nature neuroscience

Persistently active neurons in human medial frontal and medial temporal lobe support working memory

Jan Kamiński[1,5], Shannon Sullivan[1], Jeffrey M Chung[2], Ian B Ross[3], Adam N Mamelak[1] & Ueli Rutishauser[1,2,4,5]

Persistent neural activity is a putative mechanism for the maintenance of working memories. Persistent activity relies on the activity of a distributed network of areas, but the differential contribution of each area remains unclear. We recorded single neurons in the human medial frontal cortex and medial temporal lobe while subjects held up to three items in memory. We found persistently active neurons in both areas. Persistent activity of hippocampal and amygdala neurons was stimulus-specific, formed stable attractors and was predictive of memory content. Medial frontal cortex persistent activity, on the other hand, was modulated by memory load and task set but was not stimulus-specific. Trial-by-trial variability in persistent activity in both areas was related to memory strength, because it predicted the speed and accuracy by which stimuli were remembered. This work reveals, in humans, direct evidence for a distributed network of persistently active neurons supporting working memory maintenance.

The ability to store information in an active and readily available state for short periods of time is fundamental for cognition. Working memory (WM) is essential for many high-level cognitive skills, including inference, decision making, mental calculations and awareness[1]. The prevailing models of WM posit that, in the absence of external stimuli, memoranda are maintained by persistent neuronal activity[2,3]. Signatures of sustained activity have been observed in a variety of brain areas in macaques including dorsolateral prefrontal cortex (dlPFC)[4–6], parietal cortex[7], inferior temporal cortex[8,9], entorhinal cortex[10], and medial frontal cortex (MFC)[11]. It is thus thought that WM relies on different types of persistent activity provided by a distributed set of brain areas. Indeed, some brain areas in nonhuman primates exhibit stimulus-selective[4,5,7–9] persistent activity, whereas others do not[12]. However, the relationship and importance of these different kinds of persistent activity for human WM remain unclear.

In humans, a number of brain areas are thought to be essential for forming, maintaining and retrieving short-term memories[13]. Here our focus is on two of these areas: the medial temporal lobe (MTL) and the MFC. In neuroimaging studies, MFC activity is consistently related to WM[14], and it has been suggested that the MFC supports executive functions through persistent neuronal activity[15]. However, there is at present no direct neuronal evidence of such activity in the human MFC. The MTL, on the other hand, is traditionally thought to be required for the formation of new long-term memories but not for WM[16]. However, it has become clear that the MTL also plays an important role for WM[17] whenever subjects are distracted or are maintaining an amount of information close to their WM capacity. Indeed, studies of patients with damage to the MTL reveal deficits in retaining information for more than a few seconds in WM tasks under some circumstances[17]. Studies using intracranial recordings have also begun to reveal field-potential signatures of persistent activity during maintenance of WM in the MTL[18], and recordings from human lateral temporal cortex have revealed elevated activity of neurons during encoding and retrieval of short-term memories[19]. However, there is presently no direct single-neuron evidence of persistently active MTL or MFC neurons during WM maintenance in humans.

Attractors are dynamically stable patterns of neuronal activity that have been an influential framework for conceptualizing how memories are maintained by persistent activity[20,21]. In this framework, the brain has large numbers of different attractors, each of which corresponds to a different specific memory. During maintenance, at most a small subset of these attractors can be active simultaneously. Consequently, the extent to which the pattern of persistent activity during a given trial remains close to the attractor(s) should predict the quality of the memory trace as assessed by later behavior. Indeed, in macaques, the extent of drift of neural activity away from attractors formed by persisted activity is predictive of error magnitude in recalled spatial location[22]. While a major concept in theoretical work[20,21,23], there is little direct evidence for the existence and importance of attractors for WM maintenance in humans.

Here we used human single-neuron recordings performed in neurosurgical patients to investigate the properties of persistent neuronal activity and its relationship to behavior. We recorded simultaneously from two areas each in the MTL, hippocampus and amygdala, and in the MFC, dorsal anterior cingulate cortex (dACC) and presupplementary motor area (pre-SMA). As a starting point, we screened for highly selective and sparsely responsive 'concept cells'. Such cells are

[1]Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, California, USA. [2]Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, California, USA. [3]Department of Neurosurgery, Huntington Memorial Hospital, Pasadena, California, USA. [4]Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, California, USA. [5]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. Correspondence should be addressed to U.R. (ueli.rutishauser@cshs.org).

Received 1 December 2016; accepted 18 January 2017; published online 20 February 2017; corrected after print 4 May 2017; doi:10.1038/nn.4509

characterized by a highly selective and invariant response for abstract concepts, such as a particular person or building[24,25]. Here we tested whether concept cells exhibited persistent activity and whether such activity was related to WM. As in previous work[24], we used a screening task to select the images with the best response prior to starting our actual experiment. We then used only the images for which selective neurons were present in the subsequent WM task.

RESULTS

Task and behavior

Subjects (14 sessions from 13 patients) performed a WM task, (the 'Sternberg task'). In each trial, subjects were asked to memorize (encode) 1–3 images that were presented sequentially for 1 s each. After a waiting (maintenance) period of at least 2.5 s and at most 2.8 s, subjects were asked to judge whether a probe stimulus was identical to one of the 1–3 images held in memory (**Fig. 1a**). We refer to the number of images held in memory as 'load' throughout this paper. Subjects pressed either the 'yes' or 'no' button to provide their answers (button identity was reversed in the middle of the experiment). Subjects performed well: average accuracy was 89.55% ± 5.66% (**Fig. 1c**), and median reaction time (for correct trials only) was 1.09 ± 0.28 s (± s.d.). For all subsequent analysis, we used only correct trials unless indicated otherwise. Reaction time (RT) increased as a function of load (**Fig. 1d**; $F_{2,26}$ = 4.11; P = 0.025, permuted repeated-measures ANOVA). In addition, RT was significantly faster in response to probes that were held in memory (IN = 1.03 ± 0.29 s) compared to probes not held in memory (OUT = 1.2 ± 0.3 s; permuted paired t-test: t_{13} = 3.35; P = 0.005). Moreover, correct in/out decisions were made more quickly than incorrect decisions (1.09 ± 0.28 s versus 1.52 ± 0.74 s, permuted paired t-test: t_{13} = 3.81; P = 0.0005). Together, this shows that our subjects performed the task accurately and exhibited the expected patterns of RT differences[26].

Screening task

We customized the set of images to be memorized for each subject. Each subject performed a screening task 2–3 h before the experiment. During this task, 54–64 images were presented in random order six times for 1 s each. We then processed the data to identify the five images with the best image-selective responses (**Supplementary Fig. 1** and Online Methods), which were then used for the WM experiment. We will refer to these images as Images A–E throughout the manuscript (the choice of Image A, B, C, D or E was set arbitrarily but fixed for each subject).

Electrophysiology

Across all subjects, we isolated 651 (47 ± 22.1 per session) putative single units from the dACC (n = 120), pre-SMA (n = 184), amygdala (n = 195) and hippocampus (n = 152) (**Fig. 1b** shows the locations of recording sites in MNI (Montreal Neurological Institute) space). Spike sorting quality was assessed quantitatively (**Supplementary Fig. 2**). Throughout the manuscript, we use the terms neuron and cell interchangeably to refer to a putative single unit.

Classes of neuronal responses

We selected for and quantified the properties of three groups of neurons: (i) visually selective concept cells; (ii) maintenance neurons, which exhibited elevated activity in the absence of visual stimuli; and (iii) probe neurons, which responded only during retrieval. Together, these three groups of cells provide a comprehensive inventory of cellular activity during WM encoding, maintenance and retrieval.

Identification of concept cells

We identified concept cells[24] by testing whether their firing rate significantly covaried as a function of picture identity (using a permuted one-way ANOVA with x groups, where x was the number of unique images followed by a permutated t-test for the image with the maximal response among all images). In the initial screening sessions, 99 cells (of 670, 14.77%) qualified as concept cells. In the Sternberg task, 93 cells (of 465, 14.2%) qualified as concept cells (**Fig. 2a**). Most concept cells were located either in the amygdala (**Fig. 2b**; 25.4% of all amygdala cells, permutation test, P = 0.002 versus chance) or the hippocampus (12.1%, permutation test, P = 0.002 versus chance). The number of concept cells was not larger than expected by chance in the dACC (4.35%, permutation test, P = 0.21) or pre-SMA (2.55%, permutation test, P = 0.806). We will use the term 'preferred image' for the picture for which a given cell was selective and the term 'nonpreferred images' for all other pictures. The responses of concept cells were highly selective: the average depth of selectivity index was 0.68 (**Supplementary Fig. 1b** and Online Methods). During the screening task, only 9.61% ± 9.2% (± s.d.) of the presented pictures evoked a selective response in at least one cell. In contrast, during the Sternberg task, on average 56.92% ± 36.37% (± s.d.; permuted paired t-test, t_{12} = 5.48, P = 0.001) of all presented pictures elicited a response in at least one concept cell (**Fig. 2c**). Of the units that were identified as significantly selective in the screening task, 86% remained selective for the same visual stimulus during the Sternberg task. This indicates that the screening procedure was effective, that recordings were stable and that most neurons maintained their tuning between sessions.

Activity of concept cells during working memory maintenance

We found that concept cells located in the MTL (hippocampus and amygdala) remained active during memory maintenance if the preferred image of a cell was presently held in memory (**Figs. 3a**–**d** and **4a**, and **Supplementary Fig. 3a**,**b**). Note that the preferred and nonpreferred images for each cell were defined based only on the activity observed during encoding, making this analysis unbiased. We used the picture selectivity index (PSI) to quantify the activity of each concept cell during maintenance (Online Methods). During maintenance, the PSI for concept cells recorded in the amygdala (n = 57) was 49.2 ± 65.88 (permuted paired t-test, t_{56} = 4.79, P = 0.0005 versus zero; **Fig. 4b**). Similarly, the PSI for concept cells recorded in the hippocampus (n = 21) was 28.8 ± 51.85 (permuted paired t-test, t_{20} = 2.54, P = 0.0025 versus zero). Tested individually for each neuron, 44.9% (n = 35) of the concept cells in the MTL exhibited persistent activity during WM maintenance (P < 0.05, one-tailed test). This suggests that persistent activity of concept cells in the MTL supported WM. If so, we expected that the PSI would systematically covary with memory load and success or failure in maintaining a memory. We next investigated this hypothesis.

We found that the higher the load, the lower the PSI (permuted repeated-measures 1 × 3 ANOVA, $F_{2,112}$ = 7.31, P = 0.0005 for amygdala and $F_{2,40}$ = 8.48, P = 0.0005 for hippocampus; **Fig. 4c**). Amygdala neurons had significant PSI values for all loads, whereas hippocampal neurons maintained above chance values only for loads 1–2. We next tested whether the drop in PSI as a function of load could be explained by lingering visually evoked activity. If so, persistent activity should only be present for the last picture encoded[9]. However, we found no significant relationship between the position during encoding (using loads 2–3 only) and neither the PSI in load 2 (permuted paired t-test, t_{77} = 0.38, P = 0.702) nor that in load 3 (permuted repeated-measures 1 × 3 ANOVA, $F_{2,231}$ = 0.135, P = 0.873; **Figs. 3c**,**d** and **4d**). Thus,

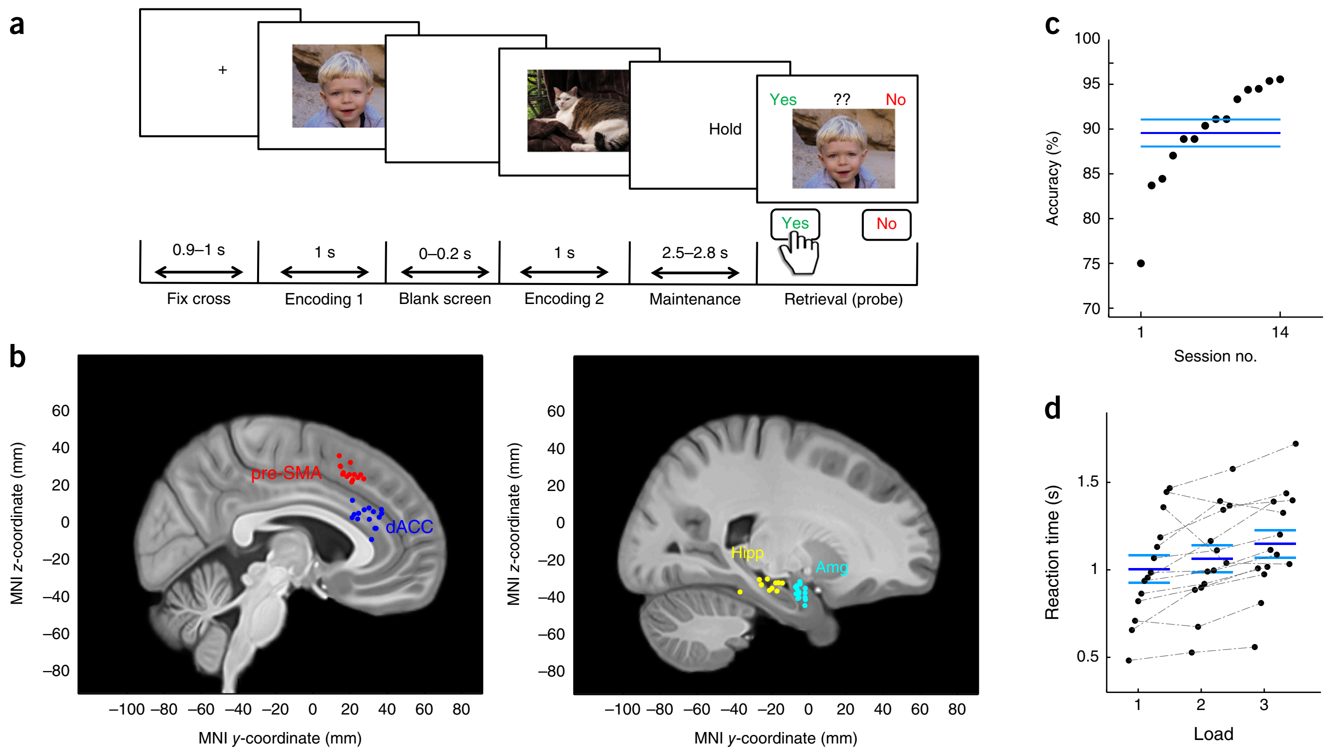

Figure 1 Task, recording locations and behavioral results. (**a**) The task. Top: examples of the screens presented to the subjects during an example trial (with load 2). Bottom: the lengths of time for which each screen was shown. Each trial consisted of 1–3 sequentially presented pictures (encoding), followed by a variable delay (holding or maintenance period). After the delay, a probe image was shown and subjects indicated whether the probe was or was not shown during the immediately preceding encoding period. (**b**) Location of recording sites in Montreal Neurological Institute's MNI152 space displayed superimposed on the California Institute of Technology's CIT168 T1w brain atlas (Online Methods). Recording locations are indicated by different colors (red, pre-SMA; blue, dACC; yellow, hippocampus (Hipp); cyan, amygdala(Amg)). (**c**,**d**) Behavioral results. (**c**) Accuracy of all sessions, rank-ordered. (**d**) Median reaction time (relative to onset of the probe image) as a function of load. Each dashed line connects an individual session. In **c** and **d**, thick and light blue lines represent the mean and s.e.m. across all sessions, respectively.

the persistent activity during maintenance could not be explained by lingering visually evoked activity.

We next tested whether the activity of concept cells during the maintenance period correlated with the correctness of the later response. For the subset of trials during which the preferred stimulus of a concept cell was held in memory, persistent activity of concept cells was significantly larger in correct compared to incorrect trials (permuted paired t-test, $t_{77} = 4.92$, $P = 0.005$; **Fig. 4e**). In contrast, when the preferred stimulus of a cell was not held in memory, there was no significant difference between correct and incorrect trials (**Fig. 4e**; permuted paired t-test, $t_{77} = 0.76$, $P = 0.48$).

Lastly, if cells showed persistent activity, it should be possible to select for concept cells using only maintenance activity (load 1 only). Indeed, we found that in both the amygdala (14.9%, $n = 29$, permutation test, $P = 0.002$) and the hippocampus (11.8%, $n = 18$, permutation test, $P = 0.002$), a significant proportion of neurons qualified as concept cells using this approach.

Together, this shows that concept cells exhibited selective persistent activity during maintenance of items in WM and that the amplitude of this activity was correlated with the quality of the memory trace.

Maintenance neurons

The second group of neurons we characterized was maintenance neurons. These neurons increased their firing rate relative to baseline during the maintenance period regardless of the stimulus that was held in memory. While we observed maintenance neurons in all recorded areas (**Fig. 5c**), they were most prominent in the two areas of the MFC ($\chi^2_1 = 25.08$; $P = 5.473 \times 10^{-6}$): pre-SMA and dACC, in which 31.3% and 21.2% of all recorded neurons were maintenance neurons, respectively. Was the activity of maintenance neurons predictive of which image(s) were held in memory? Using the same approach as in the case of concept cells (Online Methods) we found that the activity of maintenance neurons was not indicative of the images held in memory; only 5 of 126 neurons (3.97%) showed significant differences ($P = 0.829$ versus scrambled data).

We next tested whether the firing rates of maintenance neurons were associated with load and retrieval performance (RT and accuracy). Indeed, during maintenance, the firing rates of subsets of maintenance neurons in dACC (10.7%) and pre-SMA (46.3%) covaried systematically with load (**Fig. 5d** and **Supplementary Fig. 3c**). Additionally, in pre-SMA, the majority of such neurons decreased their firing rate as a function of load (9.2% versus 37%; **Fig. 5a**,**d**). In addition, the activity of 18.5% of maintenance neurons recorded in the pre-SMA and 10.3% (3 cells) of amygdala maintenance neurons differentiated between slow and fast later memory retrieval (median split computed independently for each load and for in and out conditions; median was computed individually for each subject to account for individual differences in RT; **Fig. 5e** and **Supplementary Fig. 3d**). Lastly, the mean firing rate of maintenance neurons in both dACC and amygdala, but not in pre-SMA or hippocampus, was significantly higher in trials that were later correctly remembered (**Fig. 5f**). Together, this shows that the MFC contained neurons with persistent activity during

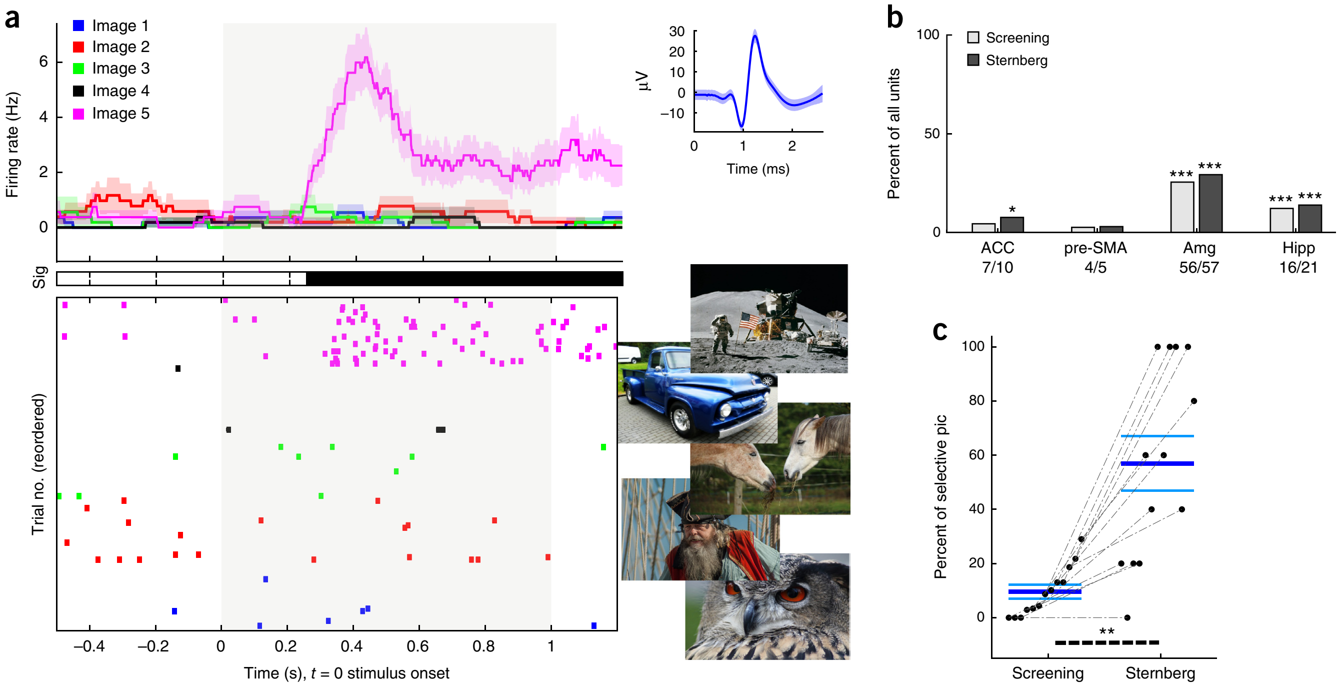

Figure 2 Stimulus-selective concept cells. (**a**) Example concept cell recorded from the amygdala. Top: poststimulus time histograms (PSTH; bin size, 200 ms; step size, 2 ms). Colors denote the different images (shown on the right). Shaded areas represent ± s.e.m. across trials. Middle: periods of significance (sig; 1 × 5 ANOVA; corrected for multiple comparisons using a cluster-size correction; Online Methods). Bottom: raster plot of trials reordered according to image identity for plotting purposes only. Image onset is at $t = 0$ (gray bar). The right inset shows the mean extracellular waveform ± s.e.m. of all spikes associated with this cell. (**b**) Percent of all recorded cells that qualified as concept cells in each area during the screening and WM (Sternberg) tasks. The numbers below each area label denote the numbers of neurons associated with each bar. We tested whether the observed percentage was higher than that expected by chance by comparing with a null distribution estimated after scrambling the condition labels (repeated 500 times; $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P \leq 0.002$). (**c**) Percentage of images shown for which we observed at least one concept cell in both tasks ($P = 0.0014$). Each dashed line connects an individual session. This shows that the screening task successfully identified responsive neurons. Thick and thin blue lines represent the mean and ± s.e.m., respectively.

maintenance. This persistent activity was not indicative of memory content but correlated with memory load and later performance.

Probe neurons

The third group of cells we characterized was probe neurons. These neurons increased their firing rate only during the presentation of the probe stimulus, relative to encoding and maintenance (**Fig. 6a**,**b** and Online Methods). The response of probe neurons was not visually selective: only 5 of 77 neurons (6.49%) showed image selectivity ($P = 0.385$ versus scrambled data). Additionally, probe neurons did not respond to the identical stimuli when presented during encoding ($P = 0.0005$, permutation test; **Fig. 6d**). We identified probe neurons in all areas, but most prominently in pre-SMA ($\chi^2_1 = 44.21$; $P = 2 \times 10^{-11}$; **Fig. 6c**), where 26.2% of neurons were probe neurons. Next, we tested whether the response of these neurons was related to visual input or movement initiation by computing the maximal firing rate after aligning to stimulus onset and to button press (RT). Aligning the response with stimulus onset resulted in significantly higher peak firing rates (mean for all probe neurons: 8.82 Hz ± 8.74) compared to the peak firing rate when aligned to button press (5.85 Hz ± 7.5, permutation t-test: $t_{76} = 8.71$; $P = 0.0005$, **Fig. 6a**,**b**). This suggests that these neurons responded to the probe itself rather than to movement initiation.

What was the relationship between the activity of probe neurons and behavior? A minority of the probe neurons recorded in pre-SMA (15.6% of probe neurons) showed a differential response as a function of the WM-based decision (in versus out; **Fig. 6e**). Similarly, a small number of neurons differentiated the button (left or right) used to communicate the decision (**Fig. 6f**). However, the large majority of probe neurons (84%) signaled neither. Instead, these neurons showed a strong but indiscriminate increase in firing rate to all pictures shown as probes. We hypothesize that probe neurons signaled a change of task phase, i.e., a switch from maintenance to retrieval of information held in WM (see "Discussion" section below).

Separability of neuronal categories

Were the three neuronal classes we identified separate or were there neurons that qualified as multiple types? Note that concept and maintenance cells were observed preferentially in separate anatomical areas (MTL versus MFC), supporting the argument that they were separate. Nevertheless, there was some overlap in the same areas (**Supplementary Fig. 4a**). However, further analysis of the correlations between effect sizes attributable to different factors showed that the three types we identified constituted three largely separate categories of neurons (**Supplementary Fig. 4b**,**c**). Lastly, we confirmed that the effects observed did not differ between neurons recorded within versus outside of epileptic tissue later resected (**Supplementary Fig. 4d**).

Decoding of information held in WM

We next assessed how well neuronal activity during maintenance predicted the identity of remembered images and/or performance in a given trial. For this, we trained a decoder on a pseudopopulation of all neurons recorded of a given kind (for example, only MFC neurons

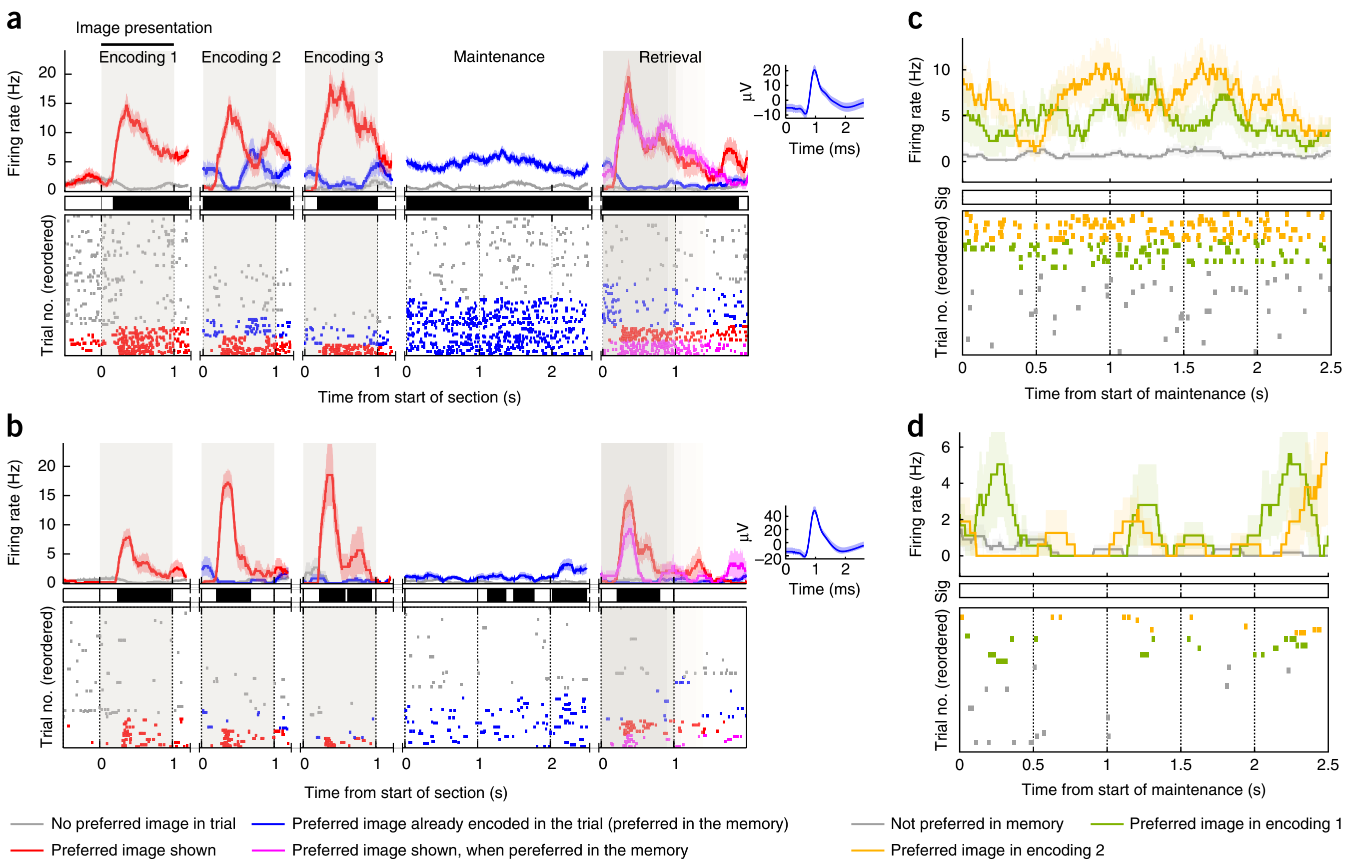

Figure 3 Concept cells are persistently active during WM maintenance. (**a**,**b**) Two example concept cells recorded from (**a**) amygdala and (**b**) hippocampus. Top: PSTH (bin size, 200 ms; step size, 2 ms). Shaded areas represent ± s.e.m. across trials. Insets: mean extracellular waveform ± s.e.m. of all spikes associated with that cell. Middle: periods of significance (black) between preferred vs. nonpreferred stimuli (corrected for multiple comparisons using a cluster-size correction; Online Methods). Bottom: raster plot of trials reordered according to condition for plotting purposes only. Both neurons show both visually evoked selective activity (red) and sustained activity (blue) during maintenance. Note that during maintenance, concept cells have elevated activity only when their preferred stimulus was held in memory (blue vs. gray), and that the sustained activity (blue) was suppressed during encoding of the nonpreferred image (i.e., encoding 3) when the preferred stimulus was already held in memory. (**c**,**d**) Maintenance activity of the same neurons shown in **a** and **b** but only for the subset of trials with load 2 (two items held in memory). There was no significant difference in activity between trials where the preferred image was shown first vs. second (encoding 1 or 2; middle panel marks periods of significance). See **Figure 4d** for a similar analysis at the population level. See **Supplementary Figure 3** for further single-cell examples.

or only concept cells) on a subset of trials and tested its performance on an independent set of test trials (Online Methods). We used this decoding approach to assess whether activity during maintenance was strong enough to be read out at the single-trial level. We found that neurons in the MTL, but not in MFC, carried information about picture identity during load 1 trials (**Fig. 7a**). Because there were multiple correct answers for loads 2 and 3, we used a separate binary decoder (trained in load 1 trials) for each image to predict whether it was currently in memory or not in load 2 and 3 trials, and we averaged performances across all five images. This allowed us to compare performances between all load conditions. The average decoding accuracy for amygdala and hippocampal neurons during load 1 was 83.5% and 63.4%, respectively (**Fig. 7a** and Online Methods; $P = 0.002$ and $P = 0.007$, respectively, estimated using scrambled data; chance was 50%). We observed a sharp drop in decoding accuracy as a function of increasing load (**Fig. 7a**) in both amygdala and hippocampus. Nevertheless, decoding performance was above chance for loads 2 and 3 in the amygdala (load 2: 60.7%, $P = 0.008$; load 3: 56%, $P = 0.039$) but not in the hippocampus (load 2: 54.2%, $P = 0.25$; load 3: 47.7%, $P = 0.73$). We also performed the same analysis using concept cells only (for load 1 trials). We found little difference in decoding accuracy between using only concept cells and using the whole population of all recorded neurons (**Fig. 7a**, $P = 0.88$ for amygdala comparison and $P = 0.32$ for hippocampus). Thus, the information decodable about picture identity from the population was carried principally by concept cells.

We next tested whether information present in the neuronal firing rate of concept cells was carried by a static or dynamic code[27,28]. We tested this by evaluating whether a decoder trained at one point in time could decode information obtained from a different timepoint. We found that regardless of the time the decoder was trained, decoding performance was significantly above chance (**Fig. 7b**): for example, the average decoding accuracy during the maintenance period of a decoder trained based on the activity during the encoding period time (300 ms after image exposition) was 0.58 (chance level = 0.2, $P = 0.002$ versus chance as estimated using a permutation test). Together, this shows that concept cells coded for the identity of the items held in memory using a static code. We next tested whether activity during maintenance was indicative of the number of items (load) held in memory. We found that decoding memory load was possible in both pre-SMA (decoding accuracy: 67.96%, $P = 0.002$, chance = 33%; **Fig. 7c**) and dACC (accuracy = 42.88%, $P = 0.027$) but not in amygdala

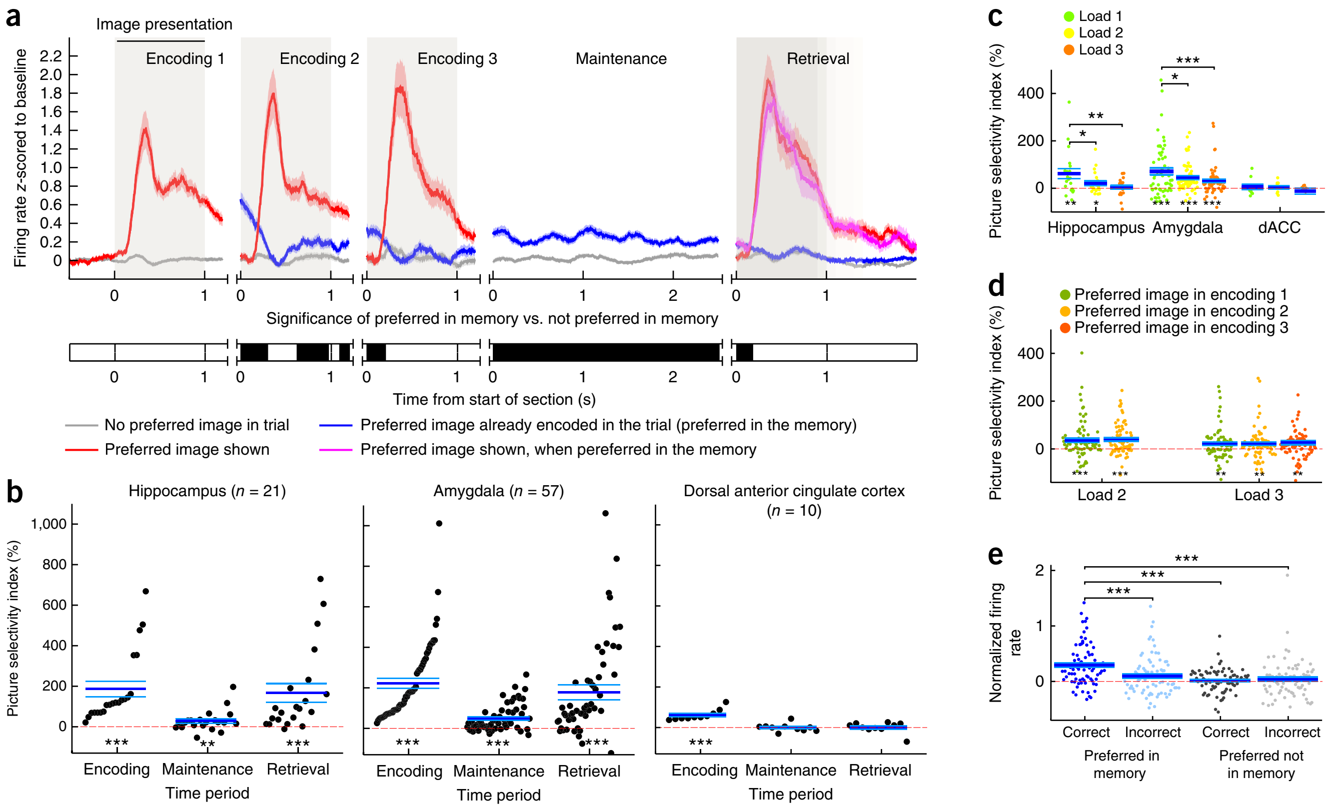

Figure 4 Population analysis of MTL concept cells. (**a**) Average firing rate of all concept cells identified in the amygdala (n = 57) in the different phases of the task. Shaded areas represent ± s.e.m. across neurons. Gray vertical bars mark periods of time during which an image was on the screen. Bottom panel marks points of time during which the activity of the cells was significantly different (black) between trials when a preferred image was in memory vs. when it was not (corrected for multiple comparisons based on cluster size; Online Methods). Colors mark different trials as indicated. For subplots **a**–**d**, only correct trials were used. (**b**) PSI during encoding, maintenance and retrieval for all identified concept cells (each data point is one neuron; data points are sorted according to the encoding phase of the task). Neurons in both amygdala and hippocampus, but not in dACC, maintained their selectivity throughout the task and showed persistent activity. Significance was computed against chance (PSI = 0). (**c**) PSI for different load conditions indicated that neurons maintained persistent activity for loads 1–3 in amygdala and 1–2 in hippocampus, but did not do so in dACC. (**d**) PSI for loads 2 and 3 as a function of whether the preferred image was shown first, second or third during encoding. This shows that images that were shown directly before the maintenance period did not have greater selectivity (load 2, $P = 0.702$; load 3, $P = 0.873$). (**e**) Relationship between firing rates of concept cells in the MTL and behavior. The firing rate was significantly higher for correct compared to incorrect trials only when the preferred stimulus was held in memory. For **c**–**e**, PSI and firing rate were calculated for the entire maintenance period. $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ as estimated with permutation tests. In **b**–**e**, thick and thin blue lines represent the mean and ± s.e.m., respectively. Pre-SMA is not shown in this figure because we did not identify any concept cells in this area.

or hippocampus (**Fig. 7c**). When we restricted the decoder to only information provided by the previously identified maintenance neurons, we found that load decoding performance in pre-SMA did not change significantly (59.8% versus 67.96%, $P = 0.061$). In contrast, in dACC, allowing access only to maintenance neurons reduced decoding performance to chance levels (performance accuracy = 38.04%, P = 0.164). Thus, activity of MFC cells did not carry information about the content of the memory itself but rather about other aspects of the memory: the current load. Additionally, we observed that in pre-SMA maintenance neurons carried most of the information about memory load. Note that the activity of concept cells in the MTL also varied as a function of load (**Fig. 4c**) but only for the preferred stimulus of a cell. In contrast, the activity of MFC cells varied as a function of load regardless of the stimulus held in memory (as the above decoding results demonstrate).

Finally, we tested whether it was possible to predict the RT taken to make the in-versus-out decision at the end of the trial based on the activity of neurons during the maintenance period (**Fig. 7d**). Stronger memories generally lead to faster, more accurate and more confident decisions[29]. We found evidence for this behavior in our task, because incorrect decisions took longer (see "Task and behavior" section above). Based on this rationale, we split correct trials into two groups based on the median RT of each subject (slow and fast groups). We did this separately for each load and in or out group (resulting in six groups total) to account for the systematic RT differences that resulted from different loads and in-versus-out decisions. We found that RT could be predicted from neurons located in pre-SMA (accuracy = 65.28%, chance = 50%, P = 0.002; **Fig. 7d**) and amygdala (accuracy = 60.86%, P = 0.0027; **Fig. 7d**) but not from dACC or hippocampal neurons. Repeating the same analysis using only maintenance neurons revealed similar performance in pre-SMA (accuracy = 59.8%, $P = 0.0157$ versus chance). In amygdala, decoding based on maintenance neurons only reduced performance to chance levels (accuracy = 52.65%, $P = 0.2$).

Together, this shows that information about the content of working memory was provided by the firing rate of concept cells in the MTL (**Fig. 7a**) and that information about the quality of memories (load, later RT) was encoded predominantly by neurons in the MFC (**Fig. 7c,d**).

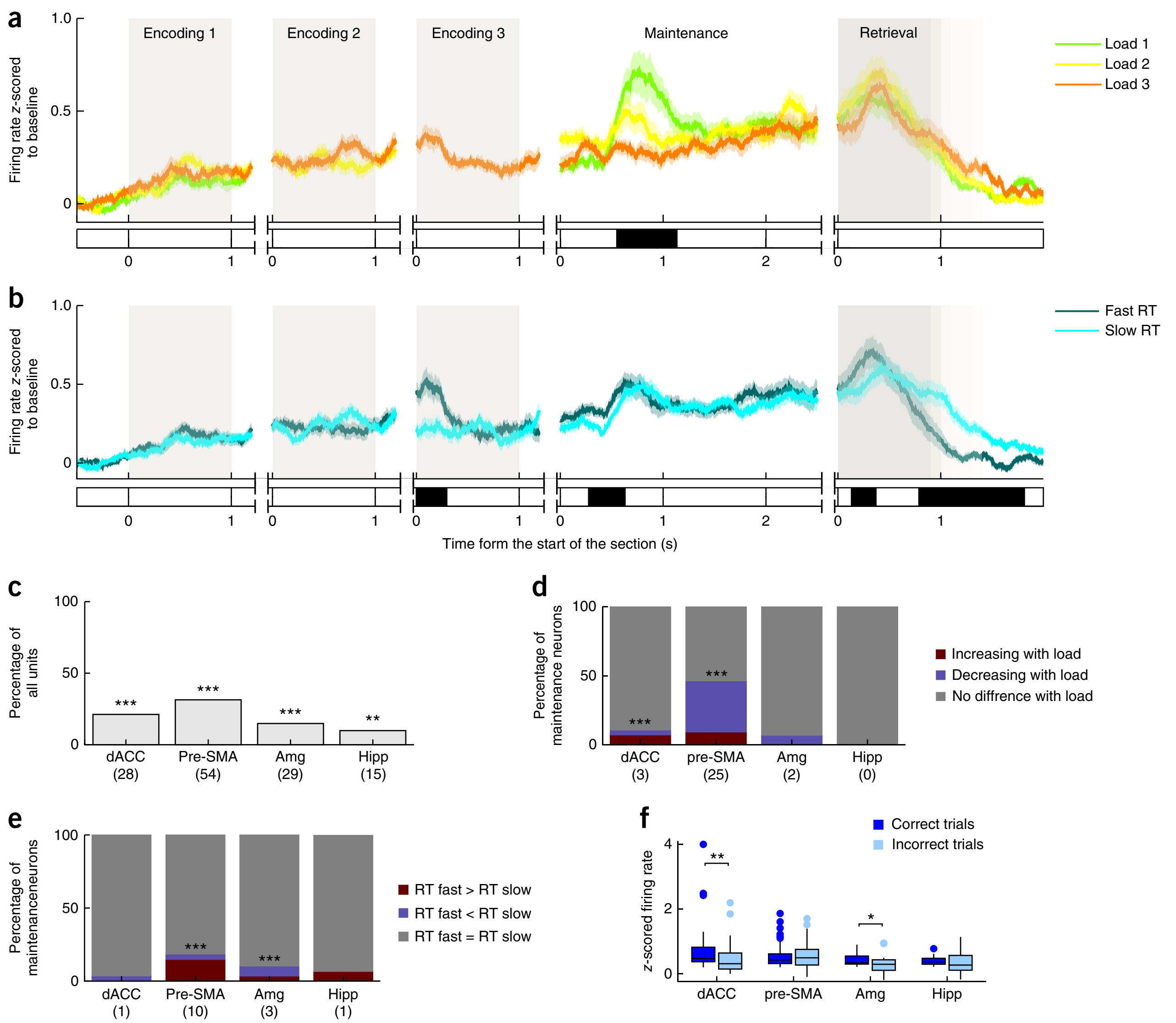

Figure 5 Persistent activity of maintenance cells in MFC. (**a**) Average firing rate of all maintenance neurons ($n = 54$) in pre-SMA for different load conditions and (**b**) for trials that ended with a fast or slow response (median split of RT). Shaded areas represent ± s.e.m. across neurons. Gray vertical bars mark image presentation. Black bars indicate significance at $P < 0.05$ of a 1 × 3 (top) permutation ANOVA and 1 × 2 (bottom) permutation *t*-test with load as the dependent variable for **a** and response time for **b**. Multiple comparisons were corrected for using a cluster-size approach (Online Methods). (**c**) Percentage of all recorded cells identified as maintenance neurons in each area. Numbers below the area label denote numbers of cells. The medial frontal areas (dACC, pre-SMA) contained significantly higher proportions of maintenance neurons ($\chi^2_1 = 21.1$; $P = 4.353 \times 10^{-6}$) compared to areas in the MTL. (**d**) Percentage of maintenance neurons whose firing rate during maintenance differed as a function of load. Notably, in pre-SMA, 37% of cells decreased their firing rate as a function of load. (**e**) Percentage of maintenance neurons whose activity during maintenance differed as a function of response time. (**f**) The firing rates of maintenance neurons in dACC and amygdala differed as a function of whether stimuli were later remembered or forgotten. In **c**–**e**, significance was assessed by comparing with a null distribution estimated using bootstrapping (Online Methods). For **f** we used a permutation test. Boxplots represents quartiles (25%, 75%); horizontal lines represent medians; whiskers show ranges up to 1.5 times the interquartile range; dots above whiskers show outliers. $*P < 0.05$, $**P < 0.01$ and $***P \leq 0.002$.

Working memory as attractors in state space

We next quantified the dynamics of neuronal activity during memory maintenance. An attractor is a location in state space, which in our case was formed by the activity of all recorded neurons. Here an attractor was thus a pattern of firing rates. Theoretical work suggests that during WM maintenance, the neural trajectories reside close to a point in this state space (the attractor) to maintain information in WM. This space could potentially assume a variety of patterns, including a static code[22], dynamic patterns[28] or combination of both[30]. Our goal was to determine whether the neuronal trajectories during the maintenance period were compatible with an attractor formed by a static code. We used demixed principal component analysis (dPCA) as a dimensionality reduction technique[31] to derive basis functions into which all neural activity was projected. The projection matrix of dPCA was computed based only on activity during encoding period 1 (first image) on all neurons. Afterwards, we projected the activity during the other encoding and the maintenance periods into this same space without recomputing the basis functions, making this analysis unbiased. We used dPCA with picture identity as the marginalized variable[31]. We found that a

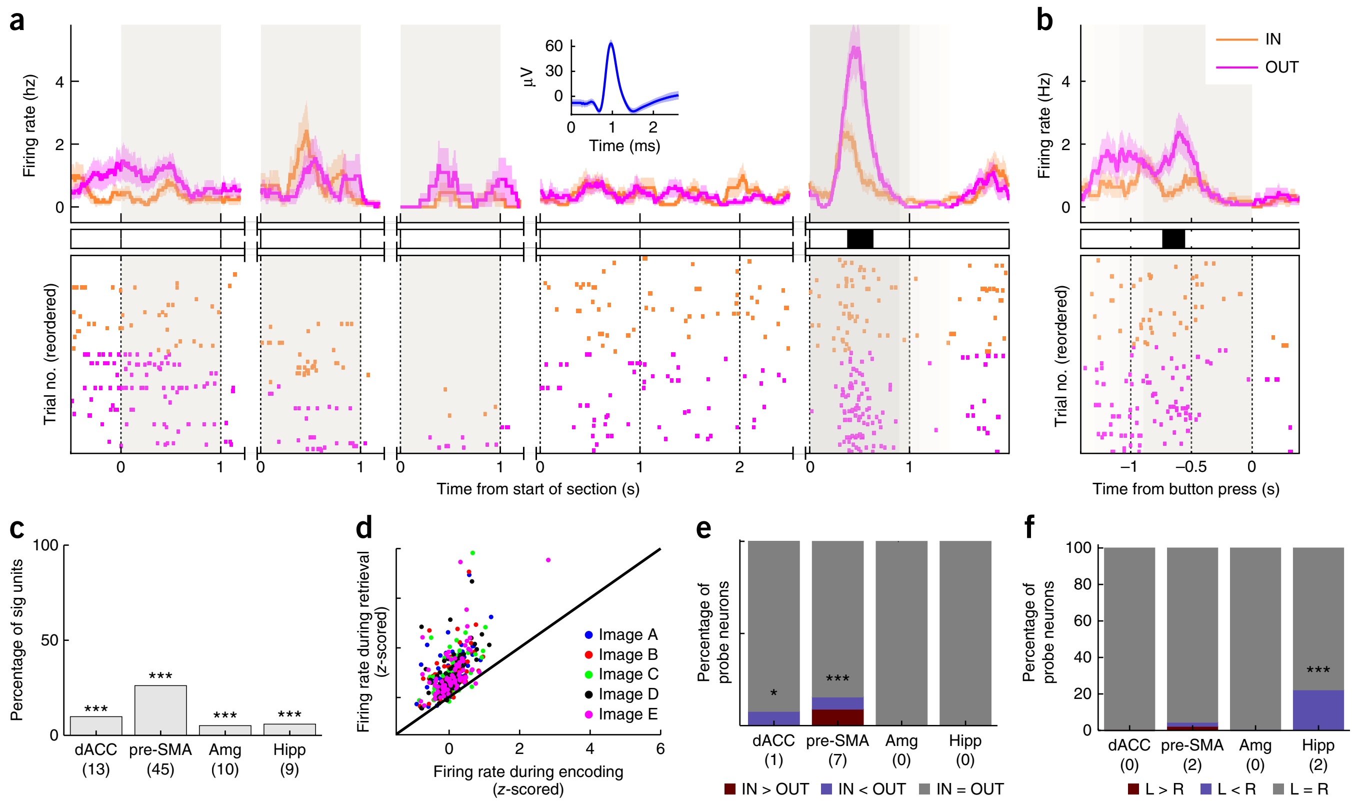

Figure 6 Probe neurons reflect WM-retrieval related evoked activity in MFC.(**a**) Firing rates of an example probe neuron recorded from the pre-SMA, shown separately for trials in which the probe was held (IN, cyan) or not held (OUT, magenta) in memory. Top: PSTH (bins, 200 ms; step size, 2 ms; shaded areas represent ± s.e.m.). Middle: points of time with significant differences between IN and OUT trials, corrected for multiple comparisons using a cluster-size approach. Bottom: raster plot of reordered trials. Gray vertical bars mark image presentation. (**b**) Same neuron as in **a**, but with response aligned to button press (button press was at 1.2 s after image presentation). Note the much-reduced peak response (1.44 Hz vs. 0.45 Hz, permuted *t*-test, $P = 0.005$). (**c**) Percentage of probe neurons in each area. Probe neurons were most prominent in pre-SMA, followed by dACC. (**d**) Probe neurons elevated their firing rates only during retrieval, not during encoding ($P = 0.0002$, permuted *t*-test). (**e**–**f**) Percentages of probe neurons in each area whose firing rates during probe (–800 to 0 ms relative to button press) differed as a function of (**e**) IN vs. OUT or (**f**) button press. Most cells showed no difference, i.e., they responded equally strongly (but selectively) to the probe stimulus. In **c**, **e** and **f** significance was assessed based on a null distribution estimated based on permuted labels. $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$.

four-dimensional space formed by four demixed principal components (components 1, 2, 3 and 5) separated the neural trajectories well (30.7% explained variance) between the five different image identities and had the highest percentage of variance attributed to picture identity (**Fig. 8a**,**b** and **Supplementary Figs. 5** and **6d**; see also **Supplementary Video 1**).

We first analyzed the rate of change (velocity) in this four-dimensional neural state. As expected, velocity was highest during encoding due to the strong visual-onset transients of concept cells (**Fig. 8c**). In contrast, velocity was substantially reduced at baseline (fixation cross). Unexpectedly, the amount of change in the activity during maintenance was reduced: the velocity during maintenance was not significantly different from that during baseline ($P = 0.233$; **Fig. 8c**). We next quantified the pairwise distance between the neural trajectories associated with different images held in memory (only for load 1 because pairwise distance was ill defined if several items were held in memory; **Fig. 8d** and see below for other loads). Distances were maximal during encoding but were maintained significantly above baseline levels during maintenance ($P = 0.0005$; **Fig. 8d**). Thus, the neural trajectories during the maintenance period were drawn to particular locations in state space, which is the definition of an attractor[21,32]. Therefore, our data suggests that neural activity during maintenance clustered around attractors (**Fig. 8d**).

We next tested whether the distance of the neural trajectories from a given attractor was correlated with behavior (accuracy and RT). We defined the attractor location based on the activity during maintenance of one item and quantified the distance of the neural activity to this activity in each trial using a distance metric (distance-to-attractor or DA; Online Methods). DA was defined as the Euclidian distance across trials between the attractor center for a particular image divided by the average distance of this attractor from all other attractors representing all the other pictures (**Fig. 8e**). We found that the DA was smaller for trials in which the item in memory was correctly remembered compared to trials where the item was forgotten (permutation test: $P = 0.002$, computed for loads 1–3 separately and averaged; **Fig. 8f** and **Supplementary Fig. 6c**). This difference was abolished when we excluded concept cells from the population, but excluding maintenance neurons did not alter this effect (**Fig. 8f**). Similarly, excluding neurons recorded in MFC did not affect this result, whereas excluding MTL neurons eliminated the observed differences (permutation test, $P = 0.002$ and $P = 0.664$, respectively; **Fig. 8f**). We also tested whether the distance from the attractor predicted later decision time. We tested this for the subset of trials in which the probe was held in memory (IN trials). Faster decisions were associated with smaller DA values, indicating that neural trajectories (during maintenance) were closer to the given

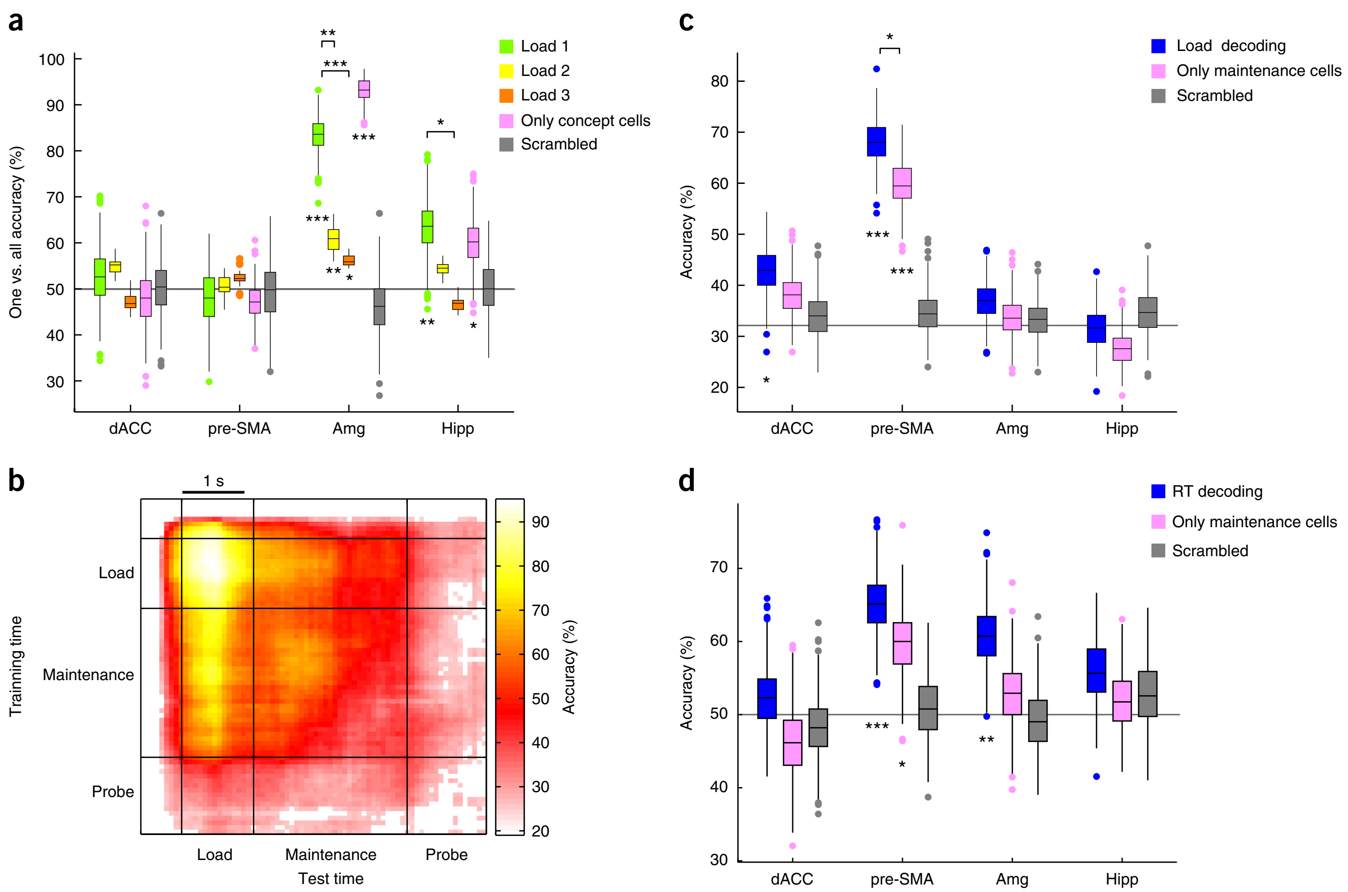

Figure 7 Population decoding from all recorded neurons during the maintenance period. (**a**) During maintenance, picture identity could be decoded from the activity of amygdala and hippocampal neurons but not from that of dACC or pre-SMA. Decoding performance was maintained when only cells identified as concept cells were considered (magenta). 'One vs. all' denotes average accuracy of decoders trained to distinguish between a given image and all the others (50% chance level). (**b**) Picture identity decoding using different time windows for training and testing. Shown is the test–retest decoding performance for load 1 trials (chance level is 20%). (**c**) The activity of neurons in the pre-SMA during maintenance was predictive of how many items were held in memory (load, 1–3). Decoding only from maintenance neurons (magenta) was sufficient. (**d**) The activity of neurons in the pre-SMA and amygdala during maintenance was predictive of later response speed. $*P < 0.05$, $**P < 0.01$ and $***P \leq 0.002$. Markers below bars indicate significance vs. chance performance, estimated by randomly scrambled labels. Significance of pairwise tests was estimated by comparisons with a null distribution of the same differences estimated from decoders trained on data with randomly scrambled labels. In **a**, **c** and **d**, boxplots represent quartiles (25%, 75%); horizontal lines indicate medians; whiskers show ranges up to 1.5 times the interquartile range; dots above whiskers show outliers.

attractor (**Fig. 8g**). As before, that was true only when concept cells and MTL were present in the population.

We next repeated the same analysis only using simultaneously recorded neurons (n = 5 subjects who had at least four simultaneously recorded concept cells). As in the pseudopopulation results, trial-by-trial, the distance to the closest attractor was predictive of whether the answer given by the subject in a trial would be correct or incorrect (permutated t-test, $t_{624} = 3.33$, $P = 0.003$; **Fig. 8h**). This shows that attractors could be seen in individual subjects and that the distance to the attractor was informative in individual trials.

In conclusion, our results show that persistent activity during maintenance created attractors and that the distance of the neural trajectory in a given trial predicted the quality of the memory trace as measured by reaction time and accuracy.

DISCUSSION

We found neurons with two different types of persistent activity in MTL and MFC during WM maintenance. In the MTL, concept cells were persistently active, and this activity was characterized by a high degree of selectivity. In the MFC, on the other hand, we observed a group of neurons (**Fig. 5c**) that tonically increased their activity during maintenance, but this activity was indifferent to the identity of the memoranda currently held in the memory.

Preceding the WM task, we used a screening task to identify visually selective 'concept cells' similar to those first identified by Quiroga *et al.*[24,33–35]. Such concept cells responded strongly and selectively only to the preferred images but not to the nonpreferred images (**Supplementary Fig. 1a**). Notably, we identified more concept cells in the amygdala (25.4%) than in the hippocampus (12.1%), a trend compatible with that observed previously[34]. It remains an open question whether concept cells also exhibit persistent activity for stimuli other than images, such as text or audio[34]. In addition to transient activation by visual input, concept cells can also be activated by free recall of episodic memories[36] and visual imagery[37]. While this has so far been interpreted as representing recall from long-term memory, we now show a direct relationship between persistent concept-cell activity and WM. Specifically, we found that the response strength (firing rate) of concept cells for images held in memory decreased as a function of memory load, an observation that fits models of persistent activity[38].

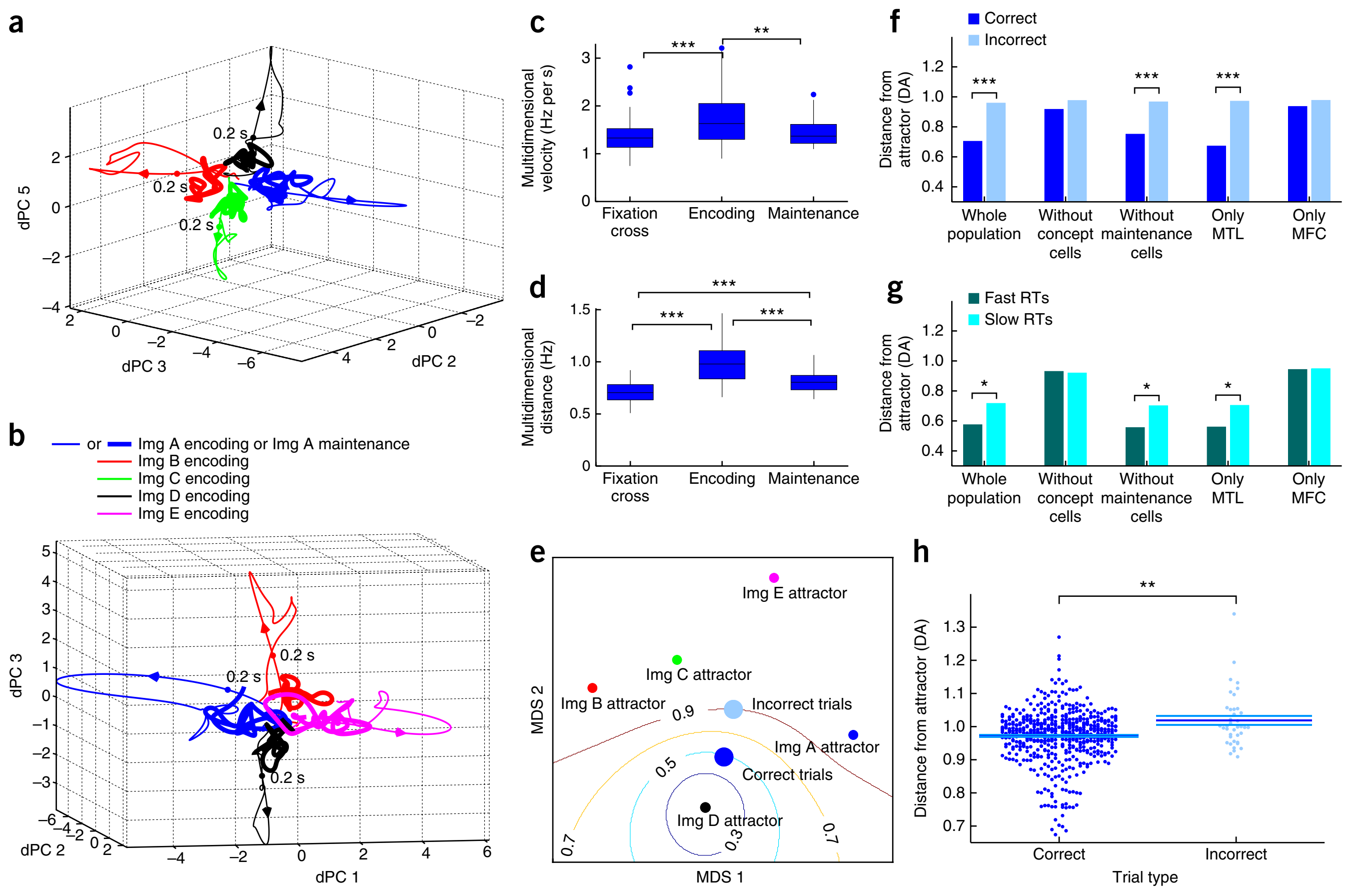

Figure 8 Persistent activity during maintenance forms attractors. (**a**) Illustration of the mean trajectories in neuronal state space formed by the three demixed principal components (dPCs) associated with picture identity during encoding (thin line) and maintenance (thick line). The dot indicates the timepoint 200 ms after image onset and the arrow indicates the direction of change. Colors mark different images (img; only 4 of the total 5 are shown for clarity). (**b**) Rotated view of **a**. (**c**) Multidimensional velocity of the population in the different phases of the task (shown for all load 1 trials). The velocity during maintenance was significantly slower than that during encoding and was not significantly different from that during baseline. (**d**) Multidimensional pairwise distance between all possible pairs of attractors during maintenance (load 1). The distance during maintenance was significantly larger than that during baseline ($P = 0.0005$). Together, **c** and **d** are indicative of attractors. The significance of the population metrics shown in **c** and **d** was computed by randomly subsampling a subset of trials and neurons (Online Methods). (**e**) Schematic representation of the distances between the attractors (attractors are defined based on correct load 1 trials) for each image (small filled dots) and the average position in state space for Image D for two behaviors: remembered (correct) and forgotten (incorrect) computed for all loads separately and averaged. This representation was determined based on multidimensional scaling of the state space (Online Methods). Isolines depict areas of equal distance from the attractor for Image D. (**f**) DA was significantly smaller for correct than for incorrect trials only when concept cells were part of the population. Note that DA < 1 indicates that the trajectory is closer to the correct attractor than to all the other attractors. (**g**) The DA corresponding to the remembered image was indicative of the speed of the response. This relationship was observed only for concept cells. (**h**) DA predicts performance on individual trials. Each dot is one trial. In **c** and **d**, boxplots represent quartiles (25%, 75%); horizontal lines indicate medians; whiskers show ranges up to 1.5 times the interquartile range; and dots above whiskers show outliers. $*P < 0.05$, $**P < 0.01$ and ***denotes $P \leq 0.002$.

We found that the dynamics of self-sustained persistent activity were well described as an attractor state. Attractor-like behavior has been observed in animal experiments as well as in theoretical work[20–22,39]. Modeling work shows that attractor networks can hold multiple pieces of information encoded by separate attractors[40] in a robust and noise-resistant manner[39]. Theoretical work further indicates that the activity of attractor networks can drift during the maintenance period due to proximity to other attractors in state space[20,39]. This prediction has been confirmed experimentally in macaque recordings with a single item held in memory, where the extent of drift away from an attractor was predictive of errors[22]. Here we present evidence that the activity of human neurons during WM maintenance is compatible with that expected if attractors were present. We show that this prediction of the model holds for discrete attractors and for multiple items held in memory, because we found that the quality of a memory trace could be predicted by the drift (distance) away from the attractor representing a given image (**Fig. 8**). In addition, we found that it was principally concept cells in the MTL that supported attractors (**Fig. 8f,g**). Here we defined each attractor based on the pattern of activity observed when only a single item was held in memory. However, when multiple items were held in memory (loads 2–3), multiple concept cells were persistently active. Consequently, the population activity was different from that defined by the single-item attractors as demonstrated by increased DA values (**Supplementary Fig. 6c**). Despite this, the distance to the attractors defined in this manner was predictive of memory content and quality even in loads 2 and 3 (**Supplementary Fig. 6c**). This was possible because of the high dimensionality of the state space, which made it possible for an individual attractor to be closest to several attractors while being far away from all the others[40]. Overall, our findings reveal a direct neuronal correlate for discrete attractors in humans. This shows that persistent activity can be viewed as forming attractors and that persistent activity is a mechanism capable of supporting information maintenance during WM.

A lack of evidence for stimulus-specific persistent activity in some studies[28,41] has led to alternative proposals on how WM is maintained. A key new proposal is that information in WM can be carried by salient states encoded through synaptic weight changes[42,43] or by oscillatory bursts[41]. Our analysis now shows that persistent activity compatible with the original model existed in the human MTL. This argues that the original WM model is a valid way to conceptualize persistent activity during Sternberg-like WM tasks. However, this does not rule out the possibility that mechanisms other than persistent activity also contribute to WM.

Our data shows that the activity of MTL neurons during WM maintenance carried information about the items held in memory. This raises questions about the role of the MTL for WM. While it has been previously observed that intracranial field potentials recorded from the MTL are modulated by load[18,44], these differences are not stimulus-selective. Subjects with bilateral MTL damage do not exhibit WM deficits in some circumstances[17], but sufficiently difficult WM tasks or the presence of distractors[17] reveal WM deficits in the same subjects. Based on this, it has been suggested that the MTL is critical in conditions when WM alone is insufficient[17]. Therefore, a possible role for the persistent MTL activity we observed is the maintenance of memory engrams so that they can be used to recover information in WM that has been lost due to shifts in attention triggered by distractors or when the capacity of WM is saturated.

We found amygdala neurons with prominent persistent activity. Notably, amygdala persistent activity was more easily decodable than activity in the hippocampus. This is in contrast to episodic memory-related activity during free recall[36], which is prominent in the hippocampus but not the amygdala. This raises the novel possibility that the amygdala could have a role in supporting WM. While encoding new long-term memories does not require the amygdala, the amygdala prominently modulates this process by enhancing or suppressing the encoding of new memories[45]. Notably, patients with amygdala lesions frequently have working memory deficits[46]. Additionally, rats with amygdala lesions have a specific impairment in visual WM tasks[47]. While it is recognized that the amygdala can boost general vigilance, and thereby facilitate execution of demanding tasks[45], our data indicate that the amygdala might directly contribute to WM maintenance by storing specific memoranda with the help of persistently active neurons.

The persistent activity of MFC neurons did not carry information about the content of the memory. Nevertheless, this activity was relevant for WM, because it was predictive of later behavior (accuracy and RT) and memory load. This is similar to findings from some recordings in macaques, which also revealed nonselective persistent activity in the MFC[12]. However, other macaque studies have revealed MFC single units that are stimulus-selective[11,48]. Whether stimulus-specific persistent activity in the human MFC exists in other tasks or for other stimuli, such as tactile vibration stimuli[11], sounds[11] or task contexts[48], remains an open question. Note that in this study we did not record from dlPFC, which is frequently associated with stimulus selective persistent activity[4,5,7]. It therefore remains an open question whether human dlPFC neurons have stimulus-specific persistent activity. Nevertheless, our finding that the activity of maintenance neurons in the MFC during delay was predictive of RT and accuracy is consistent with the hypothesized role of the MFC in attentional control[49].

The dACC and pre-SMA are both part of the cingulo-opercular system. It is thought that the activity of this network supports implementation of different task sets[15]. Indeed, lesion studies have found that subjects with MFC lesions are impaired in switching between different response and instruction sets[50]. Our task consisted of three task sets (encoding, maintenance, retrieval). A role of the MFC in controlling switches between these three task sets is compatible with our finding that differences in activity due to load and RT appeared prominently only in the initial part of the maintenance period (**Fig. 5a**,**b**). This is because in each trial, the set size was unknown to the subject. Therefore, starting the maintenance period after encoding only one picture was more unpredictable than starting it after encoding three pictures. Thus, switching to maintenance after encoding one picture would be more attentionally demanding. Indeed, we found that pre-SMA cells were most active for load 1 and less so for loads 2 and 3 (**Fig. 5**). Similarly, the firing rate of MFC maintenance neurons was most predictive of RT early during the maintenance period (**Fig. 5b**), which might also be related to switching from the encoding to the maintenance mode. We suggest that these neurons support attentional control and, together with probe neurons (see below), the implementation of task sets. Future work is needed to determine whether maintenance and concept neurons interact during WM tasks.

Finally, we also observed a group of probe neurons in pre-SMA that only became active during the retrieval phase of the task. The activity of these neurons was better explained by the time of visual stimulus-onset than the timing or nature of the motor response. Probe neurons did not respond to the same stimuli during encoding, demonstrating that they were not simply visually tuned or responsive. In the context of task sets[15], probe neurons signaled the transition to the retrieval phase of the task, i.e., the switch from maintenance to retrieval of information held in WM. Together, this reveals a direct correlate of neurons implementing transitions between different tasks sets in human MFC.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Minxa, R. Adolphs and J. Dubois for discussion and the staff and physicians of the Epilepsy Monitoring Unit at Cedars-Sinai Medical Center and the Huntington Memorial Hospital for invaluable assistance. This work was supported by the National Science Foundation (1554105 to U.R.), the National Institute of Mental Health (R01MH110831 to U.R.), the McKnight Endowment Fund for Neuroscience (to U.R.), a NARSAD Young Investigator grant from the Brain & Behavior Research Foundation (23502 to U.R.) and the Pfeiffer Foundation (to U.R.).

AUTHOR CONTRIBUTIONS

J.K. and U.R. designed the experiments. J.K. and U.R. performed experiments. J.K., S.S. and U.R. performed analysis. A.N.M. and I.B.R. performed surgery. J.M.C. provided patient care. J.K. and U.R. wrote the paper. All of the authors discussed the results at all stages of the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Baddeley, A. Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* **63**, 1–29 (2012).

2. Eriksson, J., Vogel, E.K., Lansner, A., Bergström, F. & Nyberg, L. Neurocognitive architecture of working memory. *Neuron* **88**, 33–46 (2015).
3. Goldman-Rakic, P.S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
4. Constantinidis, C., Franowicz, M.N. & Goldman-Rakic, P.S. The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nat. Neurosci.* **4**, 311–316 (2001).
5. Funahashi, S., Bruce, C.J. & Goldman-Rakic, P.S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
6. Rainer, G., Asaad, W.F. & Miller, E.K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**, 577–579 (1998).
7. Chafee, M.V. & Goldman-Rakic, P.S. Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. *J. Neurophysiol.* **79**, 2919–2940 (1998).
8. Fuster, J.M. & Jervey, J.P. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science* **212**, 952–955 (1981).
9. Miller, E.K., Li, L. & Desimone, R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.* **13**, 1460–1478 (1993).
10. Suzuki, W.A., Miller, E.K. & Desimone, R. Object and place memory in the macaque entorhinal cortex. *J. Neurophysiol.* **78**, 1062–1081 (1997).
11. Vergara, J., Rivera, N., Rossi-Pool, R. & Romo, R. A neural parametric code for storing information of more than one sensory modality in working memory. *Neuron* **89**, 54–62 (2016).
12. Constantinidis, C. & Procyk, E. The primate working memory networks. *Cogn. Affect. Behav. Neurosci.* **4**, 444–465 (2004).
13. Fuster, J.M. Cortex and memory: emergence of a new paradigm. *J. Cogn. Neurosci.* **21**, 2047–2072 (2009).
14. Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C. & Wager, T.D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
15. Dosenbach, N.U.F. *et al.* A core system for the implementation of task sets. *Neuron* **50**, 799–812 (2006).
16. Squire, L.R., Stark, C.E.L. & Clark, R.E. The medial temporal lobe. *Annu. Rev. Neurosci.* **27**, 279–306 (2004).
17. Jeneson, A. & Squire, L.R. Working memory, long-term memory, and medial temporal lobe function. *Learn. Mem.* **19**, 15–25 (2011).
18. Axmacher, N. *et al.* Sustained neural activity patterns during working memory in the human medial temporal lobe. *J. Neurosci.* **27**, 7807–7816 (2007).
19. Ojemann, G.A., Creutzfeldt, O., Lettich, E. & Haglund, M.M. Neuronal activity in human lateral temporal cortex related to short-term verbal memory, naming and reading. *Brain* **111**, 1383–1403 (1988).
20. Wang, X.J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).
21. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79**, 2554–2558 (1982).
22. Wimmer, K., Nykamp, D.Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).
23. Rutishauser, U. & Douglas, R.J. State-dependent computation using coupled recurrent networks. *Neural Comput.* **21**, 478–509 (2009).
24. Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
25. Mormann, F. *et al.* A category-specific response to animals in the right human amygdala. *Nat. Neurosci.* **14**, 1247–1249 (2011).
26. Sternberg, S. In defence of high-speed memory scanning. *Q. J. Exp. Psychol. (Hove)* **69**, 2020–2075 (2016).
27. Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.* **100**, 1407–1419 (2008).
28. Stokes, M.G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
29. Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat. Neurosci.* **18**, 1041–1050 (2015).
30. Murray, J. *et al.* Stable population coding for working memory coexists with heterogeneous neuraldynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **114**, 394–399 (2016).
31. Kobak, D. *et al.* Demixed principal component analysis of neural population data. *eLife* **5**, 1–37 (2016).
32. Milnor, J. On the concept of attractor. *Commun. Math. Phys.* **99**, 177–195 (1985).
33. Mormann, F. *et al.* Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. *J. Neurosci.* **28**, 8865–8872 (2008).
34. Quian Quiroga, R., Kraskov, A., Koch, C. & Fried, I. Explicit encoding of multimodal percepts by single neurons in the human brain. *Curr. Biol.* **19**, 1308–1313 (2009).
35. Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I. & Koch, C. Sparse representation in the human medial temporal lobe. *J. Neurosci.* **26**, 10232–10234 (2006).
36. Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R. & Fried, I. Internally generated reactivation of single neurons in human hippocampus during free recall. *Science* **322**, 96–101 (2008).
37. Kreiman, G., Koch, C. & Fried, I. Imagery neurons in the human brain. *Nature* **408**, 357–361 (2000).
38. Macoveanu, J., Klingberg, T. & Tegnér, J. A biophysical model of multiple-item working memory: a computational and neuroimaging study. *Neuroscience* **141**, 1611–1618 (2006).
39. Camperi, M. & Wang, X.J. A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. *J. Comput. Neurosci.* **5**, 383–405 (1998).
40. Laing, C.R., Troy, W.C., Gutkin, B. & Ermentrout, G.B. Multiple bumps in a neuronal model of working memory. *SIAM J. Appl. Math.* **63**, 62–97 (2002).
41. Lundqvist, M. *et al.* Gamma and beta bursts underlie working memory. *Neuron* **90**, 152–164 (2016).
42. Stokes, M.G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
43. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
44. van Vugt, M.K., Schulze-Bonhage, A., Litt, B., Brandt, A. & Kahana, M.J. Hippocampal gamma oscillations increase with memory load. *J. Neurosci.* **30**, 2694–2699 (2010).
45. Davis, M. & Whalen, P.J. The amygdala: vigilance and emotion. *Mol. Psychiatry* **6**, 13–34 (2001).
46. Buchanan, T.W., Tranel, D. & Adolphs, R. in *The Human Amygdala* (eds. Whalen, P.J. & Phelps, E.A.) 289–317 (2009).
47. Peinado-Manzano, M.A. The role of the amygdala and the hippocampus in working memory for spatial and non-spatial information. *Behav. Brain Res.* **38**, 117–134 (1990).
48. Saez, A., Rigotti, M., Ostojic, S., Fusi, S. & Salzman, C.D. Abstract context representations in primate amygdala and prefrontal cortex. *Neuron* **87**, 869–881 (2015).
49. Bush, G., Luu, P. & Posner, M.I. Cognitive and emotional influences in anterior cingulate cortex. *Trends Cogn. Sci.* **4**, 215–222 (2000).
50. Gläscher, J. *et al.* Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **109**, 14681–14686 (2012).

ONLINE METHODS

Task. We used a modified Sternberg task with images (instead of the usual digits) as material for memorization (**Fig. 1a**). Each trial started with a fixation cross shown for 900–1,000 ms. Next, we sequentially presented the images to be memorized ('encoding') in a given trial. Each picture was presented for 1 s, followed by a blank screen for 1–200 ms (randomized). Subjects were asked to memorize the 1–3 images shown in each trial. We use the terms 'encoding 1', 'encoding 2' and 'encoding 3' to refer to the 1–3 images shown in a trial. After encoding, there was a maintenance (delay) period lasting at least 2.5 s and at most 2.8 s. During this time, the word "hold" was shown on the screen. Lastly, after the end of the maintenance period, a probe stimulus was displayed. Subjects were asked to decide whether the probe stimulus shown was one of the preceding 1–3 images or not. Participants responded by pressing the green or red buttons on a response pad. Which color corresponded to 'yes' and which to 'no' was shown at the top of the screen during each probe trial. We used this approach to switch the location of the yes and no buttons in the middle of the experiment as a control. We asked subjects to respond as fast as possible. The probe picture was presented until subjects made a response. In each session, subjects performed 108 or 135 trials depending on the task variant. Pictures were shown in pseudorandom order.

The pictures used for each participant were different and were determined by the results of a screening task conducted 2–3 h earlier (except for one subject who did not undergo a screening procedure and who was shown images used with a previous subject). For the screening task, images were chosen based on a subject's interests.

During screening, we showed 54–64 images. Each image was shown six times in randomized order for 1 s. As a control, every few trials (randomized) we asked a control question related to the image shown immediately before (i.e., "Did the last image present a person/landscape/animal?"). After the screening task, we immediately analyzed the data to choose the five images with the best responses as judged by the mean response in a 200- to 1,000-ms window relative to stimulus onset (based on an F-statistic computed by a one-way ANOVA with image as a factor). In sessions in which less than five neurons showed significantly selective responses (6 of 13 sessions), we picked the remaining images according to the strongest nonselective responses. In these instances, we used a combination of the F-statistic (with image as a factor) and neuronal isolation quality (amplitude of waveform) to choose the next best nonselective response. These five images were subsequently used for the Sternberg task. Both tasks were implemented in Matlab using the Psychophysics Toolbox[51]. Note that the statistical tests run during the screening and Sternberg task were statistically independent, because they were run at distinct periods of time.

Data collection and analysis were not performed blind to the conditions of the experiments.

Subjects. Thirteen subjects participated in the study (**Supplementary Table 1**). All of them were implanted with depth electrodes for possible surgical treatment of epilepsy. They volunteered for the study and gave informed consent. This study was approved by the Institutional Review Boards of the Cedars-Sinai Medical Center, the Huntington Memorial Hospital and the California Institute of Technology. Electrodes were localized based on pre- and postoperative T1 structural MRIs. We used the following processing pipeline to transform the postoperative MRI into the same space as a template brain. We extracted the brains from the pre- and postoperative T1 scans[52] and aligned the postoperative to the preoperative scan with Freesurfer's mri_robust_register[53]. We then computed a forward mapping of the preoperative scan to the CIT168 template brain[54] using a concatenation of an affine transformation followed by a symmetric image normalization (SyN) diffeomorphic transform, computed by the ANTs software package[55]. This resulted in a postoperative scan overlaid on a version of version of the CIT168 template brain[54] registered in MNI152 space. We then used Freesurfer's Freeview program to mark the electrodes as point sets to determine where the tips of the microwires were located. For visualization only, electrode locations were projected onto the 2D sagittal plane (**Fig. 1b**).

Spike sorting and quality metrics of single units. Each macroelectrode contained eight 40-μm diameter microwires[56]. We recorded broadband (0.1–9,000 Hz filter) data from 64 channels sampled at 32 kHz using a Neuralynx Atlas system. Signals were locally referenced to one of the eight microwires in each brain area.

The raw signal was filtered with a zero-phase lag filter in the 300- to 3,000-Hz band, and spikes were detected and sorted using a semiautomated template-matching algorithm[23]. We computed several spike sorting quality metrics for all identified putative single units to assess the quality of identified units (**Supplementary Fig. 2**): (i) the percentage of interspike intervals (ISIs) below 3 ms was 0.42% ± 0.68%; (ii) the ratio between the s.d. of the noise and the peak amplitude of the mean waveform of each cluster was 6.18 ± 4.09 (peak SNR); (iii) the pairwise projection distance in clustering space between all neurons isolated on the same wire was 14.39 ± 6.2 (projection test; in units of s.d. of the signal)[57]; (iv) the modified coefficient of variation of variability in the ISI (CV2) was 0.94 ± 0.14; and (v) the median isolation distance[58] was 30.9. We calculated the isolation distance in a ten-dimensional feature space (energy, peak amplitude, total area under the waveform and the first five principal components of the energy normalizes waveforms[58]).

Statistics. Statistical comparisons were conducted using permutation tests based on a null distribution estimated from $B = 2{,}000$ runs on data with scrambled labels using the EEGLAB toolbox[59]. For comparisons with two groups, we used the permuted t-test statistic and for comparisons with more than two groups, we used F-statistics. Note that we used permutation tests throughout to avoid the assumption of normality. We used ANOVAs instead of linear regressions to test for effects of load because linear regressions introduce the additional assumption of a monotonic relationship between load and behavior. Note that the reported P-values may differ from those expected from the t- and F-distributions because P-values were based on the empirically estimated null distribution. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications in the field[56]. A **Supplementary Methods Checklist** is available.

Selection of neurons. For each recorded neuron, we ran three statistical tests to determine whether a cell qualified as a concept, maintenance or probe neuron. Tests were independent and some cells qualified as multiple types. To identify concept cells we counted spikes in a window 200–1,000 ms following stimulus onset of the first encoding period. Concept cells were selected ($P < 0.05$) using a permuted one-way ANOVA with five groups (the number of unique images). In addition, we required that the activity for the image with the maximal response be significantly larger than that of all other images ($P < 0.05$, permutation t-test). The value of this latter test was also used as the effect-size metric to compare the neuronal groups (see "Separability of neuronal categories" section, above). For the analysis of persistent activity of concept cells, we used all concept cells found during the Sternberg task regardless of whether they were also significant in the screening task.

To identify maintenance cells, we tested whether the mean firing rate during the maintenance period (0–2,500 ms) was significantly larger (permutation t-test, $P < 0.05$) than the firing rate during the 500-ms long fixation-cross period before trial start. If a maintenance cell was also a concept cell, we additionally required that its maintenance activity for the nonpreferred stimuli be significantly higher than baseline. This second criteria was to ensure that strong concept cells, which only increase their firing rate to a single stimulus, didn't automatically qualify as maintenance cells.

Probe neurons were identified by comparing the firing rate during presentation of the probe stimulus (spikes counted during the period 200–1,000 ms after probe onset) to that during both the encoding ($P < 0.05$, 200–1,000 ms) and maintenance periods ($P < 0.05$, 0–2,500 ms, permutation t-test).

Single-cell metrics. We used a Picture Selectivity Index (PSI) to quantify the response of concept cells:

$$\mathrm{PSI} = \frac{FR_{\text{preferred image}} - FR_{\text{nonpreferred image}}}{FR_{\text{baseline}}} \times 100$$

Where FR is the mean firing rate in a 200- to 1,000-ms window relative to stimulus onset for encoding and probe, 0 to 2,500-ms window for maintenance and −500- to 0-ms window for baseline. When comparing correct and incorrect trials, we computed firing rates and PSI values independently for the three load conditions (1, 2 and 3) and then averaged the results to eliminate bias resulting from the different numbers of incorrect trials for each load. We also used the depth of selectivity index[6] to quantify the sparsity of concept cells (**Supplementary Fig. 1**):

$$S = \frac{n - \left(\frac{\sum R_i}{R_{\max}}\right)}{n - 1}$$

Where n is the number of images presented, R_i is the firing rate of the neuron during the presentation of the ith picture, and $R_{\max}$ is the largest firing rate across all presented images. A neuron with $S = 0$ would respond identically to all images, whereas a neuron with $S = 1$ would respond only to a single image and not at all to others.

To compare time courses of neuronal activity, we counted spikes in 200-ms bins with a step size of 2 ms. We corrected for multiple comparisons using a cluster-based approach[60]. In this method, a cluster is defined as a group of adjacent significant tests. We tested, for all identified clusters, whether the summed value of the test statistic in a given cluster was larger than the 95th percentile of the same value estimated from scrambled labels. For comparing firing rates across cells, we standardized the firing rate of each neuron using the mean and s.d. of the firing rate of each cell during the baseline (fixation cross, −500-ms to 0-ms window before encoding 1). To evaluate whether the proportion of significant neurons in an area was larger than that expected by chance, we computed a null distribution based on randomly scrambled labels ($B = 500$) and then estimated an empirical P-value as described.

Population decoding. We used a pseudopopulation of neurons pooled across all recording sessions and subjects to determine decoding performance. To pool images from different subjects, we arbitrarily labeled the individual images of each subjects as Images A–E. We then pooled the images with the same label across all subjects. Note that each subject saw a different set of pictures. Nevertheless, neurons can be considered a pseudopopulation for decoding purposes, a procedure that is frequently used to pool single-neuron recordings across sessions and animals[61]. This is because for the learning algorithm, all that matters is that each subject saw five distinct images regardless of their identity. The decoder assigns a weight to each neuron independently, i.e., the decoder identified all neurons that signaled the presence of Image A in all subjects, regardless of the identity of Image A. This pooling procedure provides, at every point of time, an $N \times T$ matrix, where N is the total number of neurons recorded from all subjects in given brain area and T is the smallest number of trials of a given type observed in all subjects. This matrix thus represented, as a function of time, the neural state in an N-dimensional space. We then trained a decoder to separate patterns in this high-dimensional space. Spikes were counted in 1,500-ms bins moved with a step size of 250 ms. For decoding, we used a support vector machine (SVM) as implemented in the ndt toolbox[61] and the LIBSVM library[62]. We used leave-one-out cross-validation to estimate performance: one trial was randomly assigned for each neuron from each class as the test trial and the remaining trials were used for training. All possible train–test combinations were computed, and we used 500 randomly chosen train–test combinations to estimate the cross-validated testing error. For decoding image identity, we trained a classifier on the load 1 condition and then applied it to loads 2 and 3 to determine the accuracy by which the algorithm could determine whether a given image was in held in memory. To test whether any given decoding performance was significantly better than chance, we created null distributions by decoding using the same approach as described above but with scrambled labels. We repeated this computation 500 times. Note that thus the minimal possible P-value was 0.002, and we used this P-value if no value in the null distribution exceeded the observed value. Similarly, for comparisons between decoding accuracy of different decoders—for example, all neurons compared to maintenance-only neurons—we created a null distribution of differences in a given comparison by subtracting the null distribution created for the whole population (in this example) from the null distribution created for maintenance neurons only. We then estimated the P-value for this comparison by counting the number of times the null distribution exceeded the observed difference of performance between the two decoders.

Analysis of neuronal activity dynamics. To analyze the dynamics of activity of the entire neuronal population, we reduced dimensionally using demixed principal component analysis (dPCA)[29]. We used dPCA instead of PCA because PCA stretches components 'blindly', based on the percentage of explained variance as a criterion. As a result, the dimensions chosen often lack a meaningful interpretation and thus are not helpful in a particular analysis. In contrast, dPCA has an advantage in that it stretches components along dimensions, not only to explain overall signal variance but also to explain variance attributable to variables of interest (such as image identity in our case). For dPCA analysis, we used a pseudopopulation of all recorded neurons (as defined above). We used dPCA with picture identity and time as the marginalized variable. We binned neuronal firing in 2-ms non-overlapping bins and smoothed the resulting time course by convolution with a Gaussian kernel (200 ms wide). In addition, we z-scored all time-courses based on the mean and s.d. estimated from the baseline (−500 to 0 ms relative to onset of first image). We computed the basis functions of dPCA (demixing weights) based only on the data recorded 200–1,000 ms following onset of the first image (encoding 1). We rank-ordered the dPCs by their explained variance and used the first 15 dPCs. These together accounted for 51.58% of the total signal variance. To prevent overfitting, we used a regularization procedure to find the optimal lambda parameter (**Supplementary Fig. 6f**). In addition, we tested whether the percentage of explained variance for picture identity was higher than that observed for data with scrambled image identity labels. Indeed, the portion of variance that was attributed to image identity by the 15 largest dPCs of the real data was 45.13%, whereas for the scrambled data it was only 13.64% (permutation test, $P = 0.002$; **Supplementary Fig. 6e**). We projected the data from the maintenance period onto the basis functions computed from the encoding 1 period.

We used the multidimensional Euclidean distance $d(p,q)$ to quantify how different the population activity was between the neural activity vector $p(t)$ and $q(t)$ (which here are the neuronal states during two different conditions at time t):

$$d(p(t), q(t)) = \sqrt{\sum_{i=1}^{n} (p_i(\mathrm{t}) - q_i(t))^2}$$

In addition, we quantified the speed by which the population activity changed at a given point of time t as $V(t)$:

$$V(t) = \frac{1}{n} \times \frac{\sqrt{\sum_{i=1}^{n} (p_i(\mathrm{t}) - p_i(t - \Delta t))^2}}{\Delta t}$$

Where n is the number of dimensions in dPCA space ($n = 4$), Δt is 50 ms and p is the neural activity vector. We used the first four dPCs (ids 1, 2, 3 and 5) that had the highest percentage of variance attributed to picture identity during load 1 trials (**Fig. 8a**,**b** and **Supplementary Figs. 5** and **6d**; see also **Supplementary Video 1**). In contrast, dPC 4 had variance that was only attributable to time but not to image identity (**Supplementary Fig. 5**). Additionally, the next biggest dPCs (id 6) accounted for only 2.55% of the variance attributed to picture identity. We also recomputed our analysis using the 8 or 12 dPCs with the highest percentage attributed to image identity and found that results were very similar to the analysis based on 4 dPCs (**Supplementary Fig. 6a**,**b**). Both V and d are population-level metrics not computable for single neurons. To estimate the variance of V and d, we bootstrapped the confidence intervals of V and d by randomly picking a subset of 10% of trials and computed V and d for each such subset (repeated 50 times).

For each possible remembered stimulus k, we defined the location of its corresponding attractor A_k as the center (mean across time) of the neuronal trajectory observed during maintenance of that image during the load 1 condition. There were five attractors (a–e), one for each image used. To quantify the distance of a neuronal trajectory from an attractor in a given condition, we used a DA metric:

$$DA(t) = \frac{d_i(t)}{\frac{1}{\mathrm{C} - 1} \sum_{j \neq i}^{C} d_j(t)}$$

Where $d_j(t) = d(A_j, p(t))$ is the Euclidean distance of the neuronal trajectory $p(t)$ from the attractor A_j. C is the total number of attractors ($C = 5$). We computed $DA(t)$ separately for loads 1–3 and for each time point t during maintenance (0–2,500 ms), and then averaged $DA(t)$ over all points of time and loads to get a single value DA for every trial. Note that $DA = 1$ implies that the neural trajectory was equidistant between the tested (d_i) and all other attractors, which indicates no memory. On the other hand, $DA < 1$ indicates that the neuronal state was closer to one attractor than to all the other attractors.

Like d and V, DA is a population-level metric not computable for single neurons. We thus assessed the significance of differences in DA between correct and incorrect trials and fast and slow RT conditions by estimating a null distribution of DA based on data with scrambled labels. For visualization of distances between attractors in 2D space (**Fig. 8e**), we used nonmetric multidimensional scaling (MDS). We used Euclidean distance as the pairwise distance measure. MDS was used for visualization only.

For the single-subject analysis (**Fig. 8h**), we used the same approach as described for the pseudopopulation but used only neurons recorded simultaneously from a given subject. For each session, we used the first four dPCs that explained the highest proportion of variance and computed DA accordingly.

Data availability. The data that support the findings of this study are available on reasonable request from the corresponding author. The data are not publicly available because they contain information that could compromise research participant privacy/consent.

Code availability. Analysis was performed in Matlab using the publicly available software packages Osort, ndt and EEGLAB together with custom-developed analysis routines.

51. Brainard, D.H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
52. Ségonne, F. *et al.* A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**, 1060–1075 (2004).
53. Reuter, M., Rosas, H.D. & Fischl, B. Highly accurate inverse consistent registration: a robust approach. *Neuroimage* **53**, 1181–1196 (2010).
54. Tyszka, J.M. & Pauli, W.M. *In vivo* delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. *Hum. Brain Mapp.* **37**, 3979–3998 (2016).
55. Avants, B. *et al.* Multivariate analysis of structural and diffusion imaging in traumatic brain injury. *Acad. Radiol.* **15**, 1360–1375 (2008).
56. Rutishauser, U., Ross, I.B., Mamelak, A.N. & Schuman, E.M. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature* **464**, 903–907 (2010).
57. Pouzat, C., Mazor, O. & Laurent, G. Using noise signature to optimize spike-sorting and to assess neuronal classification quality. *J. Neurosci. Methods* **122**, 43–57 (2002).
58. Harris, K.D., Henze, D.A., Csicsvari, J., Hirase, H. & Buzsáki, G. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* **84**, 401–414 (2000).
59. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
60. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
61. Meyers, E.M. The neural decoding toolbox. *Front. Neuroinform.* **7**, 8 (2013).
62. Fan, R.-E., Chen, P.-H. & Lin, C.-J. Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.* **6**, 1889–1918 (2005).

 doi:10.1038/nn.4509

Erratum: Persistently active neurons in human medial frontal and medial temporal lobe support working memory

Jan Kamiński, Shannon Sullivan, Jeffrey M Chung, Ian B Ross, Adam N Mamelak & Ueli Rutishauser

***Nat. Neurosci.* 20, 590–601 (2017); published online 20 February 2017; corrected after print 4 May 2017**

In the version of this article initially published, the pink and gray symbols were switched in the key to Figure 7d and the plot of the outlier data points was mis-scaled relative to the axes. In Figure 6d, the *x* axis was numbered 0 through 1 instead of –1 through 0. The errors have been corrected in the HTML and PDF versions of the article.