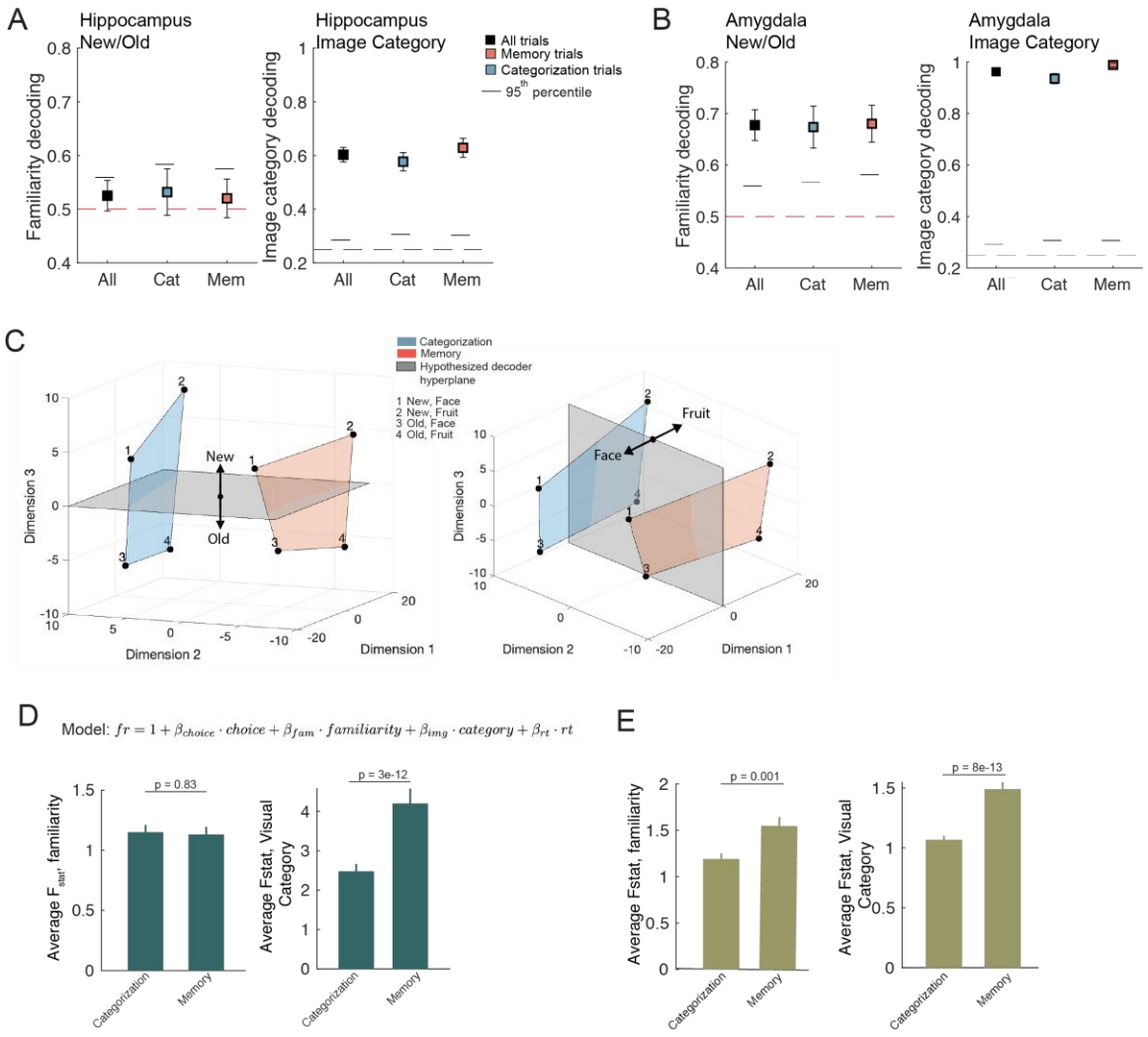

Fig. S4

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**) 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 as (D), but for the MFC neurons**. (D, E)** F-values were significantly different for familiarity in the MFC (p = 0.001, paired t-test) but not the HA (p = 0.83, paired 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).

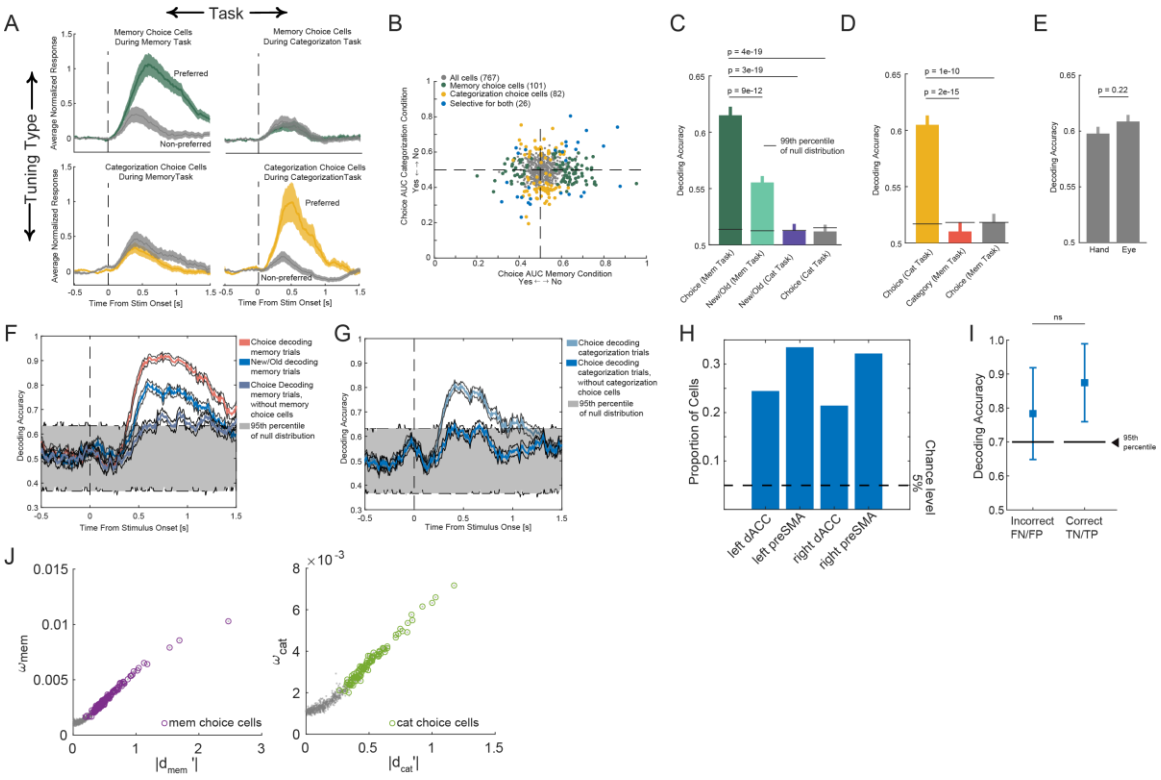

Fig. S5

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 of choice cell AUC values, computed separately for responses 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 (i.e. target vs. non-target, as defined in the preceding categorization block) 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 incorrect 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.

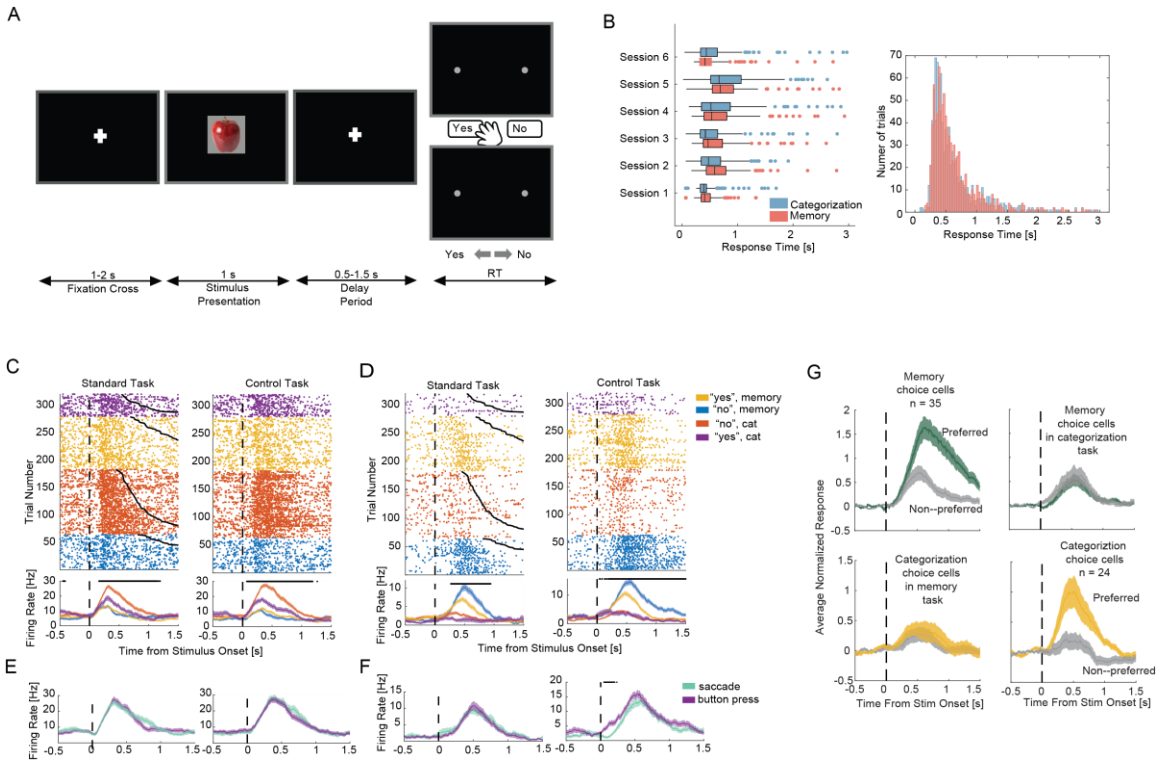

Fig. S6

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 PSTHs 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 seconds 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 PSTHs 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.

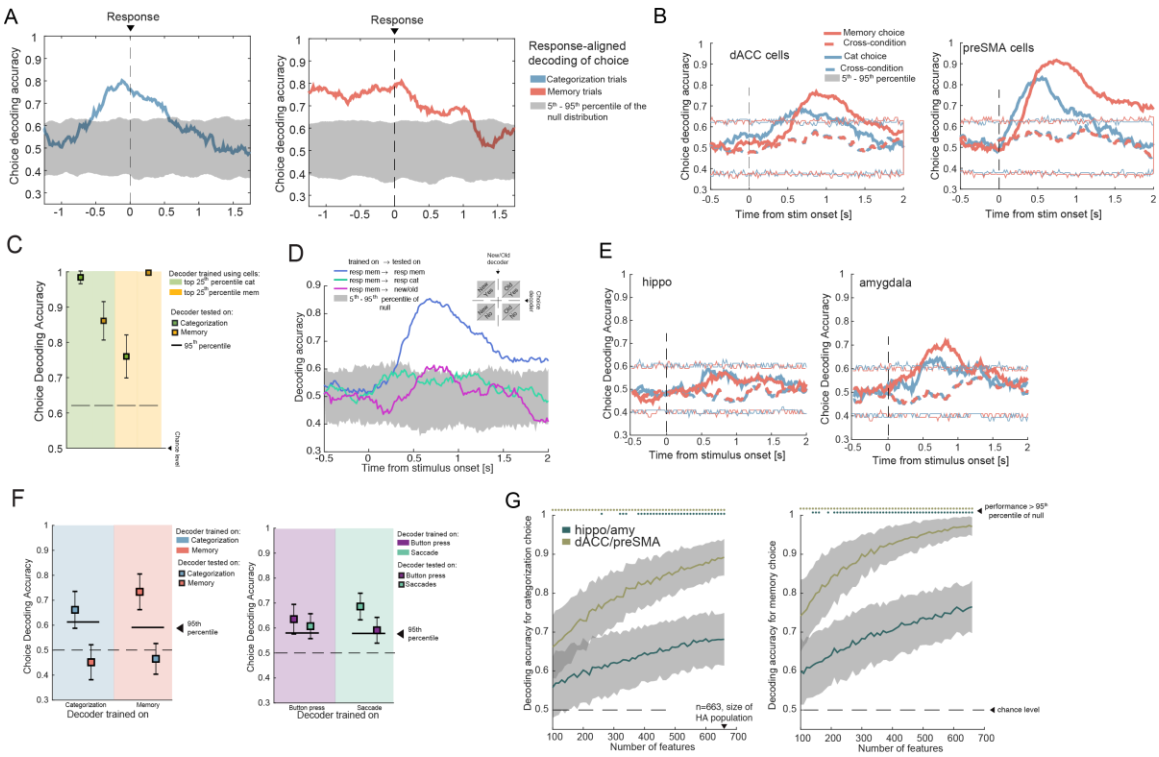

Fig. S7

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 generalization 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 1.2s] 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**).

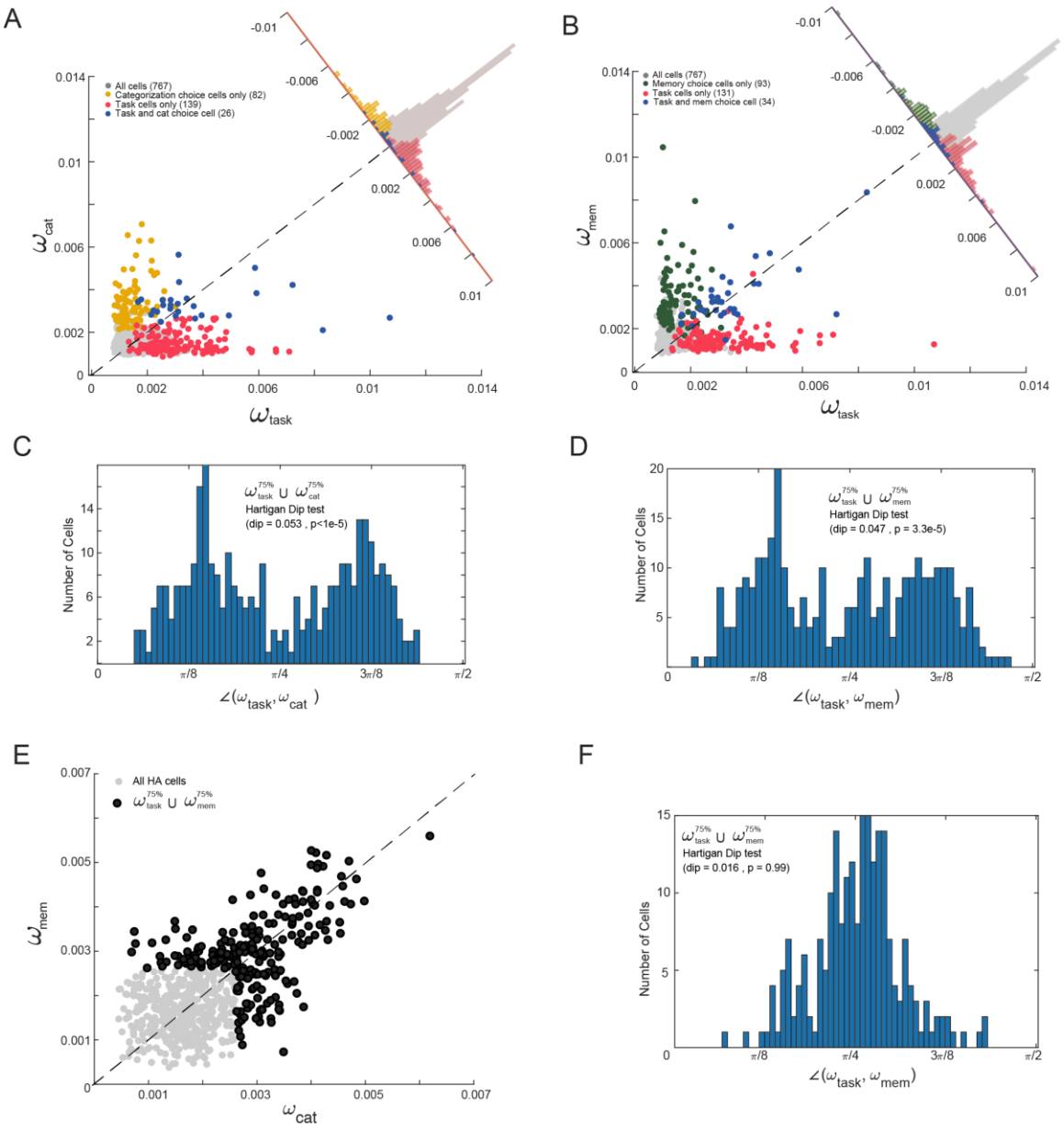

Fig. S8

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. The decoder weight is converted into a normalized measure (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**) We look at the cells that have

a high weight index for either categorization-choice or task-type decoding. Note that we are decoding choice after stimulus onset and task-type during the baseline period. We take the *union* of the sets of cells whose weight index is in the top 25th percentile for either the task-type 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-type 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 in either decoder. (**F**) Hartigan dip test for the weight index pair assigned to each cell in black from (e). The distribution is centered at $\pi/4$, which suggests that the same cells are recruited to decode image category during the memory and categorization tasks (dip = 0.016, p = 0.99).

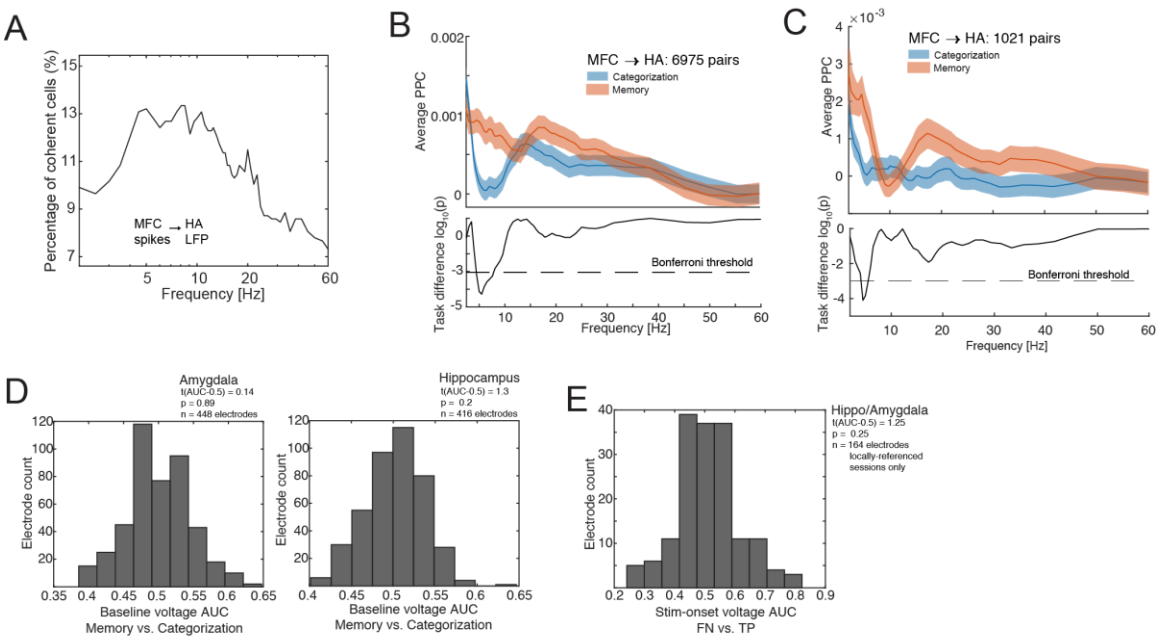

Fig. S9

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 showed 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 0s] 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] relative to stimulus onset) for all HA electrodes used in the analysis shown in **Figure 5H** (FN vs. TP). This result shows that the event-related potentials (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, in order to reduce the effect of stimulus-evoked potentials. We found no significant difference between these two conditions (p-value in figure).

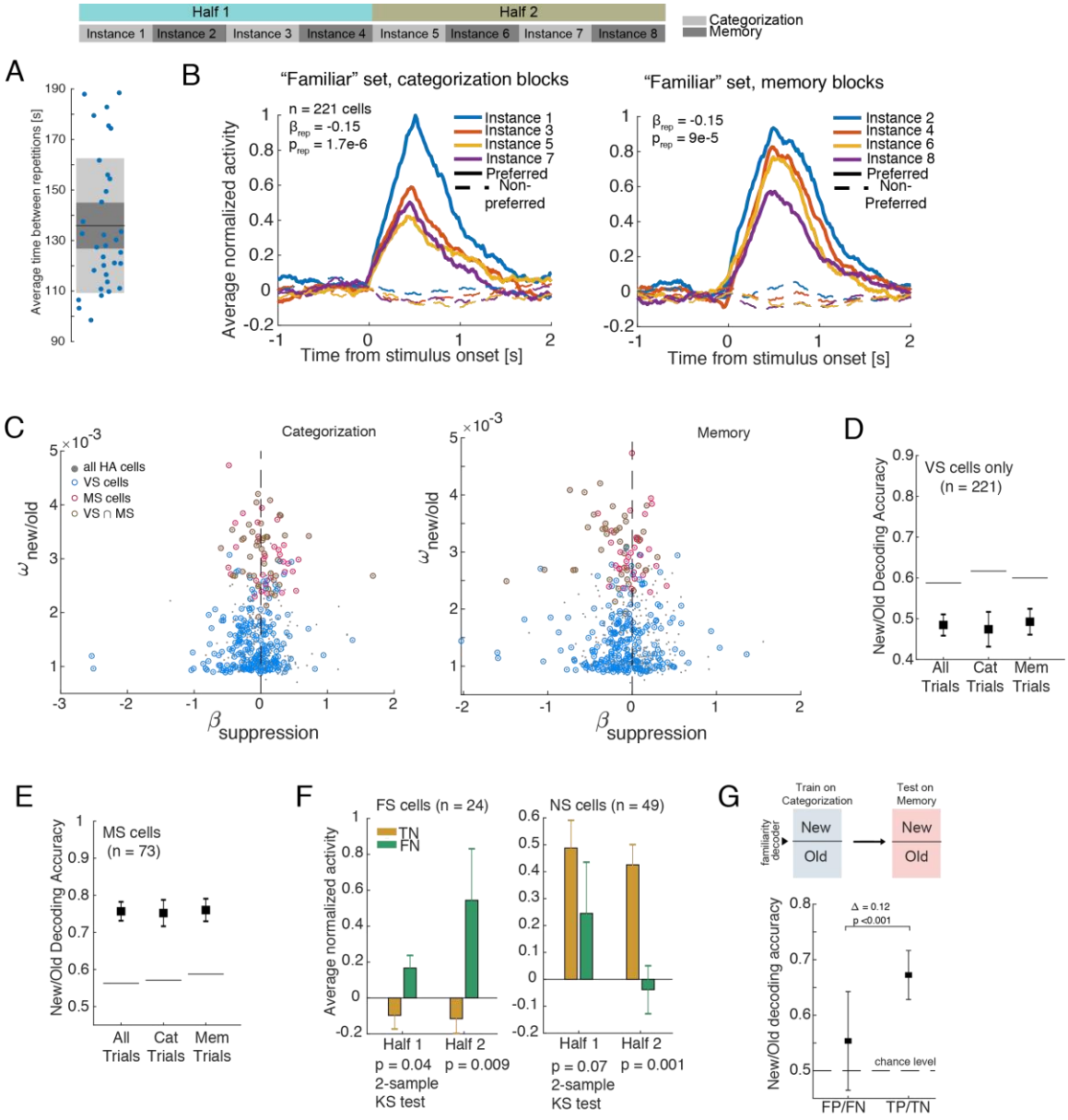

Fig. S10

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 to 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 TN vs TP trials than it is to differentiate between FN vs. FP trials (that is, [TN 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 number 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 95^{th} percentile of the null decoding distribution).

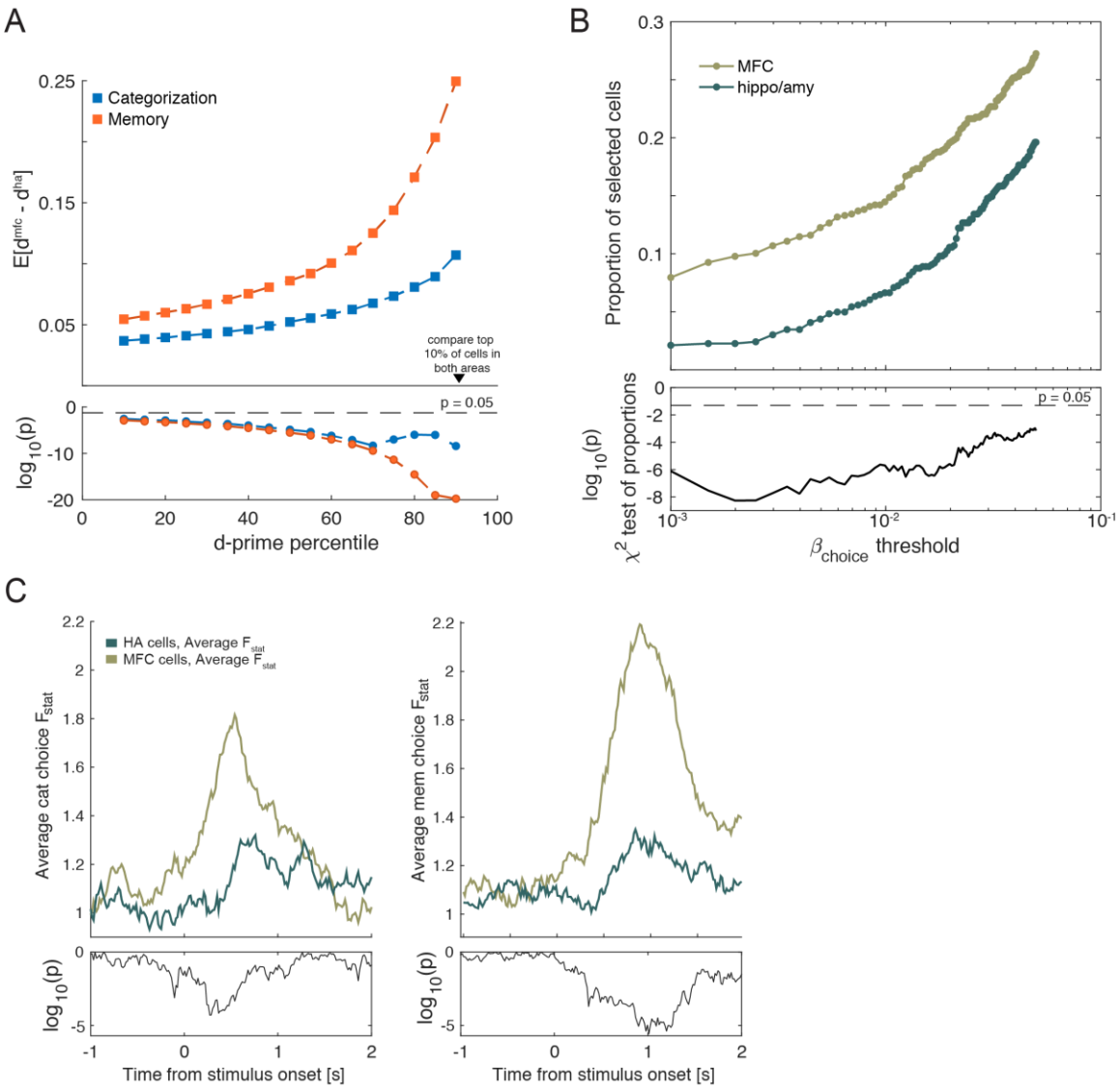

Fig. S11

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; 2-sample Kolmogorov-Smirnov test of all MFC vs. all HA d' values of all cells selected at that particular threshold, shown separately for each task). **(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 $\beta_{choice,}$ 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 is based on an ANOVA model with factors for 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 t = 0s. (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 (GPFA). 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.