nature
neuroscience

Representation of retrieval confidence by single neurons in the human medial temporal lobe

Ueli Rutishauser[1–5], Shengxuan Ye[1,4], Matthieu Koroma[1,6], Oana Tudusciuc[4,7], Ian B Ross[8], Jeffrey M Chung[2] & Adam N Mamelak[1]

Memory-based decisions are often accompanied by an assessment of choice certainty, but the mechanisms of such confidence judgments remain unknown. We studied the response of 1,065 individual neurons in the human hippocampus and amygdala while neurosurgical patients made memory retrieval decisions together with a confidence judgment. Combining behavioral, neuronal and computational analysis, we identified a population of memory-selective (MS) neurons whose activity signaled stimulus familiarity and confidence, as assessed by subjective report. In contrast, the activity of visually selective (VS) neurons was not sensitive to memory strength. The groups further differed in response latency, tuning and extracellular waveforms. The information provided by MS neurons was sufficient for a race model to decide stimulus familiarity and retrieval confidence. Together, our results indicate a trial-by-trial relationship between a specific group of neurons and declared memory strength in humans. We suggest that VS and MS neurons are a substrate for declarative memories.

Decisions are often accompanied by an assessment of how likely it is that a choice will be correct. Such confidence judgments are critical in complex environments in which decisions need to incorporate future, not yet observed outcomes based on previous actions, information and outcomes. Determining whether a stimulus is novel or familiar is a complex decision involving the comparison of sensory information with internal variables. Although the outcome is binary (familiar or not), such memory retrieval decisions in humans are typically accompanied by graded judgments of confidence. Such confidence judgments feel automatic and are often accurate[1–3]. Despite its ubiquity, the mechanism of confidence judgments about memories is not understood. One model proposes that confidence judgments require separate specialized processes that evaluate decisions after they have been made, thereby drawing on metacognitive abilities that may be unique to humans[4]. In contrast, other models propose that an assessment of uncertainty is an integral and necessary part of any decision-making process itself[5]. Confidence can therefore be assessed simultaneously and by the same process that makes the decision in the first place, a core concept of Bayesian models of decision-making[6]. Although recent studies in non-human primates and rodents have provided evidence for the latter model during perceptual decisions[3,7], nothing is known thus far about how confidence judgments for memories are made. It has proven challenging to develop procedures for animals to communicate an assessment of confidence in an experimental setting, a problem that is particularly acute for memories. We took advantage of the availability of human neurosurgical patients for single-unit recordings to study this question.

The medial temporal lobe (MTL) is required to make declarative memory-based decisions[8], and populations of neurons in the MTL whose interaction is thought to underlie this ability have been identified. For example, the response of some neurons in the primate MTL is selective for visual categories or concepts[9–12]. Others signal whether a stimulus is novel or familiar[13–16], a response that can emerge after a single exposure[13,14]. Such MS neurons represent a potential substrate for episodic memories by marking stimuli as either novel or familiar. If so, we hypothesize that their activity should correlate with memory strength and with confidence. In contrast, neurons not directly involved in memory retrieval, such as those representing visual features, should not correlate with memory strength.

We used subjective confidence ratings made by subjects during a memory recognition task to identify groups of neurons that signaled memory strength. Our results make two key contributions. First, we found that MS and VS neurons code orthogonal pieces of information about visual stimuli. Second, only the activity of MS neurons correlated trial-by-trial with memory strength. In contrast, the ability of VS neurons to differentiate different stimuli was not sensitive to memory strength.

RESULTS

Task and behavior

Subjects (44 sessions from 28 patients; **Supplementary Table 1**) performed a recognition memory test during which they rated 100 images as seen before or not[17]. 50 of the images were familiar (shown

[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 Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, California, USA. [4]Computation & Neural Systems Program, California Institute of Technology, Pasadena, California, USA. [5]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. [6]Departement de Biologie, Ecole Normale Superieure de Cachan, Cachan, France. [7]Division of Humanities and Social Science, California Institute of Technology, Pasadena, California, USA. [8]Department of Neurosurgery, Huntington Memorial Hospital, Pasadena, California, USA. Correspondence should be addressed to U.R. (ueli.rutishauser@cshs.org).

Received 1 April; accepted 13 May; published online 8 June 2015; doi:10.1038/nn.4041

~30 min before the task during a separate learning session) and the other 50 images were novel (stimulus type, familiar or novel). Images were presented for 1 s each and, after a short delay, subjects were asked to indicate whether they had seen the image before (binary decision, new or old) together with a judgment of confidence in their decision (**Fig. 1a**). Each image belonged to one of five visual categories (cars, foods, people, landscapes or animals; Online Methods).

Subjects correctly identified 69 ± 13% of familiar stimuli and reported 28 ± 17% of novel stimuli as false positives (**Fig. 1b**). Confidence ratings were systematically related to accuracy (Goodman-Kruskal gamma correlation, g = 0.36 ± 0.37, t test versus chance $P < 10^{-6}$): the higher the confidence, the better the accuracy (**Fig. 1c–g**). We computed a receiver operating characteristic (ROC) curve[18] for each session to quantify the relationship between accuracy and confidence (**Fig. 1c**). The average area under the curve (AUC) of the ROC was 0.75 ± 0.08 (**Fig. 1c,d**). Different confidence ratings resulted in performance located in different locations in ROC space (**Fig. 1c**). The ROC was asymmetric (z-ROC slope = 0.78 ± 0.33, significantly less than 1, $P < 10^{-18}$; **Fig. 1e**), as expected for declarative memories[19]. Subjects performed above chance at all levels of confidence and the majority of decisions were made with high confidence (**Fig. 1f,g**). Subjects assigned medium and low confidences more rarely and with approximately equal likelihood (**Fig. 1f**). For a balanced statistical comparison between confidence levels with approximately equal trial numbers, we used two levels of confidence for the neuronal analysis: high and low. Trials with intermediate ratings were re-assigned a high or low confidence rating depending on the proportion of trials (irrespective of performance) made with medium confidence (Online Methods). The resulting two confidence ratings were associated with different retrieval accuracy (**Fig. 1h**).

The decision time (DT, time from question onset till response) varied systematically as a function of confidence and accuracy (repeated-measure ANOVA model, **Fig. 1i–l** and Online Methods). Correct high-confidence decisions were faster than low-confidence decisions (1.54 ± 0.11 s versus 2.49 ± 0.20 s, main effect of confidence $F_{1,30} = 25.74$, $P < 10^{-4}$; **Fig. 1i**). Correct familiar decisions were faster than correct novel decisions regardless of confidence (**Fig. 1i**). This was also true for incorrect trials: high-confidence incorrect decisions were faster than low-confidence incorrect decisions (**Fig. 1j**). Correct decisions were made with higher confidence than incorrect decisions (1.95 ± 0.06 s versus 1.65 ± 0.05 s; main effect of correctness, $F_{1,41} = 58.3$, $P < 10^{-8}$; Wilcoxon signed-rank test correct versus incorrect, $P < 2.74 \times 10^{-9}$; **Fig. 1k**). Also, correct decisions were made quicker than incorrect decisions for familiar stimuli (1.41 ± 0.13 s versus 1.78 ± 0.14 s, significant interaction $F_{1,30} = 8.51$, $P < 0.05$, $n = 31$ subjects). Because incorrect decisions were made more slowly and with lower confidence, we matched the average confidence in correct and incorrect trials. We found that correct decisions were made faster even after matching confidence (incorrect versus correct: 1.89 ± 0.15 s versus 2.21 ± 0.19 s, Wilcoxon signed-rank test after matching for confidence $P < 0.01$; **Fig. 1l**). Together, these findings show that subjects accurately assessed the quality of their memories

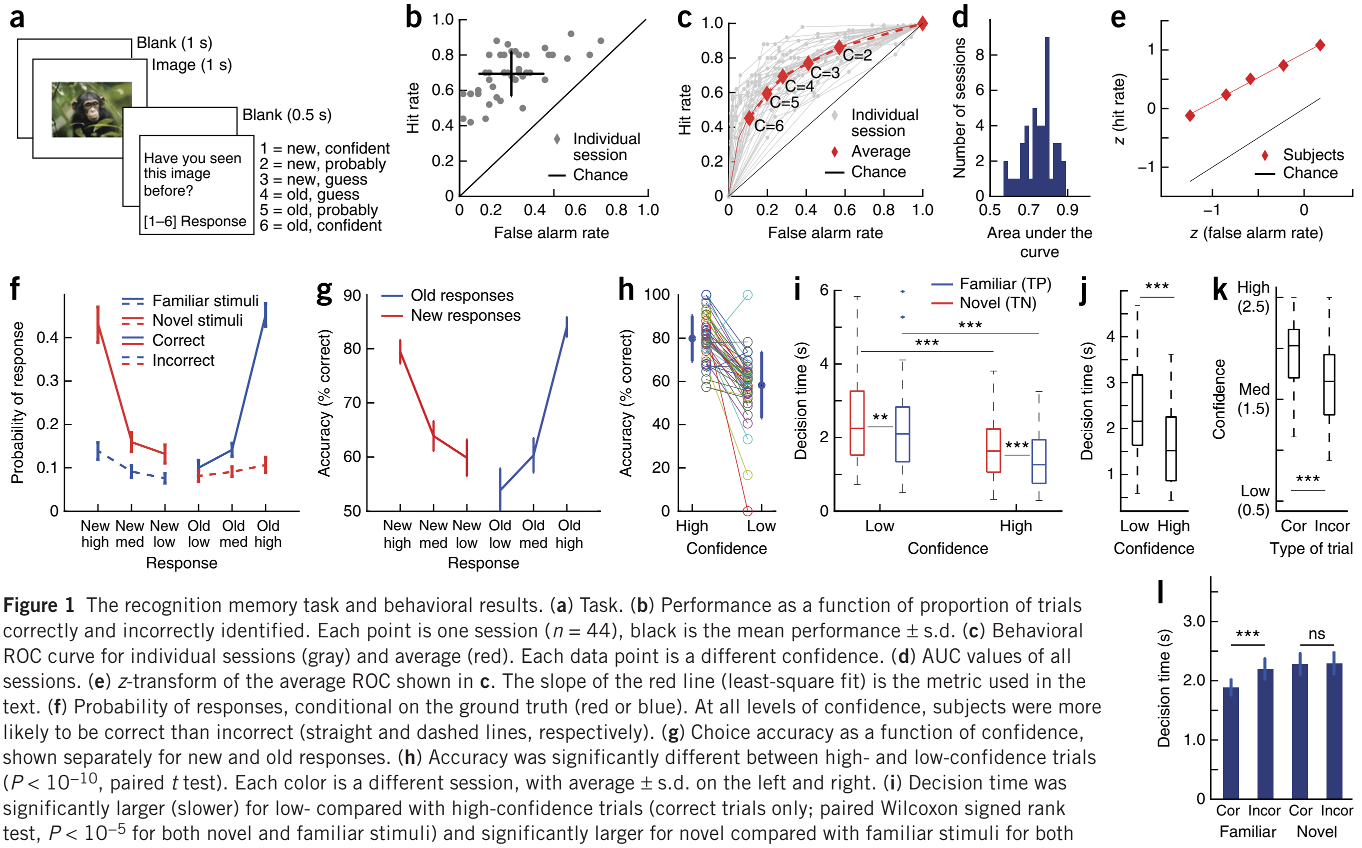

Figure 1 The recognition memory task and behavioral results. (**a**) Task. (**b**) Performance as a function of proportion of trials correctly and incorrectly identified. Each point is one session (n = 44), black is the mean performance ± s.d. (**c**) Behavioral ROC curve for individual sessions (gray) and average (red). Each data point is a different confidence. (**d**) AUC values of all sessions. (**e**) z-transform of the average ROC shown in **c**. The slope of the red line (least-square fit) is the metric used in the text. (**f**) Probability of responses, conditional on the ground truth (red or blue). At all levels of confidence, subjects were more likely to be correct than incorrect (straight and dashed lines, respectively). (**g**) Choice accuracy as a function of confidence, shown separately for new and old responses. (**h**) Accuracy was significantly different between high- and low-confidence trials ($P < 10^{-10}$, paired t test). Each color is a different session, with average ± s.d. on the left and right. (**i**) Decision time was significantly larger (slower) for low- compared with high-confidence trials (correct trials only; paired Wilcoxon signed rank test, $P < 10^{-5}$ for both novel and familiar stimuli) and significantly larger for novel compared with familiar stimuli for both low and high confidences (P = 0.01 and $P < 10^{-4}$, respectively). (**j**) Decision time was significantly slower for low- compared with high-confidence incorrect trials (paired Wilcoxon singed rank test, $P < 10^{-6}$). (**k**) Errors were made with less confidence than correct trials ($P < 10^{-8}$, paired Wilcoxon signed rank test). (**l**) Correct familiar decisions were made faster than incorrect decisions (paired comparison matched for confidence, Online Methods). (**i**–**k**) Boxplots represent quartiles (25%, 75%), line is median, whiskers show range up to 1.5 times the interquartile range, and dots above whiskers show outliers. $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$. P values are uncorrected for multiple comparisons. ns indicates non-significant (P = 0.60). Error bars in **f**, **g** and **l** represent ± s.e.m. across subjects.

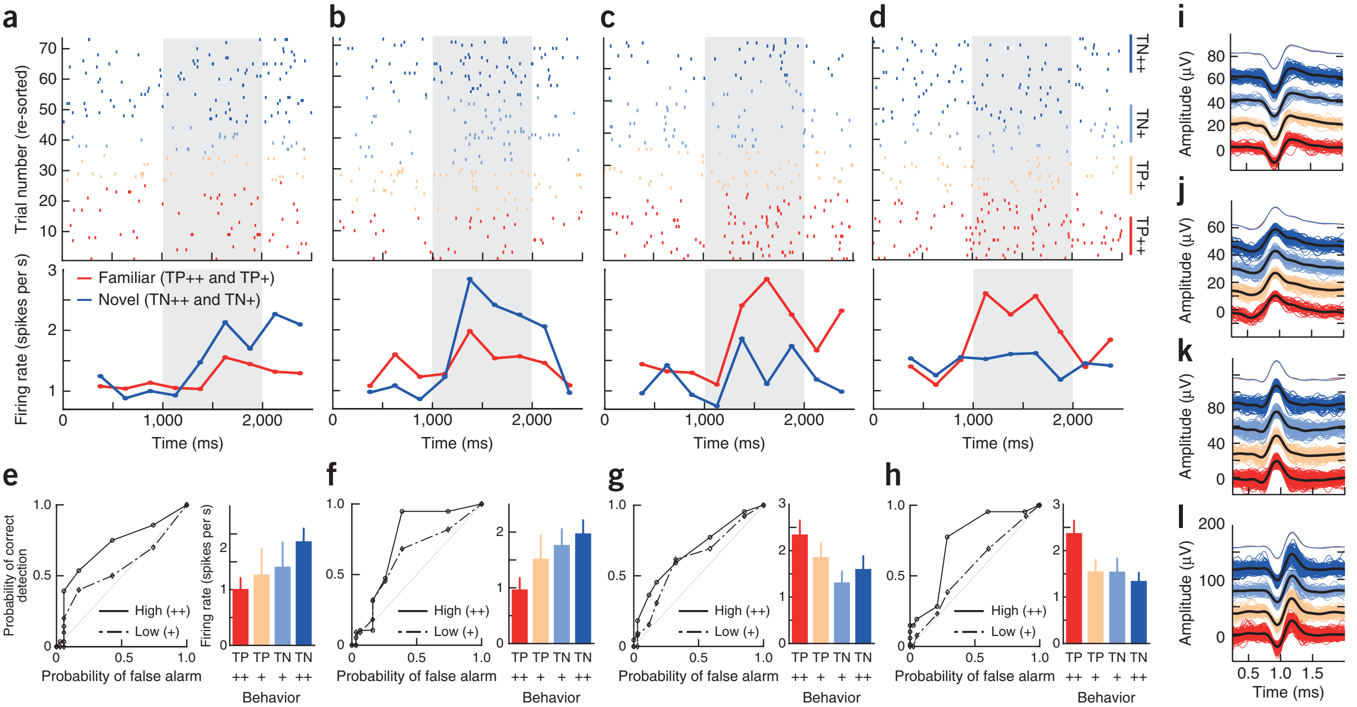

Figure 2 MS neurons. (**a**–**d**) Raster (top) and post-stimulus time histogram (PSTH) (bottom) of four example neurons, NS (**a**,**b**) and FS (**c**,**d**). Stimulus onset was at 1,000 ms (gray). Trials were re-sorted by behavior for display purposes: familiar high confidence (TP++), familiar low confidence (TP+), novel low confidence (TN+) and novel high confidence (TN++). Error trials are not shown. In the PSTH, trials were grouped according to TP and TN. (**e**–**h**) Single-neuron ROC curves (left) and mean rate (right) for the neurons shown in **a**–**d**. Bar plots show the mean rate in a 1.5-s window starting 200 ms after stimulus onset. Error bars are ± s.e.m. across trials. (**i**–**l**) Waveforms of spikes associated with the four different trial types for each neuron, in same order as in **a**–**d**. Top, mean waveforms superimposed. Bottom, all individual waveforms associated with the spikes shown in **a**–**d**. Color code is identical to that used in **a**–**d**.

(**Fig. 1h**) and that the relationships between DT and confidence were as expected for declarative memory retrieval decisions[1].

We selected subsets of sessions for analysis on the basis of behavioral metrics alone. Two groups were selected: group 1 (patients with above chance retrieval performance, n = 38 sessions, AUC = 0.81 ± 0.10, g = 0.39 ± 0.29) and group 2 (patients who were able to distinguish between high-and low-confidence memories, 26 sessions, AUC = 0.84 ± 0.08, g = 0.38 ± 0.27).

Electrophysiology

We isolated 1,065 putative single units from the amygdala and hippocampus in 44 sessions (on average 24 per session). Units were carefully isolated[17,20] and recording and spike sorting quality were assessed quantitatively (**Supplementary Fig. 1**). The average firing rate was 1.84 ± 2.66 Hz (**Supplementary Table 2**). Throughout the manuscript, we use the term neuron to refer to a putative single unit. Neurons were sensitive to the onset of visual stimuli as expected[9,17]: 30% (321/1,065) of the neurons responded when comparing baseline with post-stimulus periods ($P < 0.05$, two-tailed t test, 1 s each). Note that the analysis that follows was not restricted to visually responsive neurons.

Single-neuron signatures of memory

We first tested whether the neuronal response following stimulus onset depended on whether the stimulus was novel (not seen before) or familiar (seen before) stimuli. We found that the response of 8.5% (81 of 954, $P < 10^{-5}$, Bernoulli; correct trials only in Group 1, n = 38 sessions, **Supplementary Table 2** and **Supplementary Fig. 2**) of all neurons differed between novel and familiar stimuli. This was true for both amygdala (43/577, 7.5%) and hippocampal (38/377, 10.1%) neurons. We refer to these neurons as being MS[13]. Similar to previous experiments[13,14,21], we identified two types of MS neurons (**Fig. 2**). The first had a higher firing rate to novel than to familiar stimuli (45 of 81 neurons; **Fig. 2a**,**b**), whereas the second type had an increased firing rate for familiar than novel stimuli (36 of 81 neurons; **Fig. 2c**,**d**). We refer to these neurons as novelty- and familiarity-selective (NS and FS) neurons, respectively[13].

We next performed a single-neuron ROC analysis for every MS neuron and calculated its AUC. The AUC specifies the probability by which an ideal observer could predict the choice (novel or familiar) of a subject by counting spikes in an individual trial. Note that some studies refer to this metric as choice probability[22]. Only MS neurons from patients that were able to differentiate high from low confidences were considered (group 2, 65 of 664 units (9.8%) were MS units; **Fig. 2e**–**h** and **Supplementary Fig. 4**). The average AUC for all MS neurons, considering all correct trials, was 0.64 ± 0.04 (different from chance by design, as the neurons were selected to be different in the first place; what is important here is only the magnitude). We next computed AUC values using only high- or low-confidence trials. Note that the selection of MS neurons does not consider confidence, making this comparison independent. AUC values were significantly larger for high- than for low-confidence trials for all MS neurons together (0.66 ± 0.007 versus 0.60 ± 0.010; **Fig. 3a**–**c**) and for NS and FS neurons separately (**Fig. 3d**,**e**). This was true for both hippocampal and amygdala neurons (**Supplementary Table 2**), for neurons recorded from the left and right hemisphere alone, and when evaluating the differences using a bootstrap rather than parametric statistics (**Supplementary Fig. 3**). These differences could not be attributed to different units that might have been merged into one single cluster: the mean

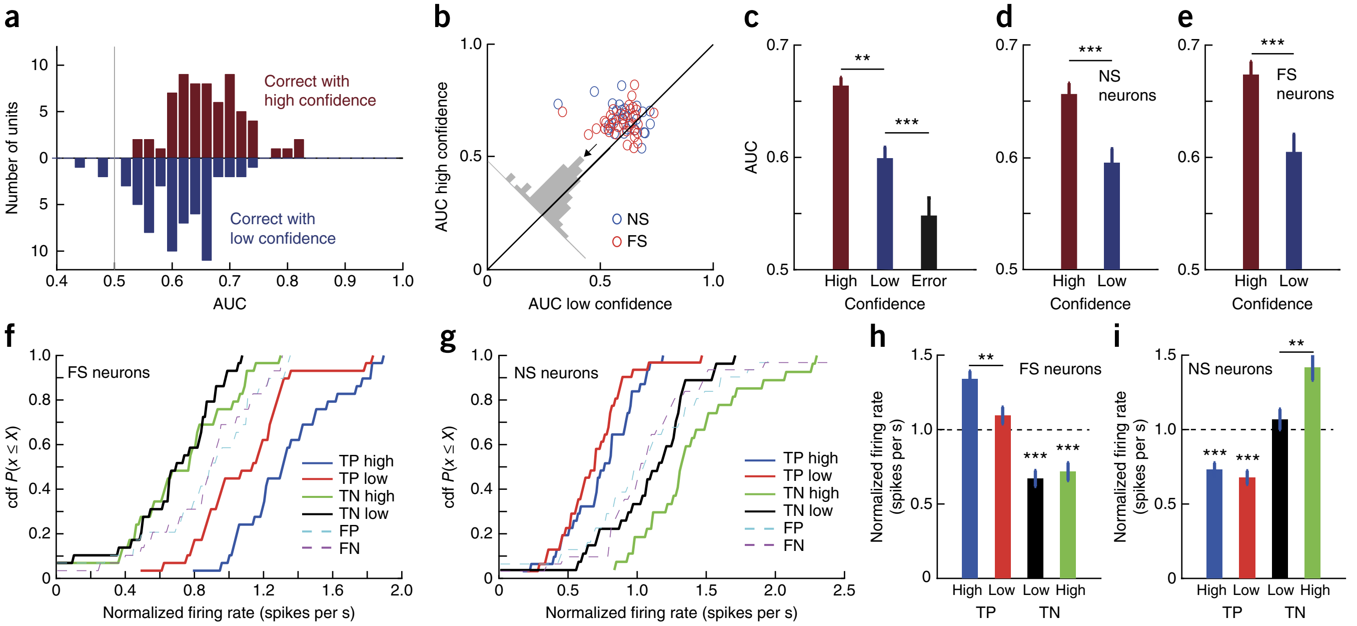

Figure 3 The response of MS neurons is modulated by subjective confidence. (**a**–**e**) Single-neuron ROC analysis. (**a**) AUC of MS neurons for high (red) and low (blue) confidence ($n = 65$ units; the two distributions were significantly different, $P = 0.001$). (**b**) Pairwise comparison of AUC values. For 49 of 65 units, the AUC was high > low ($P < 10^{-4}$, sign test). The average difference was above the diagonal (inset). (**c**) Average AUC for high- and low-confidence correct and error trials (FN; FN versus low, $P = 0.0056$; high versus low, $P < 10^{-5}$, pairwise t test). (**d**,**e**) AUC for high-confidence trials was significantly larger for both NS ($n = 29$) and FS ($n = 36$) neurons ($P = 0.0001$ and $P = 0.0003$, respectively). (**f**–**i**) Comparison of firing rate using baseline normalized responses and grouped by behavior. (**f**) Activity of FS neurons differentiated high- from low-confidence familiar trials ($n = 29$, TP high versus TP low, $P = 0.0094$, Kolmogorov-Smirnov test), but not novel trials ($n = 30$, TN high versus TN low, $P = 0.74$, Kolmogorov-Smirnov test). (**g**) Activity of NS neurons differentiated high- from low-confidence novel trials (TN high versus TN low, $P = 0.03$, Kolmogorov-Smirnov test), but not high from low familiar trials (TP high versus TP low, $P = 0.22$, Kolmogorov-Smirnov test). The cumulative distributions function (cdf) is shown in **f** and **g**. (**h**,**i**) Mean normalized response across neurons. (**h**) FS neurons had significantly higher firing rate for TP high compared with TP low trials (paired t test, $P = 0.0014$). (**i**) NS neurons had significantly higher firing rate for TN high compared with TN low trials (paired t test, $P = 0.0002$). *** indicates significant difference from baseline ($P < 10^{-4}$). True positive (TP) and negatives (TN) are correctly remembered familiar and novel stimuli. False positives (FP) and false negatives (FN) are wrongly identified novel and familiar stimuli. Errors are ± s.e.m. across neurons. **$P \leq 0.01$, ***$P \leq 0.001$.

waveforms associated with each of the four trial types were indistinguishable (**Fig. 2i**–**l**). Comparing forgotten (false negative, FN) trials with truly novel trials revealed an AUC larger than chance (0.55 ± 0.020, $P = 0.0048$ versus chance of 0.50; **Fig. 3c**), but significantly smaller than that for low-confidence correct decisions (0.60 ± 0.010, $P = 0.0056$). This indicates that MS neurons carry a memory signal that is strongest for high-confidence correct trials, intermediate for low-confidence trials and weakest for FN trials (**Fig. 3c**).

We performed a number of controls to exclude possible confounds. Using MS neurons from non-epileptic areas revealed a similar difference ($n = 40$, AUC = 0.66 ± 0.01 versus 0.61 ± 0.01, $P = 0.00066$), as did using only neurons in epileptic tissue (later resected, AUC = 0.67 ± 0.01 versus 0.61 ± 0.02, $P = 0.0041$). Equalizing the number of trials in the high- and low-confidence groups did not change the result (AUC = 0.67 ± 0.01 versus 0.60 ± 0.02, $P < 4 \times 10^{-5}$). Finally, randomly re-assigning confidences and keeping the novel and familiar labels intact abolished the high and low difference as expected (AUC = 0.65 ± 0.01 versus 0.65 ± 0.01, $P = 0.81$; **Supplementary Fig. 3m**–**o**).

We next compared the response patterns of FS and NS neurons. The previous ROC analysis was not sensitive to whether one or both terms constituting the difference are modulated. We therefore directly compared the normalized number of spikes fired by FS and NS neurons as a function of behavior. By design, FS and NS neurons responded maximally to familiar and novel stimuli, respectively (**Fig. 3f,g**). The response of FS and NS neurons significantly differed (FS neurons, $P = 0.0094$ and $P = 0.74$; NS neurons, $P = 0.22$ and $P = 0.03$; P values are for TP and TN trials, respectively) between high- and low-confidence trials, but only for the trial types to which the neurons increased their firing rate. Thus, the response of FS neurons differed between high- and low-confidence trials only for familiar stimuli and vice-versa for NS neurons (**Fig. 3f,g**). In addition, both FS and NS neurons decreased their firing rate to novel and familiar stimuli, respectively (**Fig. 3h,i**). The magnitude of this decrease, however, was insensitive to confidence. Thus, NS and FS neurons signal confidence asymmetrically because only the trial type to which they increase their firing rate relative to baseline is modulated by confidence. This conclusion relies on an absence of firing rate reduction below baseline, which is difficult to detect as a result of low baseline firing rates. However, note that this very problem would be faced by an imaginary downstream neuron receiving input from FS and NS neurons.

Single-neuron signatures of visual information

Each image that we showed to the subjects belonged to one of five investigator-defined visual categories (cars, foods, people, landscapes and animals). The response of 17.5% (186 of 1,065) of units was significantly modulated by category (one-way ANOVA, $P < 0.05$; **Fig. 4**), a proportion similar to what has been reported previously[9] (**Supplementary Table 2** and **Supplementary Fig. 2**). We refer to this group as VS neurons.

The two populations were independent: 15 of 186 VS neurons were also MS neurons (8%) and 15 of 87 (17%) of MS neurons were also VS neurons (χ^2 test of independence, $P = 0.91$; this also applies when considering only neurons from groups 1 and 2 and when excluding neurons with firing rates <1 Hz). A small group of neurons (15 of 1,065, 1.5%)

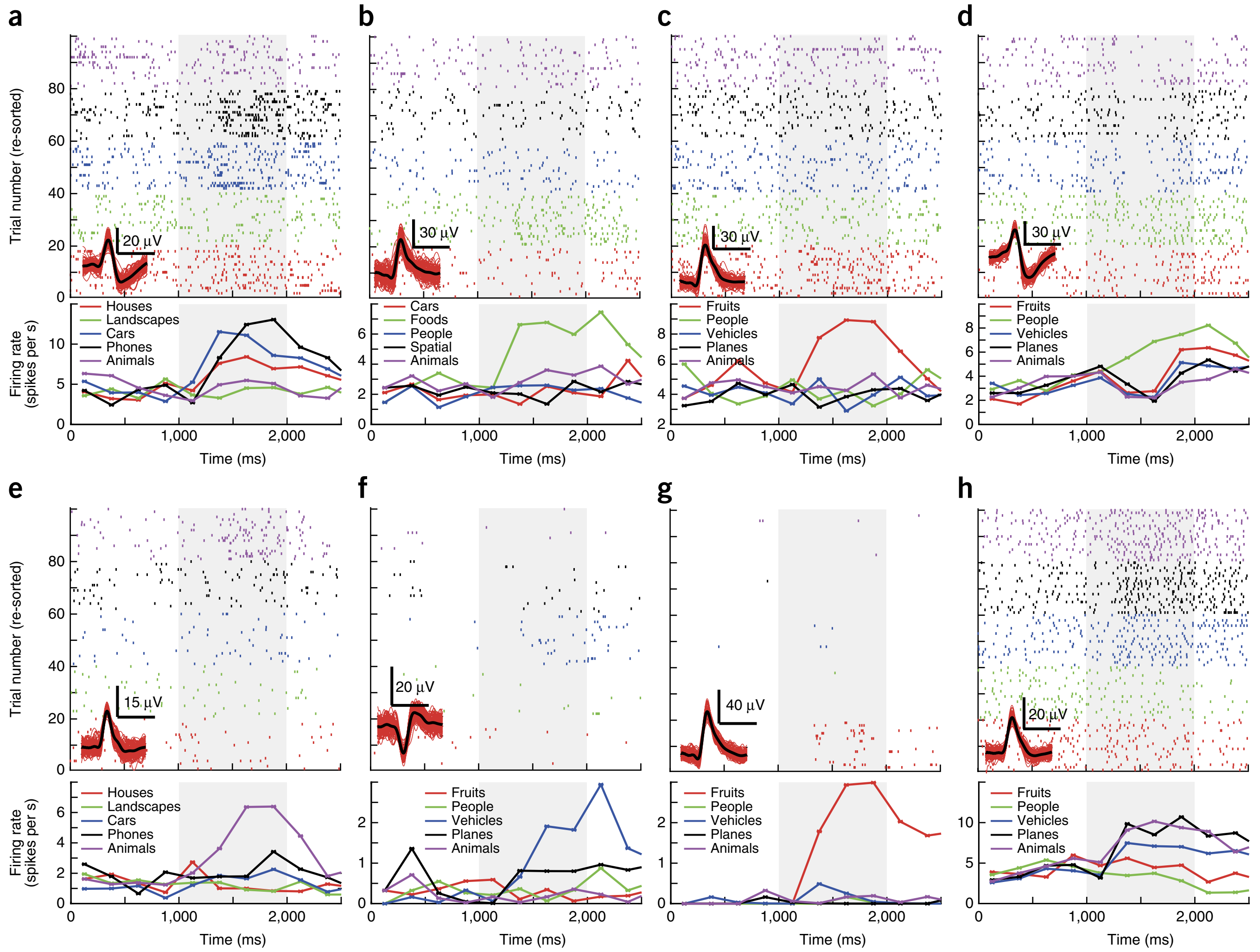

Figure 4 VS neurons. (**a**–**h**) For each, the raster (top) and PSTH (bottom) is shown. Trials are re-sorted for illustration purposes. Visual identity (category) is indicated by color, and the legends show the corresponding label (variable). The inset (bottom left of raster) shows waveforms associated with the indicated neuron (in red are 100 randomly chosen individual waveforms; in black, mean waveform; horizontal scale bar represents 1 ms). **a**,**b**,**d**,**f** and **c**,**e**,**g**,**h** are from the hippocampus and amygdala, respectively. All units are from different sessions. Some units responded with a firing increase only to one category (**b**,**c**,**e**–**g**), whereas others showed a mixed response (**a**,**d**,**h**). Stimulus onset was at 1,000 ms (gray). Significance of selection criteria (1 × 5 ANOVA) was 7×10^{-5} (**a**), 10^{-6} (**b**), 0.004 (**c**), 0.003 (**d**), 5×10^{-9} (**e**), 0.0004 (**f**), 3×10^{-12} (**g**) and 4×10^{-9} (**h**). PSTH bin size was 250 ms.

were both MS and VS cells (**Supplementary Fig. 4**), a proportion larger than would be expected by chance (chance level 0.25%, $P = 0.001$; **Supplementary Fig. 2**) and compatible with independence of memory and visual selectivity. We analyzed VS and MS neurons without excluding those that code for both.

Did the response of VS neurons depend on memory strength? To answer this question, we first identified the most and least preferred stimulus category for each VS neuron (for example, the neuron shown in **Fig. 4e** best differentiates between animals and houses). We then used single-neuron ROC analysis to quantify how well the response of each VS neuron discriminated between these two categories for four different trial types: novel, familiar, and high and low confidence. Using only correct trials from neurons in group 2 (128 of 664 were VS neurons; **Supplementary Table 2**) we found that AUC values did not differ as a function of confidence or familiarity (**Fig. 5**). The same conclusions held when excluding low-firing rate neurons (**Supplementary Fig. 5**). This indicates that the ability of a VS cell to identify its preferred category does not depend significantly on stimulus familiarity or confidence. This conclusion relies on the absence of a significant difference, which does not exclude the possibility that our data does not have enough statistical power to detect an existing difference. However, note that, using the same number of trials and time window, MS neurons showed a strong difference. In addition, the pairwise comparison between the two conditions (high and low, new and old) is based on trials for which the neuron carried information to begin with (the preferred category), assuring that the individual AUC values were well above chance.

VS neurons discriminate before MS neurons

We next estimated the first point of time at which the response of VS and MS neurons differed between different visual categories and novel and familiar stimuli, respectively. We compared the cumulative sum of the spike trains, a method that provides an estimate of the differential latency of a neuron with millisecond precision[15] (Online Methods). The average differential latencies of VS and MS neurons were 272 and 461 ms, respectively (relative to stimulus onset; **Fig. 6a**,**b**). Thus the response of MS neurons was delayed by 189 ms relative to VS neurons.

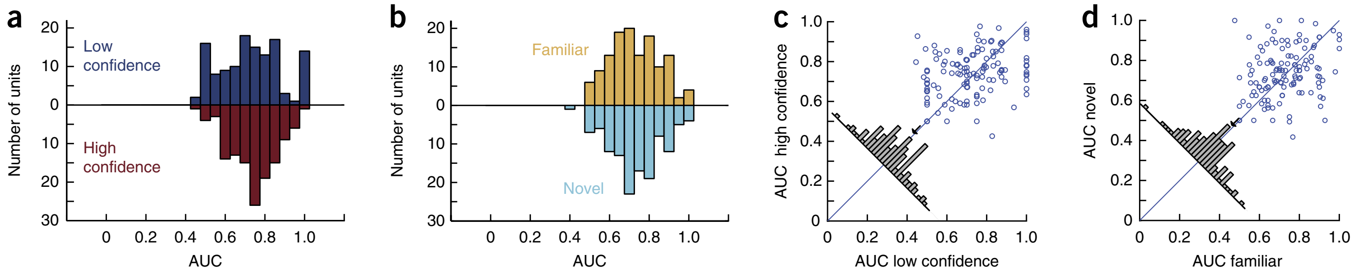

Figure 5 The ability of VS neurons to differentiate visual stimuli is not influenced by confidence judgment or novelty of the stimulus. (**a**) AUC of VS neurons for low- and high-confidence trials ($P = 0.31$, bootstrap test). (**b**) AUC of VS neurons for novel and familiar trials ($P = 0.54$, bootstrap test). (**c**) Pairwise comparison of AUC values as a function of confidence ($P = 0.53$, pairwise sign test). (**d**) Pairwise comparison of AUC values as a function of familiarity ($P = 0.41$, pairwise sign test). In **c** and **d**, every data point is one VS neuron (n = 128 in total). All pairwise comparisons showed no significant difference. Only correct trials were considered throughout.

Differential coding of visual category and memory

We next considered all recorded neurons together ($n = 664$, group 2). We fit a moving-window regression model for every single unit (using correct trials only) to estimate how much of the neuronal variability could be attributed to the factors visual category and familiarity (**Fig. 6**). We estimated the effect sizes[23] by ω^2 as a function of time (Online Methods). The population conveyed information about both the visual categories and the familiarity of the stimuli (**Fig. 6c**). VS neurons signaled information earlier and did not provide novelty information (**Fig. 6g**). In contrast, MS neurons signaled information about the novelty of the stimulus, but not its categorical identity (**Fig. 6f**). To analyze neuronal activity regardless of time, we averaged the effect size in a 1.5-s time window starting 0.2 s after stimulus onset. Units classified as MS and VS neurons tended to have high effect sizes only for novelty and familiarity or category, respectively (**Fig. 6h–k**). The effect sizes were not correlated, indicating that a neuron coded either familiarity and novelty or category, but not both (**Fig. 6d**,**e**). This was true for MS, VS and all other neurons (r = 0.04, −0.003 and −0.008, respectively; $P > 0.86$; **Fig. 6e**). Thus, a neuron was informative about one, but not both, of the variables. We also used a regression model

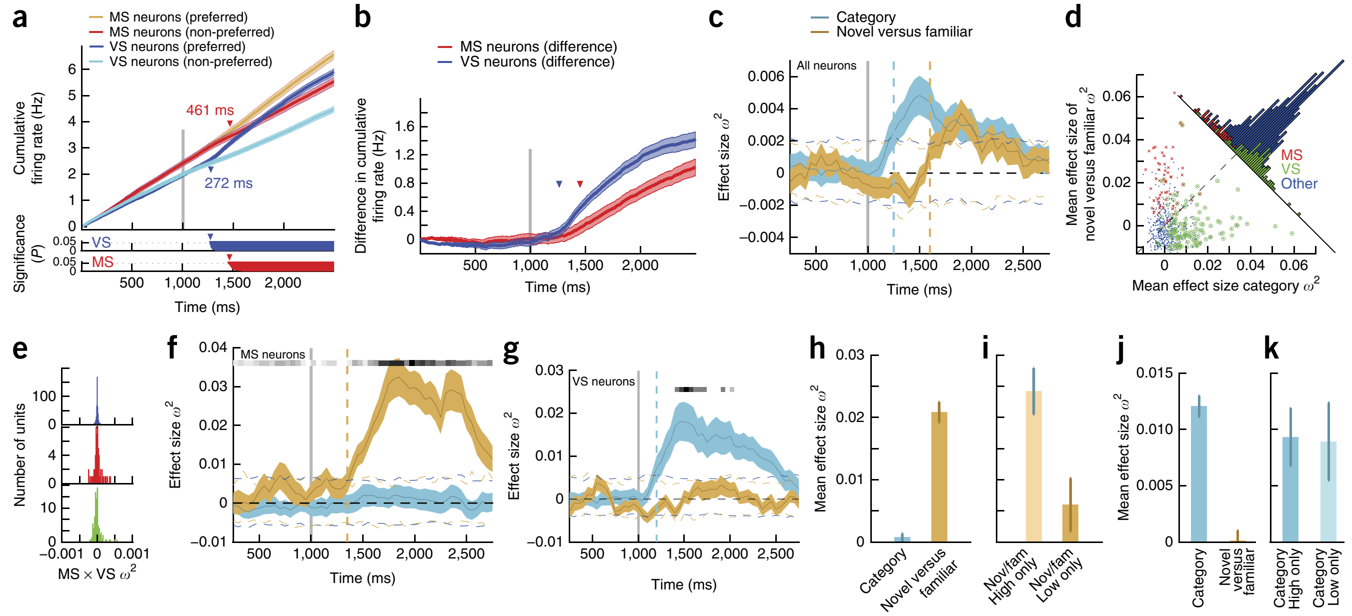

Figure 6 MS and VS neurons signal at different times and only MS neurons are sensitive to confidence. (**a**,**b**) Cumulative firing rate for MS and VS neurons. Pairwise comparison (**a**, bottom; cluster-corrected P values) between the preferred and non-preferred stimulus revealed differences in time course. (**b**) Pairwise difference for both populations. (**c**–**k**) Effect size estimation for populations of neurons based on a regression model. ω^2 is used to estimate effect size. (**c**) Time course of effect size, averaged across all neurons ($N = 664$) and computed separately for the variable category (blue) and novel and familiar (yellow). Dashed horizontal lines indicate the 99% confidence intervals of the null distribution. Dashed vertical lines indicate first time point significantly above the 99% confidence interval. Stimulus onset was at 1,000 ms (gray line). (**d**) Average effect size (1.5-s window starting 200 ms after stimulus onset) of category and novel versus familiar regressors for each neuron. (**e**) Product of ω^2 for regressors novel and familiar and category for MS, VS and other neurons. There was no significant correlation ($P > 0.86$ for all, t test versus 0). (**f**) Data presented as in **c**, but for MS neurons only. MS neurons did not distinguish categories. (**g**) Data are presented as in **c**, but for VS neurons only. VS neurons did not distinguish novel from familiar stimuli. Grayscale horizontal line in **f** and **g** indicates proportion of significant units (from white to black) at every point of time, based on the 99% confidence interval. (**h**) MS neurons had significantly larger effect size for regressor novel and familiar compared with category ($P = 0$). (**i**) Effect size of MS neurons was significantly modulated by confidence ($P = 0.0049$). (**j**,**k**) Average effect size for VS neurons was significantly larger for category information ($P = 0.0049$, **j**) and was not sensitive to confidence ($P = 0.81$, **k**). All P values are paired t tests. Bin size was 500 ms, step size was 50 ms, error bars and shaded regions represent ±s.e.m. across neurons.

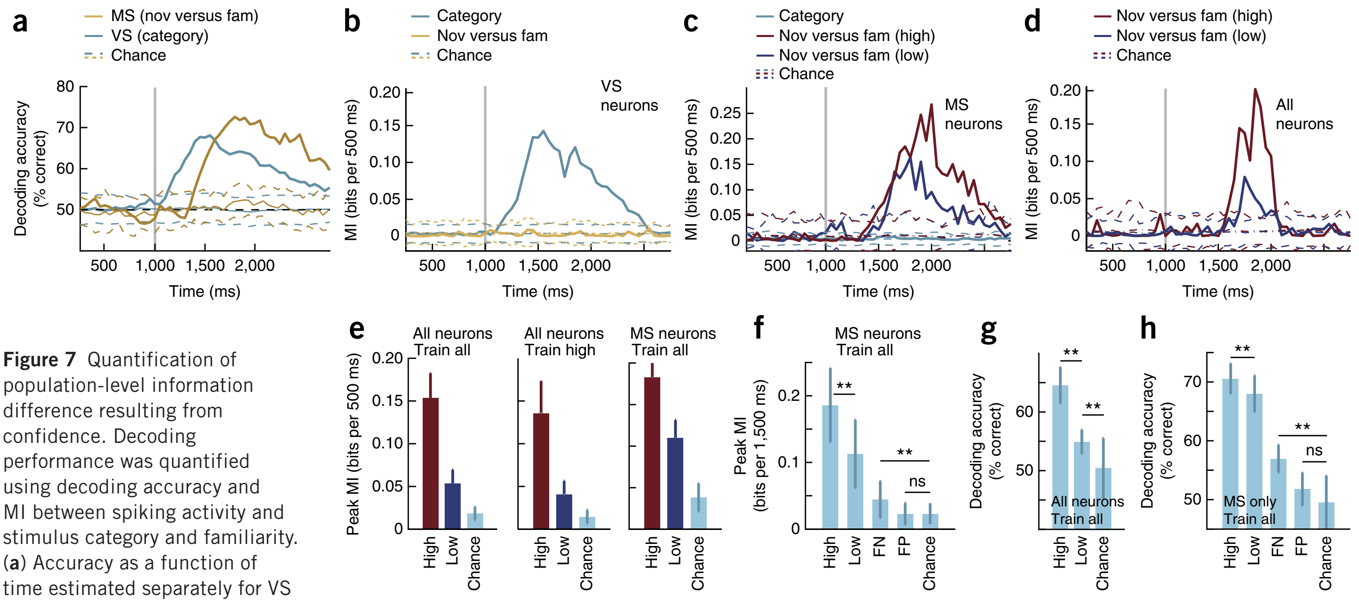

Figure 7 Quantification of population-level information difference resulting from confidence. Decoding performance was quantified using decoding accuracy and MI between spiking activity and stimulus category and familiarity. (**a**) Accuracy as a function of time estimated separately for VS and MS neurons while decoding visual category and familiarity, respectively. (**b**,**c**) VS neurons ($n = 128$) and MS ($n = 59$) neurons only signaled category and familiarity information, respectively. (**c**) Spiking of MS neurons contained more information about familiarity for high-confidence trials. (**d**) Spiking activity of all recorded neurons ($n = 606$) together contained more information for high-confidence trials. (**e**) Statistical comparison of MI for high- and low-confidence trials. A subset of $n = 200$ (all) and $n = 20$ (MS) units was chosen at random from the entire population (bootstrap, 50 runs) and the peak MI was estimated for each run. More information was available for all neurons (left) as well as for MS neurons only (right), regardless of whether the decoder was trained with all (left) or only high-confidence (middle) trials (high versus low and low versus chance is $P < 0.001$ for all). (**f**) Decoding of error trials using a subset of $n = 30$ MS neurons chosen at random from the population. Decoder was trained on all correct trials and separately evaluated on high and low confidence as well as forgotten (FN) and false positive (FP) trials. Performance for FN was above chance ($P = 0.003$), but FP was not ($P = 0.98$). FN performance was significantly lower than low confidence ($P < 10^{-5}$). (**g**,**h**) Quantification of overall readout ability (1.5-s window), regardless of time, for all neurons (**g**) and MS neurons only (**h**; ns, $P = 0.06$). (**e**–**h**) Error bars represent ±s.d. across bootstrap runs. Dashed lines in **a**–**d** show the mean ± 99% confidence interval of the null distribution. **$P < 0.01$, ns indicates not significant ($P > 0.05$).

with an interaction term, which did not explain any additional variance (**Supplementary Fig. 6**). Comparing the effect size between trials that were recognized with high and low confidence revealed that the information conveyed by MS neurons (**Fig. 6h**,**i**) was sensitive to subjective confidence, whereas that conveyed by VS neurons was not (**Fig. 6j**,**k**). Note that the estimated effect size of a neuron did not depend on spike sorting quality (**Supplementary Fig. 1h**,**i**).

Estimate of information content

What distinguishes a high- from a low-confidence memory? We used a population decoder to estimate the amount of information provided in single trials as a function of confidence and accuracy. The decoder had access to a pseudo-population of neurons and was trained and tested on subsets of independent trials. The resulting estimates are generalization errors, permitting comparisons such as whether training the decoder with a condition (that is, high confidence) generalizes to other conditions (that is, low confidence). Applying this method to all recorded VS and MS neurons revealed that visual information carried by VS neurons could be decoded earlier than memory information carried by MS neurons (**Fig. 7a**). This extends the earlier finding to single-trial decoding. To quantify the information available we used the mutual information (MI) between the spiking response and stimulus identity and familiarity (Online Methods). This again revealed an early and late component that was carried by VS and MS neurons (**Fig. 7b**,**c**). We next trained a decoder that had access to all recorded neurons using only high-confidence trials and tested its performance on both high- and low-confidence trials (**Fig. 7c**,**d**). Although this decoder based its decisions on neurons signaling high-confidence memories, low-confidence trials could still be decoded, albeit the amount of information available was reduced by ~70% (0.14 ± 0.04 versus 0.04 ± 0.02 bits; **Fig. 7e**). Thus, the population response identified for high-confidence trials is still informative for low-confidence memories. Training a decoder on all trials regardless of confidence and testing it on high and low confidence trials separately revealed similar results (0.15 ± 0.03 versus 0.05 ± 0.02 bits; **Fig. 7e**). This result also holds when only considering MS neurons (**Fig. 7e**). We conclude that the amount of information available in the entire population, in bits, is approximately threefold higher for high- than for low-confidence memories. We next estimated the MI during error trials. This revealed that, when a stimulus was forgotten (false negative, FN), the spiking activity of MS neurons still contained information about the familiarity of the stimulus (**Fig. 7f**). Although more than expected by chance (0.044 versus 0.023 bits, 1.97-fold more information), this was less than that available for low-confidence correct trials (**Fig. 7e**). Forgotten trials thus form a continuum with the low- and high-confidence correct trials, a property that is expected of a memory strength signal. Note that, in contrast with MI, decoding accuracy cannot be used to compare amounts of information. Nevertheless, a similar qualitative pattern of readout ability was revealed by decoding accuracy (**Fig. 7g**,**h**).

Differences in electrophysiological signatures

We next compared the shape of the extracellular waveforms (EWs) associated with each neuron to investigate whether VS and MS cells might be physiologically different. The trough-to-peak time *d* (**Supplementary Fig. 7a**) was bimodally distributed across all recorded neurons (**Supplementary Fig. 7a**,**b**), indicating at least two types of EWs: short and long (mode 0.3 ms and 0.8 ms; **Supplementary Fig. 7b**,**c**). Considering *d* separately for particularly well-isolated MS and VS neurons (projection test distance > 10 s.d.; all conclusions

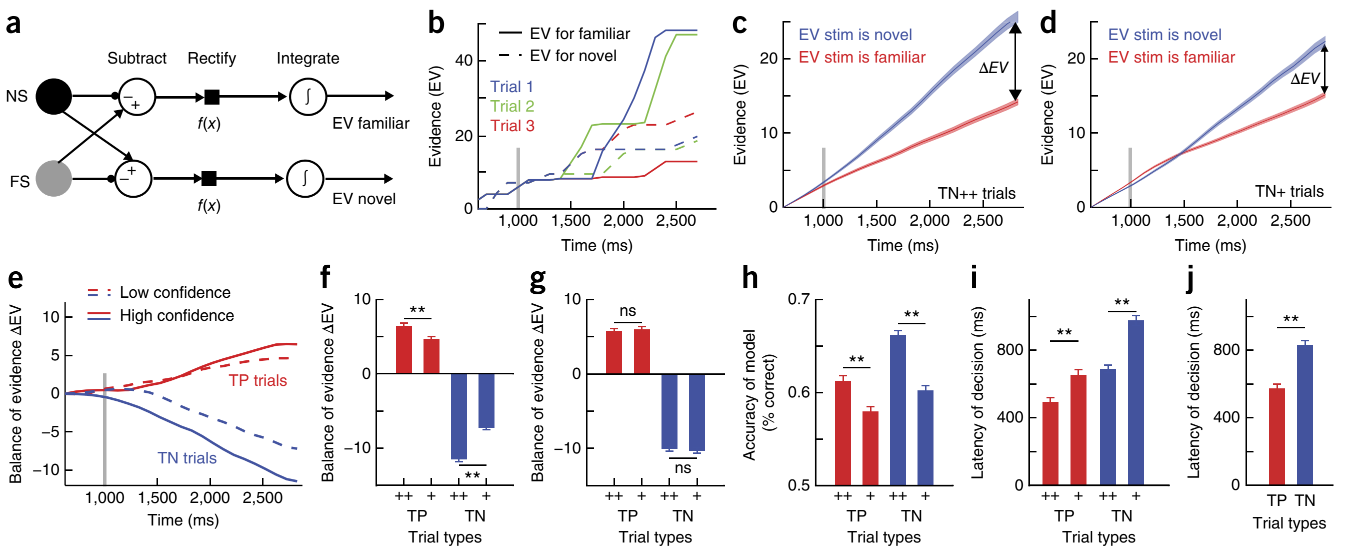

Figure 8 Computational model to decide the familiarity and confidence of a stimulus. (**a**) Circuit diagram of a race model that integrates the difference of the output of an NS and FS neuron. (**b**) Model output for three familiar (TP) trials for an example pair of neurons. Decision was made correctly for trials 1 and 2, incorrectly for trial 3. (**c**,**d**) Model output for all (FS,NS) neuron pairs (n = 951) for novel (TN) trials for high (**c**) and low confidence (**d**), respectively. Note how the balance of evidence, ΔE, was larger for high-confidence trials. Shading represents 99% confidence intervals across pairs of neurons. Marked time points are the centers of each bin (bin size = 250 ms). (**e**) ΔE as a function of time for all four trial types. Here, $\Delta E = EV_{\mathrm{fam}} - EV_{\mathrm{nov}}$, making ΔE negative for TN trials. (**f**) Average ΔE for the last time point in **e** for all neuron pairs (n = 951). ΔE was significantly larger for high- relative to low-confidence trials (pairwise t test, $P < 10^{-6}$). (**g**) Control, random reassignment of confidences abolished the difference while keeping new and old performance intact (P = 0.56 and 0.45, respectively). (**h**) Single-trial model performance for determining the familiarity of a stimulus. Performance was higher for high- compared with low-confidence trials (pairwise t test, $P < 10^{-5}$). (**i**) Latency to reach a decision, as a function of confidence. High-confidence trials had significantly shorter latency ($P < 10^{-14}$ and P = 0.00022 for TN and TP, respectively; paired t test across all cell pairs). (**j**) Familiar (TP) trials were faster than novel (TN) trials ($P < 10^{-11}$, paired t test). All error bars represent ± s.e.m. across all neuron pairs. $^{**}P < 0.01$, ns indicates not significant.

remain valid without this criteria) revealed that only the EWs of VS neurons were significant bimodally distributed (Hartigan's dip test, P = 0.004 for VS neurons and P = 0.34 for MS neurons; **Supplementary Fig. 7d**). In contrast, 72% of all EWs of MS neurons were short (**Supplementary Fig. 7f**). The proportion of long and short EWs was significantly different for MS, but not VS, neurons (χ^2 comparison of proportions, $P = 2.2 \times 10^{-5}$ and P = 0.12 for MS and VS neurons, respectively; **Supplementary Fig. 7f**). At the same time, both VS and MS neurons had low firing rates and did not differ according to other spike train metrics (modified coefficient of variation (CV_2) and burst index; **Supplementary Tables 2** and **4**). In conclusion, both MS and VS neurons had low firing rates, but MS neurons had mostly short EWs. Thus, we hypothesize that MS neurons are anatomically distinct from VS neurons.

Decision-making model

Is the information provided by MS neurons sufficient to decide both whether a stimulus is familiar as well as the confidence in that decision? To answer this question, we constructed a biologically plausible race model[24]. The model evaluates whether the difference $D(t)$ between one FS and NS neuron is negative or positive (**Fig. 8a**). If positive, the accumulated evidence (EV) for the stimulus being familiar is increased and vice-versa for negative $D(t)$. At the end of the trial, the decision is familiar if $EV_{\mathrm{fam}} > EV_{\mathrm{nov}}$, and novel if otherwise. The confidence in the decision is proportional to the 'balance of evidence' $\Delta E = |EV_{\mathrm{fam}} - EV_{\mathrm{nov}}|$ (ref. 25). We evaluated the performance of this model for all n = 954 pairs of NS and FS neurons, separately for correctly recognized familiar (TP) and novel (TN) items (**Fig. 8b–h**). The model reliably distinguished between high- and low-confidence trials (**Fig. 8c–f**) and EV and ΔE were correlated with behavioral performance. The model's ability to distinguish between novel and familiar stimuli was better for high- than for low-confidence trials (**Fig. 8h**). In addition, ΔE was correlated trial by trial with confidence, both for behaviorally correct and incorrect trials (Spearman correlation; correct, 0.042 ± 0.13, $P < 10^{-20}$, n = 957 pairs; incorrect trials 0.047 ± 0.17, P = 0.0033 versus 0, n = 130 pairs). Of the two EV values, only the larger value (the winner) correlated with confidence (0.05 ± 0.13, $P < 10^{-30}$), whereas the EV value of the smaller (looser) value did not (0.002 ± 0.16, P = 0.68). We also used the model to evaluate the decision latency by setting for each cell pair, a fixed decision threshold ΔE_{Th} (Online Methods). The first time when ΔE exceeded this threshold, the race was aborted and the latency noted. This model made decisions more quickly for trials that were made with high confidence (**Fig. 8i**) and made familiar decisions more quickly than novel decisions (**Fig. 8j**). This pattern is similar to that observed behaviorally (**Fig. 1i**). Together, our findings show that a simple readout mechanism can reliably, and on single trials, make two decisions simultaneously using only information provided by MS neurons.

DISCUSSION

We systematically compared two populations of neurons in the human MTL: VS and MS neurons. The former signaled information about the identity of the visual stimuli, whereas the latter signaled the familiarity of the stimuli. VS neurons discriminated between stimuli ~190 ms earlier than MS neurons and only the activity of MS neurons was correlated with memory strength, as expressed by a confidence judgment. Together, our results suggest that only MS neurons are directly involved in memory retrieval. The proportion of MS neurons identified here was similar to those identified previously[13,14,26]. However, using confidence ratings revealed several important new aspects of these neurons. In particular, this revealed that NS and FS neurons coded information asymmetrically: their firing rate is only informative about the confidence of the trial types to which they increase their firing rate (**Fig. 3**). In contrast, we found that the activity of the VS neurons was not sensitive to memory strength and that they were functionally distinct from MS neurons. In addition, our data is an independent reproduction of the initial description of VS neurons[9]. 1.5% of all neurons qualified as both VS and MS neurons. Although rare, our large data set shows that the probabilities of a neuron to become a VS or MS neuron are independent of each other. Such neurons have been hypothesized to represent a distributed sparse code for memories[27,28], but, given their rarity, it will be necessary to use closed-loop procedures to investigate them systematically.

Our conclusions rest on single-neuron ROC analysis, a sensitive method for quantifying the amount of information available in individual trials[29]. ROC analysis does not assume a particular distribution of the spike counts, which is important because spike counts are Poisson distributed. Using mutual information, we further estimated that the amount of information present in the population is about threefold higher in high- relative to a low-confidence trials. Note that low-confidence decisions were nevertheless correct; what was missing was additional information required to reach a high-confidence choice. In addition, low-confidence decisions were slower, a signature of recognition memory that has been observed even when not asking for a confidence[1].

Confidence judgments are subjective. Consequently, the strength associated with a certain confidence varies between subjects. Our analysis, however, is insensitive to this because it relies on a within-neuron comparison between high- and low-confidence trials. As a result, all that is required for our analysis to be valid is that subjects apply a threshold regardless of its value. For statistical reasons, we focused our analysis on two levels of confidence only. A third level is FN trials, which can be considered a 'very low' confidence. Our results indicate that these three levels are represented by MS neurons. Clearly, subjects are capable of using more than two confidence levels[1] and it remains an open question whether each of these can be separated by MS neurons.

Could the neuronal differences between high and low confidence be attributed to fluctuations in attention during retrieval? The specificity of the neuronal effects argues against this possibility, as a global attentional effect would affect all neurons equally. In particular, it would be expected to improve the reliability of visual category information[30]. Instead, we found no difference in the coding reliability of VS neurons.

In psychology, global models of recognition memory[1,31,32] have as their underlying decision variable a familiarity or strength signal that pools memory strength among many associations or items. In these models, the familiarity signal itself does not contain information about the memory apart from signaling its familiarity. MS neurons had the same property and are therefore candidates for the familiarity signal predicted by these models. This will make it possible to directly test key hypothesis made by these influential quantitative models of memory[32].

We used a simple integrator-type model to explore which decisions could be supported by the difference in firing rate between a pair of FS and NS neurons. Integration of the difference of two neurons with opposite tuning is statistically optimal in many situations[24]. Our model differs from drift-diffusion (DDM) models[24,33] because it has two integrators, only one of which increases its value depending on the sign of the difference. FS and NS neurons are not anti-correlated (**Fig. 3f–i**), and the two integrators are therefore not redundant, as is assumed in DDM models. The difference of the two integrators is the 'balance of evidence'[5,7,25]. In contrast, a standard DDM model has only one decision variable[34] and therefore has no mechanism for estimating the quality of a decision beyond the time taken to reach the decision threshold[3]. We found that integration-to-bound decision models are applicable to memory-based decisions because this model can make confidence decisions based only on the activity of MS neurons. No human neurons that represent the difference FS-NS or the integrator values EV have yet been identified, but our model makes specific predictions that will facilitate their discovery. A key technique to identify signatures of evidence accumulation has been to present sensory stimuli of different strength[22,35]. We relied on internal variability in memory strength only, but we expect that combing these two approaches will be an important future avenue.

EWs have been used to classify cells as inhibitory or excitatory[36–38], but no definitive data on the validity of this distinction exists for humans. The EW differs as a function of the location of the electrode relative to the cell, but, given that our electrodes were implanted blindly, this is unlikely to account for the difference. Large pyramidal cells can have shorter waveforms than smaller pyramids[39], and in rats particularly short waveforms are hypothesized to represent axonal activity[40]. In addition, backpropagation of action potentials widens the EW[41] and the propensity for backpropagation varies between cell types. Consequently, an intriguing possibility is that MS cells are morphologically and/or physiologically different from VS cells, but this hypothesis remains to be confirmed.

In addition to the hippocampus, we identified VS and MS cells in the amygdala, confirming previous reports of memory signals in the human amygdala[13,14,26]. Although the amygdala is not necessary for declarative memory, it is crucial for many aspects of learning[42] and is sensitive to stimulus novelty[43]. Given this, it is not surprising that VS and MS cells are also present in the amygdala. We used natural scenes as stimuli, some with emotional content. It remains an open question whether MS cells in the amygdala are specifically modulated by the emotional content of the stimuli. It also remains an open question whether MS cells are modulated by recency rather than novelty. Lists of words are frequently used in recognition memory[1] tests, but most physiological studies thus far have used natural scenes. Notably, a recent study using words reported cells tuned to recently seen words, but not broadly tuned cells of the kind we observed[27].

Assessing the quality of one's own memory (an internal state) is thought to require metacognition[44], the existence of which in animals is debated[5,45,46]. Although only humans can verbally declare their confidence, experiments with indirect measures reveal that several species can utilize a 'don't know' option[3,7,47,48] alone or in combination with post-decision wagering[3,49] to prevent the learning of an association instead of a confidence judgment. The amount of effort expended has also been used to infer confidence[50]. Theoretically, degrees of uncertainty are central components of neural computation[5,6]. Together, there is emerging evidence that an assessment of uncertainty is an integral part of neuronal decision making in general. Here, we found that MS neurons in humans support assessments of uncertainty in memory-based decisions because they carry a graded representation of memory strength that is reflected in the subjective confidence ratings made by the subjects.

METHODS

Methods and any associated 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. Kaminski, R. Adolphs, C. Anastassiou, U. Maoz, J. Wertheimer and W. Einhaeuser for discussion, Z. Fu for spike sorting, C. Heller for performing some of the surgeries, the staff of the Epilepsy Monitoring Units at Huntington Memorial Hospital and Cedars-Sinai for invaluable assistance, particularly J. Schmidt. We thank K. Birch and H. Babu for assistance with patient care and surgery, and L. Philpott and M.-T. Le for neuropsychological testing. This work was supported by the Cedars-Sinai Medical Center Department of Neurosurgery (to U.R.), National Institute of Mental Health Conte Center at Caltech (P50 MH094258), and the Gustavus and Louise Pfeiffer Research Foundation (to U.R.).

AUTHOR CONTRIBUTIONS

U.R. and A.N.M. designed the experiments. U.R. and O.T. performed experiments. U.R., M.K. and S.Y. performed analysis. A.N.M. and I.B.R. performed surgery. J.M.C. provided patient care. U.R. and A.N.M. 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. Kahana, M.J. *Foundations of Human Memory* (Oxford University Press, New York, 2012).
2. Petrusic, W.M. & Baranski, J.V. Judging confidence influences decision processing in comparative judgments. *Psychon. Bull. Rev.* **10**, 177–183 (2003).
3. Kiani, R. & Shadlen, M.N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
4. Smith, J.D., Shields, W.E. & Washburn, D.A. The comparative psychology of uncertainty monitoring and metacognition. *Behav. Brain Sci.* **26**, 317–339, discussion 340–373 (2003).
5. Kepecs, A. & Mainen, Z.F. A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1322–1337 (2012).
6. Pouget, A., Dayan, P. & Zemel, R.S. Inference and computation with population codes. *Annu. Rev. Neurosci.* **26**, 381–410 (2003).
7. Kepecs, A., Uchida, N., Zariwala, H.A. & Mainen, Z.F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
8. Squire, L.R., Stark, C.E. & Clark, R.E. The medial temporal lobe. *Annu. Rev. Neurosci.* **27**, 279–306 (2004).
9. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* **3**, 946–953 (2000).
10. Viskontas, I.V., Quiroga, R.Q. & Fried, I. Human medial temporal lobe neurons respond preferentially to personally relevant images. *Proc. Natl. Acad. Sci. USA* **106**, 21329–21334 (2009).
11. Logothetis, N.K. & Sheinberg, D.L. Visual object recognition. *Annu. Rev. Neurosci.* **19**, 577–621 (1996).
12. Rolls, E.T. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* **27**, 205–218 (2000).
13. Rutishauser, U., Mamelak, A.N. & Schuman, E.M. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. *Neuron* **49**, 805–813 (2006).
14. Rutishauser, U., Schuman, E.M. & Mamelak, A.N. Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. *Proc. Natl. Acad. Sci. USA* **105**, 329–334 (2008).
15. Xiang, J.Z. & Brown, M.W. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology* **37**, 657–676 (1998).
16. Wilson, F.A. & Rolls, E.T. The effects of stimulus novelty and familiarity on neuronal activity in the amygdala of monkeys performing recognition memory tasks. *Exp. Brain Res.* **93**, 367–382 (1993).
17. 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).
18. Green, D. & Swets, J. *Signal Detection Theory and Psychophysics* (Wiley, 1966).
19. Manns, J.R., Hopkins, R.O., Reed, J.M., Kitchener, E.G. & Squire, L.R. Recognition memory and the human hippocampus. *Neuron* **37**, 171–180 (2003).
20. Rutishauser, U., Schuman, E.M. & Mamelak, A.N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, *in vivo*. *J. Neurosci. Methods* **154**, 204–224 (2006).
21. Viskontas, I.V., Knowlton, B.J., Steinmetz, P.N. & Fried, I. Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal regions. *J. Cogn. Neurosci.* **18**, 1654–1662 (2006).
22. Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S. & Movshon, J.A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13**, 87–100 (1996).
23. Hentschke, H. & Stuttgen, M.C. Computation of measures of effect size for neuroscience data sets. *Eur. J. Neurosci.* **34**, 1887–1894 (2011).
24. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J.D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
25. Vickers, D. *Decision Processes in Visual Perception* (Academic Press, New York, 1979).
26. Fried, I., MacDonald, K.A. & Wilson, C.L. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* **18**, 753–765 (1997).
27. Wixted, J.T. *et al.* Sparse and distributed coding of episodic memory in neurons of the human hippocampus. *Proc. Natl. Acad. Sci. USA* **111**, 9621–9626 (2014).
28. Marr, D. Simple memory: a theory for archicortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **262**, 23–81 (1971).
29. Macmillan, N.A. & Creelman, C.D. *Detection Theory* (Lawrence Associates, Mahwah, New Jersey, 2005).
30. Zhang, Y. *et al.* Object decoding with attention in inferior temporal cortex. *Proc. Natl. Acad. Sci. USA* **108**, 8850–8855 (2011).
31. Wixted, J.T. Dual-process theory and signal-detection theory of recognition memory. *Psychol. Rev.* **114**, 152–176 (2007).
32. Clark, S.E. & Gronlund, S.D. Global matching models of recognition memory: How the models match the data. *Psychon. Bull. Rev.* **3**, 37–60 (1996).
33. Gold, J.I. & Shadlen, M.N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
34. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
35. Hanks, T.D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223 (2015).
36. Viskontas, I.V., Ekstrom, A.D., Wilson, C.L. & Fried, I. Characterizing interneuron and pyramidal cells in the human medial temporal lobe *in vivo* using extracellular recordings. *Hippocampus* **17**, 49–57 (2007).
37. Mitchell, J.F., Sundberg, K.A. & Reynolds, J.H. Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* **55**, 131–141 (2007).
38. Peyrache, A. *et al.* Spatiotemporal dynamics of neocortical excitation and inhibition during human sleep. *Proc. Natl. Acad. Sci. USA* **109**, 1731–1736 (2012).
39. Vigneswaran, G., Kraskov, A. & Lemon, R.N. Large identified pyramidal cells in macaque motor and premotor cortex exhibit "thin spikes": implications for cell type classification. *J. Neurosci.* **31**, 14235–14242 (2011).
40. Robbins, A.A., Fox, S.E., Holmes, G.L., Scott, R.C. & Barry, J.M. Short duration waveforms recorded extracellularly from freely moving rats are representative of axonal activity. *Front. Neural Circ.* **7**, 181 (2013).
41. Stuart, G., Schiller, J. & Sakmann, B. Action potential initiation and propagation in rat neocortical pyramidal neurons. *J. Physiol.* (*Lond.*) **505**, 617–632 (1997).
42. Hamann, S. The human amygdala and Memory. in *The Human Amydala* (eds. Whalen, P.J. & Phelps, E.A.) 177–203 (The Guilford Press, New York, 2009).
43. Weierich, M.R., Wright, C.I., Negreira, A., Dickerson, B.C. & Barrett, L.F. Novelty as a dimension in the affective brain. *Neuroimage* **49**, 2871–2878 (2010).
44. Metcalfe, J. Metamemory. in *Learning and Memory: a Comprehensive Reference* (ed. Roediger, H.L.) 349–362 (Elsevier, Oxford, 2008).
45. Metcalfe, J. Evolution of metacognition. in *Handbook of Metamemory and Memory* (eds. Dunlovsky, J. & Bjork, R.) 29–46 (Psychology Press, New York, 2008).
46. Hampton, R.R. Rhesus monkeys know when they remember. *Proc. Natl. Acad. Sci. USA* **98**, 5359–5362 (2001).
47. Perry, C.J. & Barron, A.B. Honey bees selectively avoid difficult choices. *Proc. Natl. Acad. Sci. USA* **110**, 19155–19159 (2013).
48. Foote, A.L. & Crystal, J.D. Metacognition in the rat. *Curr. Biol.* **17**, 551–555 (2007).
49. Middlebrooks, P.G. & Sommer, M.A. Metacognition in monkeys during an oculomotor task. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 325–337 (2011).
50. Fortin, N.J., Wright, S.P. & Eichenbaum, H. Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature* **431**, 188–191 (2004).

ONLINE METHODS

Electrophysiology and electrodes. Broadband extracellular recordings were filtered 0.1 Hz to 9 kHz and sampled at 32 kHz (Neuralynx). We recorded bilaterally from the amygdala and hippocampus (32 channels in total, see ref. 17 for details). One microwire in each macroelectrode served as a local reference (bi-polar recording). Electrodes were localized based on post-operative MRI images[17]. Electrode locations were chosen according to clinical criteria alone. Only electrodes localized to the hippocampus or amygdala were included. Protocols were approved by the institutional review boards of the Cedars-Sinai Medical Center, Huntington Memorial Hospital and the California Institute of Technology.

Patients. 28 patients who were evaluated for possible surgical treatment of epilepsy using implantation of depth electrodes volunteered for the study and gave informed consent. We evaluated all patients using standard neuropsychological tests (**Supplementary Table 1**). All included patients had clearly distinguishable spiking activity on at least one electrode in the areas of interest.

Task. Details of the task have been published previously[17]. The task consisted of two blocks: learning and retrieval, with a 15–30-min delay in between with a distractor task. During learning, 100 novel and unique images were shown. During recognition, a subset of 50 of these images were shown again (now familiar, old) together with 50 novel images (novel, new). Patients identified each image as novel or familiar on a 1–6 confidence scale (**Fig. 1a**). Only the data from the retrieval block of the task is reported here. Before the experiment, subjects performed a short training version of the same task but with different images. Some that performed multiple sessions of the task were recorded on different days with different sets of images. Images shown were 9° × 9° degrees in size. After offset of the image, the screen was blank and followed by the question screen 0.5 s later (**Fig. 1a**) that was displayed till an answer was provided. Stimuli were photographs of natural scenes of five different visual categories (animals, people, cars/vehicles, outdoor scenes/houses and flowers/food items). There were the same numbers of images presented in each category. The task was implemented in MATLAB using the Psychophysics Toolbox[51].

Behavioral analysis. The decision time (DT) is the time between onset of the question screen and the button press. We excluded DTs >30 s as well as those which are more than 3 s.d. away from the mean (for each subject) for all DT analysis (1.74% ± 1.00% of trials, ± s.d. across subjects, were removed; **Fig. 1h**–**k**). All DT comparisons were pairwise within-subject comparisons. We excluded sessions which did not contribute at least one data point to each category of a comparison (number of sessions for **Fig. 1i**–**l** are 38, 42, 44 and 31, respectively). All findings reported in **Figure 1i**–**l** remain when using all 44 sessions and non-paired statistics (data not shown). To analyze behavioral performance and proportion of responses (**Fig. 1b**–**h**), all trials regardless of DT were included. Note that the proportion of responses (**Fig. 1f**) remains virtually unchanged when applying the same exclusion criteria as used for the DT analysis.

The association between confidence and retrieval accuracy was assessed using the Goodman-Kruskal gamma coefficient g[52], whose value is between −1.1. The relation $V = 0.5g + 0.5$ converts g into the probability V that a confidence judgment is accurate[52]. On average, $V = 0.67 \pm 0.18$ (±s.d.).

We used a three-way repeated measure ANOVA with in-between factors memory (novel/familiar), confidence (high/low), and accuracy (correct/incorrect) to quantify the relationship with DT. The repeated factor was subject number. Pairwise *post hoc* comparisons were done using a Wilcoxon signed-rank test.

The behavioral ROC was calculated as a function of confidence as described previously[17]. The slope of a line fitted with least-squares regression to the *z*-transformed ROC was used to assess the degree of asymmetry of the ROC[53]. We reassigned the intermediate confidence level (2, 5) to either the low or high confidence level to collapse the 6 confidence levels to 4 levels. For every session, the intermediate confidence was assigned to either the low or high confidence group, based on which assignment produced a more equal proportion of high and low trials. This re-balancing was based on number of trials alone.

We assigned sessions to two groups. Group 1 consists of all sessions where patients performed at least 10% above chance. Group 2 is a subset of group 1 and contains only sessions where patients accurately discriminated between high and low confidence memories (minimal accuracy for high 70% and low 55%). Using random subsets of 50% of the trials or only the first or second half of the trials resulted in identical group assignments.

Spike detection, sorting and quality metrics. The raw signal was filtered with a zero-phase lag filter in the 300–3,000-Hz band and spikes were detected and sorted using the semiautomated template-matching algorithm OSort[20]. Channels with interictal epileptic activity were excluded. We computed several spike sorting quality metrics for all units (**Supplementary Fig. 1**): i) percentage of interspike intervals (ISIs) below 3 ms was 0.24% ± 0.45%, ii) the ratio between the peak amplitude of the mean waveform of each cluster and the s.d. of the noise was 5.6 ± 3.6 (peak SNR), iii) the pairwise projection distance in clustering space between all neurons isolated on the same wire was 16 ± 11 (projection test[54]; in units of s.d. of the signal), iv) the modified coefficient of variation of variability in the ISI (CV2) was 0.93 ± 0.21 ($P = 0.72$, not significantly different from 1, as expected from a Poisson process), and v) the isolation distance[55,56] (**Supplementary Fig. 1g**; $n = 746$, median was 35.0; compare with **Supplementary Fig. 2b** in ref. 57 and **Fig. 7** in ref. 56). The isolation distance quantifies, for every cluster, how far apart it is from the other clusters and the noise. We calculated the isolation distance in a ten-dimensional feature space[56] (energy, peak amplitude, total area under the waveform and first five principal components of the energy normalizes waveforms). To quantify whether our results depend on sorting quality, we correlated the effect size metric ω^2 with the isolation distance (**Supplementary Fig. 1h**,**i**).

Selection of units. We counted spikes in a 200–1,700-ms window relative to stimulus onset. MS neurons were selected based on a significant difference between correctly identified novel and familiar stimuli in this period ($P < 0.05$, two-tailed, boostrap comparison of means with 1,000 runs). A MS neuron was FS if the mean if all familiar trials was larger than all novel trials and NS otherwise. VS neurons were selected using a 1 × 5 ANOVA with the factor visual category (1–5) based on the identical spike counts and with $P < 0.05$.

Single-neuron analysis. We used non-overlapping bins of 250-ms width. PSTH diagrams were smoothened, for display only, with a causal exponential kernel with $\lambda = 150$ ms. All analysis and statistics was based on un-smoothened data.

Single-neuron ROC analysis. Neuronal ROCs were constructed based on the spike counts in a 1.5-s-long window, starting 200 ms after stimulus onset. We varied the detection threshold between the minimal and maximal spike count observed, linearly spaced in 25 steps. The AUC of the ROC was calculated by integrating the area under the ROC curve[18].

For MS neurons, ROC analysis was performed to quantify how well individual neurons distinguished between novel and familiar trials. Only neurons with at least ten correct novel and familiar trials each were included in the ROC analysis. A separate ROC analysis was performed for high and low confidence trials. For confidence comparisons, only neurons that had at least two trials of each of the four confidence levels were included. To perform a fair comparison, only one of the two groups used for the ROC analysis was modified according to confidence while the other was kept constant. For FS neurons, the fixed group was all TN trials (regardless of confidence) which was compared with high-confident TP and low-confident TP trials separately. For NS neurons, the fixed group was all TP trials which were compared with high-confident TN and low-confident TN trials separately.

For VS neurons, we first identified, based on all trials regardless of behavior, a binary contrast (such as category 2 versus 5, preferred versus non-preferred) that a neuron distinguished best by testing all ten possible contrasts and picking the one with the maximal AUC. We subsequently estimated the AUC for this best contrast using only novel, familiar, high, and low confidence correct trials.

Statistical comparisons between AUC values were made using two-tailed parametric tests (paired t test and paired sign-tests, as indicated). For bootstrap comparisons, we performed B = 1,000 bootstrap runs to estimate the null distribution and estimated the P value empirically by counting how many values in the null distribution were larger than the observed value. When no null distribution value exceeded the observed value, we set the P value to 1/B.

To calculate a normalized firing rate (**Fig. 3f**–**i**), we divided the firing rate by the mean firing rate of the neuron in the entire task. For the cumulative distribution comparisons (**Fig. 3f**–**i**), we only included neurons that had at least two trials in each of the six behavioral categories.

Differential latency. We binned spike trains into 1-ms bins and computed the cumulative sum. We then averaged the cumulative sums of all individual trials of a neuron that belong to the same condition. To allow averaging of all MS neurons, NS neurons were inverted so that the preferred response of all MS neurons was a firing rate increase. For VS neurons, the best contrast was used as determined by ROC analysis. We then compared, at every point of time, whether the cumulative sums of a group of neurons were different ($P < 0.05$, pairwise t test). We repeated this procedure after randomly scrambling the labels to estimate the null distribution. Corrections for multiple comparisons were performed using a cluster-size correction. The maximal number of consecutively significant data points in the null distribution was used as the minimal cluster size. The first point of time of the first significant cluster was used as the estimate of the differential latency[15]. Note that this method is not sensitive to baseline firing rate differences between neurons because the latency estimate is pairwise for each neuron individually.

Regression analysis. We used the regression model $S(t) = \alpha_0(t) + \alpha_1(t)N + \alpha_2(t)C$ to estimate whether the firing rate S was significantly related to the factors novelty/familiarity (N) or category (C). Both factors were binary (0/1) to make the effect size comparable. We quantified the effect size of each regressor using the effect size metric ω^2, which is better suited for our purposes than more traditional variance explained or p-value metrics[23]. This is because ω^2 is not biased for small numbers of trials and tends toward zero if a factor has no explanatory power[58]. To estimate ω^2 for the factor category regardless of tuning of a neuron, we fit 5 models to each neuron, each contrasting one category with the remaining four. We then averaged the resulting ω^2. Spike counts $S(t)$ were computed for a 500-ms window that was moved in steps of 50 ms. Here,

$$\omega_i^2 = \frac{[SS_i - df_i \times MSE]}{[SS_{\text{tot}} + MSE]}$$

where SS_i is the sum of squares of factor I, SS_{tot} the total sum of squares of the model and MSE the mean square error of the model. We fit the model and calculated effect sizes using the effect size toolbox functions mes1way and mes2way[23]. We averaged $\omega^2(t)$ across all neurons (**Fig. 6**). The null distribution was estimated by randomly scrambling the labels and fitting the same model. This was repeated 1,000 times to estimate the 99% confidence interval of the null distribution. Estimates of latency were based on the first time the actual value was located outside of the 99% confidence interval. To estimate potential interactions, we also fit the model $S(t) = \alpha_0(t) + \alpha_1(t)N + \alpha_2(t)C + \alpha_3(t)N \times C$ and estimated $\omega^2(t)$ for each main factor and the interaction (**Supplementary Fig. 6**).

Population decoding. We pooled all recorded neurons into a pseudo-population. Firing rates were z-scored individually for each. We used a maximal correlation coefficient classifier (MCC) as implemented in the ndt toolbox[59]. The MCC estimates a mean template $\overline{x}_i$ for each class i and assigns the class $i^* = \arg\max_i corr(x^*, \overline{x}_i)$ for test trial x^*. We used tenfold cross-validation, that is, for each iteration ten trials from each class where chosen randomly from each neuron. One trial from each class was used for testing and the remaining nine for training. All possible train/test splits were tested and this process was repeated 50 times with different subsets of trials, resulting in a total of 500 runs. Spikes were counted in bins of 500-ms size and advanced by a step size of 50 ms. For each point of time, a different classifier was trained. We converted the resulting confusion matrix into mutual information (MI) $I(S;R)$[60] to estimate the information that the overall population response R provides, in a single trial, about the stimulus S. We estimated the null distribution by repeating above procedure 200 times after randomly scrambling the labels. To estimate the variability of MI across different neurons we repeated above procedure after selecting a group of 200 (all units) or 20 (MS neurons) with replacement from the overall group. We repeated this procedure 50 times, each time estimating the peak MI (**Fig. 7e**). To estimate whether the same subset of neurons is informative about high- and low confidence trials we trained decoders using all or only high confidence trials, and subsequently tested the decoders with only high or low trials. For decoding of error trials, which are relatively rare, we used larger bin sizes and smaller number of trials (**Fig. 7f–h**). Thus, we used sixfold cross-validation (5 training trials, 1 testing), a bin size of 1.5 s with step size of 50 ms and estimated the variability across neurons by randomly sub-selecting with replacement a group of 30 MS neurons. We again used the peak MI of each run and repeated this procedure 500 times (**Fig. 7f**). For estimating overall readout ability (**Fig. 7g,h**), we used a single 1.5-s-long time window starting 200 ms after stimulus onset.

Waveform analysis. The trough-to-peak time d[37] is the time between the trough and the point of time of maximal amplitude after the trough of the mean waveform. The mean waveform is the average of all spikes assigned to the cluster. For visualization, all waveforms were normalized to their maximal amplitude and were inverted if their maximum was positive. A spike waveform was considered short if $d < 0.6$ ms.

Spike-train variability. Variability was quantified for each neuron using two metrics: the modified coefficient of variation (CV2) and the burst index (BI). The BI is equal to the proportion of ISIs less than 10 ms long and the CV2 was used as defined in ref. 61. The CV2 is insensitive to underlying rate changes and is thus the appropriate metric to use in place of the normal CV[62].

Decision making model. The input to the model is the spiking activity $S_k^{i,j}(t)$ of a NS and FS neuron i and j in trial k. The difference $D(t)_k = S_k^j(t) - S_k^i(t)$ is then integrated over time. Spikes are counted in bins of 250 ms, advanced with a step size of 100 ms. Firing rates of neurons were z-scored using the mean and s.d. of the baseline (1 s before stimulus onset). The model has two state variables $EV_{\text{fam}}(t)$ and $EV_{\text{nov}}(t)$, which accumulate as following:

$$EV_{\text{fam}}(t) = \int_0^t f(D(t))$$

and

$$EV_{\text{nov}}(t) = \int_0^t f(-D(t))$$

where $f(x) = \max(0, x)$ is a rectification nonlinearity (**Fig. 8a**). The decision is familiar if $EV_{\text{fam}}(t) > EV_{\text{nov}}(t)$ and novel otherwise. Except for **Figure 8i,j**, the decision was made 2.5 s after stimulus onset. The balance of evidence is $\Delta E(t) = EV_{\text{fam}}(t) - EV_{\text{nov}}(t)$. We evaluated the model for all possible pairs ($n = 951$) of NS and FS neurons that had at least 3 behaviorally correct trials in each category (TP high/low, FN high/low). For each, we evaluated every possible pair of trials within the same behavioral category. As a control, we randomly scrambled the high and low-confidence labels for each neuron while keeping the trial identity (new/old) labels intact. This abolished the difference in balance of evidence as expected (**Fig. 8g**). To correlate ΔE and EV with performance, we computed for every cell pair separately the Spearman correlation coefficient between confidence (high or low) with $|\Delta E|$ $t = 2.5$s. We evaluated this trial-by-trial correlation for all trials remembered correctly by the subject and the model (excluding errors made by the model) as well as all trials where the subject was incorrect (errors). To make this comparison unbiased, we used the same number of high and low confidence trials by subsampling the larger group randomly. To evaluate the decision latency of the model, we terminated the decision when $|\Delta E(t)| > \Delta E_{\text{Th}}$. The decision time was equal to the first point of time at which this condition was satisified. ΔE_{Th} was set to 50% of the $|\Delta E|$ value reached at 2.5 s for every cell pair.

A **Supplementary Methods Checklist** is available.

51. Brainard, D.H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
52. Nelson, T.O. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.* **95**, 109–133 (1984).
53. Ratcliff, R., Gronlund, S.D. & Sheu, C.F. Testing global memory models using ROC curves. *Psychol. Rev.* **99**, 518–535 (1992).
54. 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).
55. Harris, K.D., Henze, D.A., Csicsvari, J., Hirase, H. & Buzsaki, G. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* **84**, 401–414 (2000).
56. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A.D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131**, 1–11 (2005).
57. Diba, K. & Buzsaki, G. Hippocampal network dynamics constrain the time lag between pyramidal cells across modified environments. *J. Neurosci.* **28**, 13448–13456 (2008).
58. Olejnik, S. & Algina, J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychol. Methods* **8**, 434–447 (2003).
59. Meyers, E.M. The neural decoding toolbox. *Front. Neuroinform.* **7**, 8 (2013).
60. Quian Quiroga, R. & Panzeri, S. Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci.* **10**, 173–185 (2009).
61. Rutishauser, U. *et al.* Single-neuron correlates of atypical face processing in autism. *Neuron* **80**, 887–899 (2013).
62. Holt, G.R., Softky, W.R., Koch, C. & Douglas, R.J. Comparison of discharge variability *in vitro* and *in vivo* in cat visual cortex neurons. *J. Neurophysiol.* **75**, 1806–1814 (1996).