Chapter 3. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex

3.1 Introduction[3]

One prominent feature of nervous systems is the ability to distinguish novel from familiar stimuli. A rapid assessment of stimulus novelty is a prerequisite for certain kinds of learning (Davis et al., 2004; Kohonen and Lehtio, 1981; Li et al., 2003; Stark and Squire, 2000; Yamaguchi et al., 2004). For instance, conditioned taste aversions (CTA) and some forms of conditioned fear can be acquired in a single learning trial. Crucially, successful conditioning depends on the novelty of the conditioned stimulus (CS) (see Welxl, 2000 for a review). Pre-exposure to the CS severely diminishes associative learning (a.k.a. “latent inhibition”). Further, conditioning is also reduced if only some aspects of the CS are novel while others are familiar. The sensitivity to CS novelty, but not the taste aversion itself, is blocked by hippocampal lesions (Gallo and Candido, 1995). The novelty dependence of single-trial learning in the CTA paradigm points to the importance of a rapid assessment of stimulus novelty or familiarity.

The medial temporal lobe (MTL) is crucial for the acquisition of declarative memories and some functional imaging techniques have shown activation of MTL structures associated with either novel or familiar stimuli (Stark and Squire, 2000; Stern et al., 1996; Tulving et al., 1996; Yamaguchi et al., 2004). Lesion studies have repeatedly demonstrated that

[3] The material in this chapter is based on Rutishauser, U., Mamelak, A.N., and Schuman, E.M. (2006a). Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. Neuron *49*, 805-813.

MTL damage impairs or abolishes behavioral, electrographic, and skin responses to novel stimuli (Kishiyama et al., 2004; Knight, 1996; Yonelinas et al., 2002). While these studies suggest a role of the MTL in novelty detection, the cellular basis for this discrimination has yet to be described. We report here that single neurons in the human MTL can alter their firing behavior to discriminate between novel and familiar complex stimuli following a single trial, exhibiting rapid plasticity as a result of single-trial learning.

3.2 Results

3.2.1 Task paradigm and behavioral results

We recorded single neuron activity using microwires implanted in the human hippocampus-amygdala complex (Figure 3-1A,B; see Table 3-1 for electrode locations), while subjects performed a object learning and recognition task. The delay between the learning and the initial recognition period was approximately 30 min, during which time the subject performed a different, cognitively demanding task. During learning, subjects were shown 12 different visual images. Each image was presented once, randomly in one of four quadrants on a computer screen (Figure 3-1C). Subjects were instructed to remember both the identity and the position of the image(s) presented. During the recognition period, subjects saw either previously viewed (familiar) or new images (novel) presented at the center of the screen (Figure 3-1D). For each image, the subject was asked to indicate whether the stimulus was new (novel) or old (familiar). Note that the novelty of a stimulus is only defined by whether it has been seen before or not (contextual). No other attributes of the stimulus changed. For each image identified as familiar, the subject was also asked to identify the quadrant in which the stimulus was originally presented (spatial recollection). Subjects correctly identified, on average, 88.5 ± 2.8% of all familiar and novel items during recognition (Figure 3-5). Subjects correctly recalled the quadrant location for 49.5 ± 8.0% of the familiar stimuli.

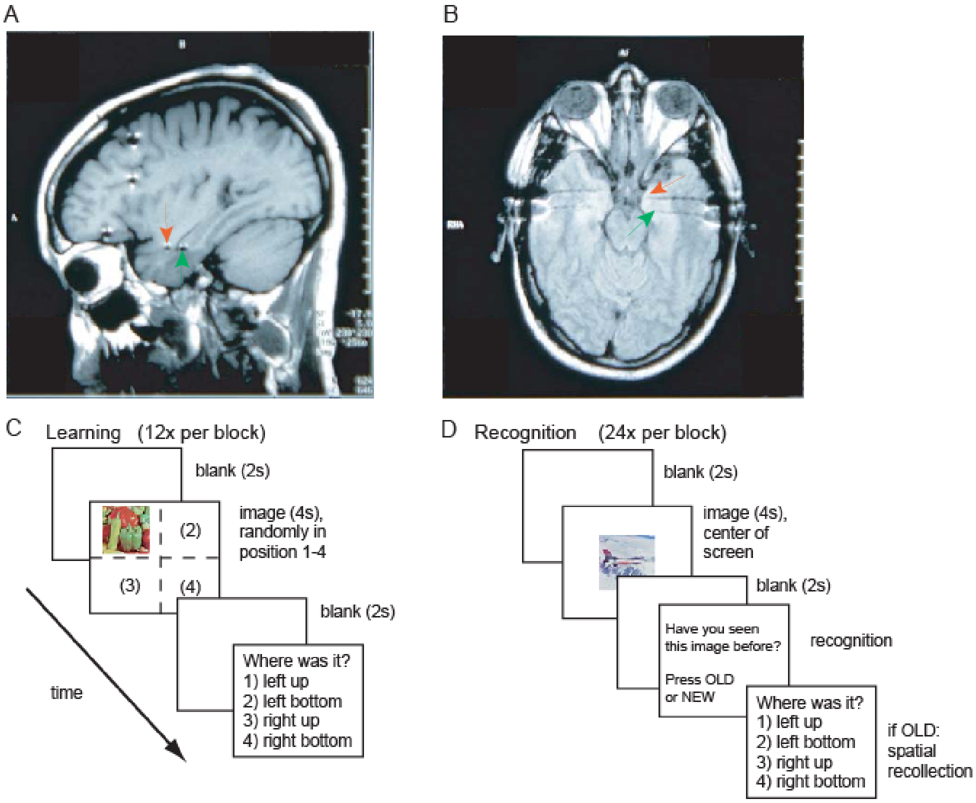

Figure 3-1. Electrode placement and task design.
(**A**) Saggital and (**B**) axial post-implantation structural MRI of one patient. The electrodes implanted in the amygdala (red) and the hippocampus (green) are indicated with arrowheads. The experiment has a learning (**C**) and a recognition block (**D**). Learning trials consisted of 12 images presented in one of 4 quadrants on the screen. 2 seconds after the stimulus was removed and replaced by a blank screen, the subject was asked to report in which quadrant the stimulus was presented. During recognition trials (30 min later), the subject was shown the 12 old images mixed with a set of 12 new images and asked to indicate whether the image had been viewed before (old) or not (new). After classifying an image as "old", the subject was also asked to indicate where the picture was during learning (spatial recognition).

3.2.2 Neural representations of single-trial learning, novelty, and familiarity

We analyzed the response of every neuron recorded (total number of neurons across all subjects = 244) during the baseline, stimulus presentation, and post-stimulus delay period. A neuron was considered selective if it exhibited an altered firing rate as a function of the stimulus (novel vs. familiar) ($p < 0.05$, bootstrap, see methods) and as a function of the task (learning vs. recognition phase). Neurons that increased their firing when exposed to novel vs. familiar stimuli were classified as signaling "novelty", whereas neurons that increased their firing to familiar stimuli were classified as signaling "familiarity" (Figure 3-2). Additionally, we classified responding neurons according to *when* they increased their firing: during the stimulus presentation of the stimulus or during the post-stimulus period (Figure 3-6D). Note that neurons signaling "novelty" increased their firing to new stimuli during the learning phase *and* also increased their firing to new stimuli presented during the recognition phase.

Are individual neurons capable of signaling that learning has occurred? If this is the case, then once the subject learns something about a stimulus (e.g., that it has been seen before) the firing properties of the neuron should reflect this knowledge. In our task, any knowledge about whether the specific stimulus presented has been seen before must result from a single trial experience. We indeed found subsets of neurons that showed enhanced or depressed firing rates on the second of two stimulus presentations, indicating the capacity for single-trial learning of familiarity. There are two different patterns of responses we observed that indicate single trial learning. One set of neurons ("familiarity detectors") exhibited enhanced firing when previously viewed stimuli were presented a second time during the recognition phase of the experiment. An example of this type of response is shown in Figure 3-2, where the neuron does not exhibit any appreciable response to the stimuli when first presented (Figure 3-2B) but when these same stimuli are presented a second time a dramatic increase in firing rate was observed (Figure 3-2D). These cells, which form a class of “familiarity” detectors, thus exhibit single-trial learning, exhibiting memory for a stimulus that was presented only one time. The other class of cells increased firing only for the first presentation of the stimulus ("novelty detectors”, see Figure 3-8 for an example). All told, 40 neurons consistently signaled either novelty (n = 23) or familiarity (n = 17) (Figure 3-3A,B). To characterize the firing differences of all neurons, we used two measures: i) average firing rate increase relative to baseline for new or old stimuli (depending on type of neuron), and ii) the average firing rate difference between new vs. old stimuli. For both measures, spikes were counted in the entire 6 s period following stimulus onset. We find that neurons increase firing on average 47% relative to baseline and the average firing difference between old vs. new stimuli is 76% (Figure 3-3C). The larger difference when comparing new vs. old firing indicates that in addition to increasing firing to the preferred stimulus (e.g., familiar), neurons decrease firing for the other stimulus type (e.g., novel). The large change in firing rate observed was induced by a single presentation of the stimulus and as such, these neurons provide a potential source for the rapid single-trial memory exhibited behaviorally by the subjects.

Do the observed neuronal changes reflect either a priming or a habituation response, or alternatively, do they reflect a form of long-term memory? If the former is the case, one would expect that, if presented with the same familiar stimuli (as well as new stimuli) 24 h later, the neuronal response to the familiar stimulus would be diminished. On the other hand, if the response reflects long-term memory, the altered firing pattern should still be observed the next day. To address this, we conducted a recognition session on the second and/or third day of recording, presenting subjects with the stimuli learned the previous day (4 sessions total in 3 patients) as well as a new set of stimuli. The time delay between the learning and the second recognition session was approximately 24 h (including one night of sleep). The behavioral performance (recognition and recollection) of these 3 patients did not differ significantly after a 30 min or 24 h time delay. Unfortunately, single-unit microwire recordings do not allow one to unambiguously determine whether the same individual neurons can be recorded on two sequential days. As such, we asked whether individual neurons, recorded 30 min or 24 hrs after the stimulus presentation, showed differences in firing to old vs. new stimuli. We then compared the average response strength per neuron after 30 min and 24 h time delays. We found that neither the average response strength per neuron nor the average increase in firing rate relative to baseline (Figure 3-3D) differed significantly for the two different time delays (2-way ANOVA with groups neuron type (Novelty/Familiarity) and time delay (30 min/24 h), $p < 0.05$). These neurons thus reflect the memory of the stimulus learned 24 h earlier but do not exhibit any further increases in firing rate (see discussion). The majority of neurons (37 of 40) exhibited a significant response within the first 2 s after stimulus onset (Figure 3-7C). Does the response strength decrease as a function of trial number? We found that neither novelty nor familiarity neurons significantly reduce their response strength over the duration of the experiment, during either learning or recognition (1-way ANOVA with block-nr and $p < 0.05$ reveals no significant effects for blocks of 1, 2, 3, or 4 trials). In addition, we found both types of neurons, familiarity and novelty detectors, in the amygdala as well as the hippocampus (Figure 3-6). However, the overall incidence of these neurons was significantly less in the amygdala when compared to the hippocampus: 19.7 ± 4.9% (n = 11) of all hippocampal neurons and 8.3 ± 2.7% (n = 12) of all amygdala neurons were classified as either novelty or familiarity neurons (n is number sessions, $p < 0.05$).

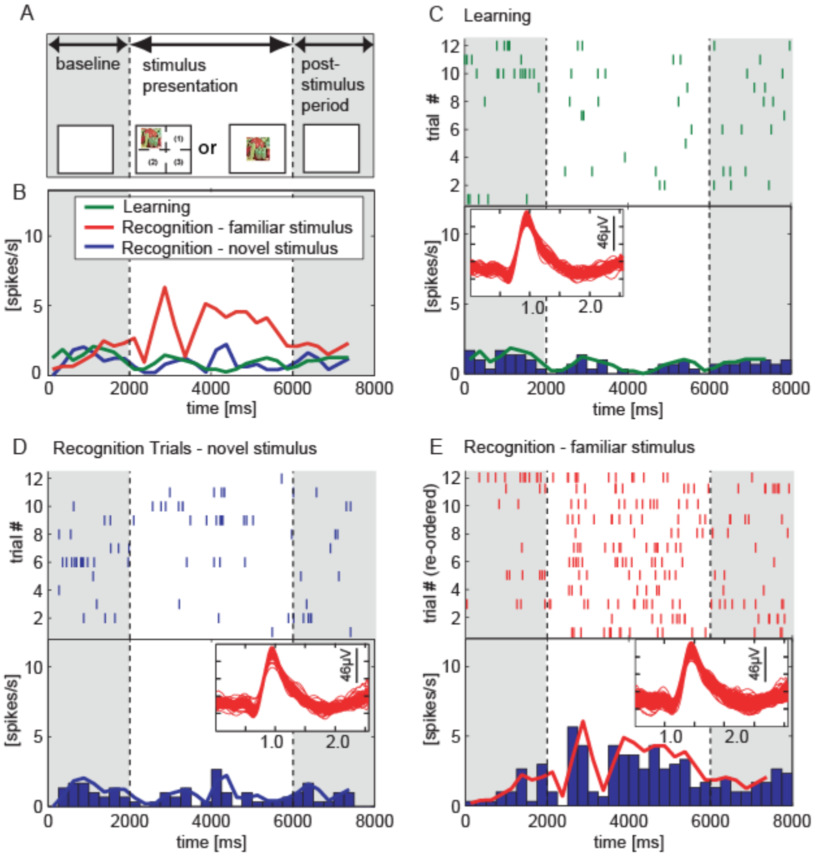

Figure 3-2. Example of a single hippocampal neuron during learning and recognition.

(**A**) Schematic representation of the experiment. Baseline (blank screen) from 0 to 2s, stimulus presentation from 2 to 6s, and post-stimulus period (blank screen) from 6 to 8s. (**B**) Average responses (spikes/sec). **(C-E)** The top portion of each figure shows the rasters depicting individual spikes. The stimulus was presented during the epoch defined by the dashed vertical lines. The bottom portion of each figure shows the binned histograms across all trials. Insets show overlays of all spike waveforms during the phase of the experiment depicted. (**C**) Responses during each learning trial. (**D**) Responses during the recognition phase for all new (not previously viewed) stimuli. (**E**) Responses during the recognition phase for all previously viewed (old) stimuli. Trials were randomly ordered during the experiment but are shown in (E) in the same order as during learning (C). This neuron increases its firing rate for stimuli seen before (E) but not for stimuli viewed for the first time (novel during both learning and recognition) (C and D).

Note that in C and E, the exact same visual stimuli are presented to the subject (12 images). When the stimuli are presented the first time (C), the neuron does not respond, whereas for the second presentation (E) it responds strongly.

3.2.3 Single neuron and population decoding

We analyzed how reliably these neurons can signal novelty or familiarity with an ideal-observer model. The model has access to the number of spikes fired during the 6 s period following stimulus onset. Using this information, a "decision" is made as to whether the subject is viewing a novel or a familiar stimulus. By parametrically varying the threshold (number of spikes) above which a single trial was considered novel or familiar, we conducted a receiver operator characteristic (ROC) analysis for each single neuron (Figure 3-7) and compared the true and false positives ratio at different thresholds. As a summary measure, we computed the area under the curve (Britten et al., 1996), which is the probability of correctly predicting whether the subject is currently viewing a novel or familiar stimulus (probability is between 0 and 1.0; 0.5 represents chance performance). We found that our neurons have an average single-trial single-neuron prediction probability of 0.72 ± 0.02. The population average is significantly above the chance level, which is determined by randomly shuffling the novel/familiar labels while keeping the spike trains intact. An observer that only has access to a single neuron can thus predict with on average 72% success whether a subject is seeing a familiar or novel stimulus.

How much information does the population of all recorded neurons contain about the familiarity of a stimulus? While ROC analysis quantifies how much information a single neuron conveys about the stimulus, it remains to be investigated how well this information can actually be decoded from a population of neurons on a single-trial basis. Single trials are highly variable and noisy. Does combining multiple neurons allow more accurate decoding than observing only a single neuron? Only if the signal or the noise were uncorrelated among neurons would one expect an improvement in decoding accuracy.

To address these questions, we used a simple population decoder which has access to all simultaneously recorded neurons that were previously identified as signaling novelty or familiarity. The decoder does not know the identity (novelty or familiarity detector) of the neurons. The only information available to the decoder is the number of spikes each neuron fired in the 6 s period following stimulus onset. The weighted sum (Figure 3-4A) of all spike counts is used to predict whether, for a given trial, an Old or New stimulus was presented. The weights are estimated from a set of labeled trials (Old or New) using multiple linear regression (see Methods).

We evaluated the properties of the classifier by considering only behaviorally correct recognition trials. For each recording session, we trained the classifier with all trials except a randomly chosen one (the "left-out trial"). Afterwards, we tested the classifier's performance by using it to predict whether the "left-out trial" was Old or New. Repeating this procedure many times for each session gives an accurate estimate of classifier performance (leave-one-out cross validation, see Methods). Additionally, we restricted the number of neurons that the classifier has access to. We found that the average single-trial classification performance increases from 67% correct for one neuron to 93% when 6 simultaneously recorded neurons are considered (Figure 3-4B, red line). A 1-way ANOVA reveals a significant effect of number of neurons (F = 6.6, p = 0.0001). Repeating the same procedure using randomly scrambled labels for the test trial results in a chance (50%) level performance (Figure 3-4B, black line). This analysis shows that it is beneficial for an "ideal" decoder to look at multiple neurons simultaneously. This indicates that the spikes fired by individual neurons signaling familiarity are uncorrelated in the sense that each of them contributes additional information that can be used to increase the accuracy of decoding.

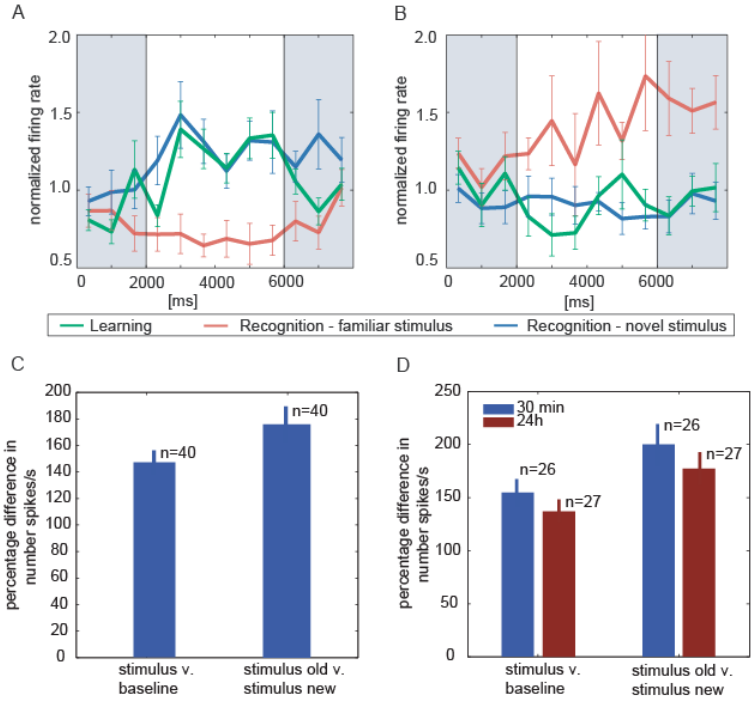

Figure 3-3. Population summary of all responding neurons.
Learning trials are in green, recognition old (familiar) trials are in red and recognition new (novel) trials are in blue. Neurons were classified according to which stimulus (old or new) they exhibited an increased firing rate and when they increase their firing (during either the stimulus or post-stimulus period or both). (**A,B**) Population average of all novelty (n=18) and familiarity neurons (n=10) which signal during the stimulus period. (**C**) Summary of response, quantified either as percentage firing rate difference during the 6 s post-stimulus period for old vs. new stimuli (right) or as percentage rate change relative to baseline (left). Note that the average rate increase of 75% is the result of a single stimulus exposure — the stimulus is learned after one trial. (**D**) Comparison of response for different time delays between learning and recognition. Shown is the average response strength with 30 min and 24 h delay. There is no significant difference in response strength for 30 min and 24 h delay (ttest, $p < 0.05$) nor is there a difference for novelty and familiarity detectors (not shown, 2-way ANOVA, $p < 0.05$). All errorbars are ±s.e. and n specifies number neurons.

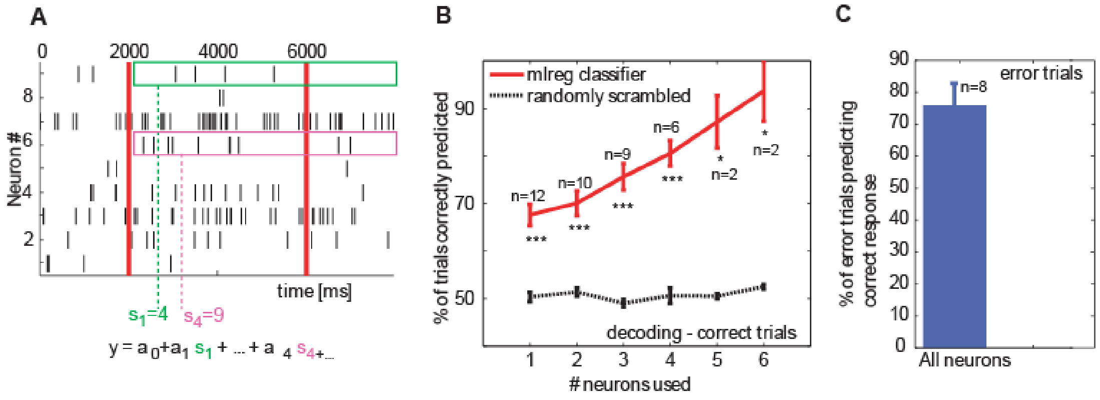

Figure 3-4. Population decoding from simultaneously recorded neurons. (**A**) Illustration of the decoding approach. Spikes of each neuron that signals novelty/familiarity (9 neurons in this example) are counted in the 6 s period following stimulus onset (first red line). Each neuron is assigned a weight determined by multiple linear regression. For a given trial, y predicts whether the trial is "Old" or "New". (**B**) Performance of the single trial-predictor as a function of number of simultaneously recorded neurons. Decoding performance increases when information from multiple recorded neurons is considered. The number of neurons used for decoding has a significant effect on performance of the decoder (1-way ANOVA, p < 0.001). n indicates the number of recording sessions. (**C**) The population decoder as trained in (B) applied to error trials. For 75% of all error trials in each session it predicts the correct response, that is, the neurons have better memory than the patient has behaviorally. The maximum number of available neurons is used for each session (mean number of neurons = 4.5). Only sessions that have at least 2 error trials are included (8 sessions). Errorbars are s.e. per session (n = 8) and the mean per session is significantly different from chance (p < 0.01).

3.2.4 Relations between neural responses and behavior

What is the relationship between the familiarity/novelty responses of individual neurons and the behavioral performance of the subject? The neuronal activity associated with behavioral errors allows us to answer this question. In our experiments, there were two kinds of error trials: i) recognition (novel vs. familiar) errors and ii) spatial recollection (which quadrant) errors. Below we investigate each type of error separately, beginning with spatial recollection errors.

There have been conflicting accounts as to whether retrieval-related activity in the hippocampus is related to familiarity recognition or recollection (Cameron et al., 2001; Stark and Squire, 2000; Yonelinas et al., 2002). One hypothesis states that the hippocampus is not involved in the retrieval of pure recognition memory, that is, memory without a recollective component. To investigate this issue, we examined neural activity during trials with successful recognition but failed recollection (spatial location of stimulus). We found that the subsequent successful spatial recollection is not required for neurons to exhibit familiarity responses. In fact we observe novelty and familiarity selective neurons in subjects who perform at chance levels for spatial recollection: In 4 (of 12) sessions, spatial recollection performance was at chance (21.7 ± 15.8%) and yet we found that 12 of the total 68 recorded units (17%) signaled novelty or familiarity. Thus, despite the fact that these patients weren't able to correctly recollect the spatial location in any of the trials, the same percentage of cells signaled novelty as in the other sessions. Also, for the sessions in which spatial recollection performance was above chance, we repeated our analysis including only trials associated with failed spatial recollection. Of the original 30 neurons, 26 remained significant (see Methods for details). We thus conclude that successful recollection is not required to observe a novelty/familiarity response in the hippocampus.

How is the neuronal activity during the stimulus presentation related to errors in recognition? Recognition of pictures is a highly automatic and reliable form of memory and subjects are usually very confident in their responses. This results in a small number of errors even when a large stimulus set is used, which has prevented analysis of such error trials in the past (Xiang and Brown, 1998). In our experiments, however, we record from many neurons simultaneously and can thus use a population decoder that allows accurate single-trial decoding (see discussion above). For each recording session, we trained the population decoder using all behaviorally successful trials. Afterwards, we used it to investigate what it would predict for the spiking activity observed during error trials. What might the population decoder (classifier) predict for an error trial? The classifier could: i) be at chance, ii) mimic the subject's (incorrect) response, or iii) predict the (correct, but not chosen) response. Each outcome would be informative: i) if it is at chance, these neurons do not contain any information about the stimulus on error trial; ii) if it predicts the behavioral response given, these neurons would likely represent some form of decision taken by the patient or motor planning activity related to the key the patient used to indicate the response; iii) if it predicts the correct response, these neurons would likely represent some form of high-fidelity memory. The third possibility is intriguing because it would suggest that these neurons exhibit "better memory" than the subject's behavioral response indicated. Since we are interested in the fraction of error trials per session that predict a certain outcome, we consider only sessions which contain at least 2 error trials (8 out of 12 sessions with a total of 33 error trials). For each session, we trained a classifier with all available neurons (on average 4.5) that signaled novelty/familiarity using all behaviorally correct trials and used it to predict the outcome of each error trial. We find that the classifier predicts the actual correct response for 75±7% of all error trials. The classifier is thus able to correctly predict the correct response in 75% of all cases even when the subject responded incorrectly (Figure 3-4C). These neurons thus have better memory than the patient exhibited behaviorally. This also suggests that the neuronal activity reported here does not represent some form of motor activity related to the subject's intended or actual response.

3.3 Discussion

3.3.1 Novelty and familiarity detectors in the human brain

We identified single neurons in the human hippocampus and amygdala that signal novelty or familiarity with an increase in firing rate. Several other groups have described non-human primate neurons that gradually (over many trials) decrease their response magnitude as specific stimuli become more familiar (Asaad et al., 1998; Fahy et al., 1993; Li et al., 1993; Rainer and Miller, 2000; Rolls et al., 1993). These types of neurons have also been observed in rodents (Berger et al., 1976; Vinogradova, 2001). The opposite pattern, neurons that increase their response magnitude for familiar stimuli, have largely not been observed in the primate brain (Fahy et al., 1993; Heit et al., 1990; Rolls et al., 1993; Xiang and Brown, 1998), and only rarely in humans (Fried et al., 1997). Also, studies investigating the relative proportion of novelty/familiarity- selective neurons in different areas of the MTL have usually failed to find any such neurons in the non-human primate hippocampus (Riches et al., 1991; Xiang and Brown, 1998) or, in one case, found only a very small proportion of such cells (Rolls et al., 1993). In contrast, we found a large proportion (17%) of familiarity/novelty-sensitive neurons, with an approximately equal number of neurons that increased firing for novelty or familiarity in the human hippocampus and amygdala. It has been speculated that the apparent absence of novelty/familiarity neurons in the primate hippocampus can be attributed to the lack of a spatial component in the tasks used (Riches et al., 1991; Xiang and Brown, 1998). To address this point, we used a non-spatial (old/new) and spatial recollective component in our task and find that the responses observed do not depend on successful spatial recollection. Another crucial difference is the behavioral task. Our task consists of a learning and recognition block with an interposed time delay. During the delay, other tasks are conducted. Others have used a serial recognition task where learning and recognition trials are intermixed and as such, there is no time delay that would permit a diversion of cognitive resources. It is possible that the emergence of the neuronal response requires time to develop. In our experiments, the firing rate increase can be observed after an initial delay of 30 min and remains equally strong for at least 24 h. This indicates that these neurons represent some form of long-term memory. Also note that the response strength does not increase further between 30 min and 24 h delays. The ability to correlate neuronal responses with human behavior may also be critical: we used an abstract task that can be rapidly learned thus facilitating the detection of these rapidly changing neuronal responses. In contrast, in non-human primates a simple associative memory task can take many trials for animals to reach criterion and learning-induced changes in hippocampal activity show a similar prolonged temporal profile (Wirth et al., 2003).

Could it be that the different findings are caused by eye movements? Most primate studies require the animal to fixate. In our experiments, subjects are free to move their eyes as they like. This is to make the task as natural as possible. Owing to clinical constraints, we were unable to record eye movements but there are several pieces of evidence which argue that eye movements cannot explain our results. The first few fixations made on any picture are mostly dominated by the statistics of the stimulus and do not change as a function of the familiarity of the stimulus (Noton and Stark, 1971). Also, a previous study of human MTL neurons found no influence of the fixated location of the picture on the visual response properties (Kreiman et al., 2002).

Others have reported that some neurons in the human MTL (Kreiman et al., 2000a) and the primate cortex (Li et al., 1993) are sharply tuned to the visual category of stimuli. Here, we used stimuli from many different visual categories (e.g., planes, cars, bottles, animals, mountains, people, computers, cameras, houses, books, chairs, and trucks) with one example per category. While the small stimulus set required for this kind of memory experiment prevents us from testing large numbers of stimuli from different categories, the response observed is invariant to at least a majority of the visual categories we have used. Thus, the neurons we describe here are capable of signaling the familiarity of the stimulus regardless of its visual category. One possibility is that the neurons preserve their tuning to categories and additionally increase or decrease their firing to indicate familiarity in an additive way. If this were the case, we would only detect broadly tuned units because narrowly tuned units would respond to a very limited set of stimuli. The neuronal responses we describe could thus serve as "general" novelty detectors that serve to establish the significance of behavioral stimuli during the acquisition of new or consolidation of existing memories (Lisman and Otmakhova, 2001).

Recognition and recollection are two largely distinct memory processes. Here we study recognition memory, but to allow a comparison with earlier human studies of recall/recollection we have included a spatial recollective component. Importantly we find that the response to the second presentation of the stimulus does not depend on whether spatial recollection is successful. This is in agreement with an earlier study of recollective memory which found that recall success is not correlated with the response of hippocampal neurons (Cameron et al., 2001). Also note that (Cameron et al., 2001) used the same stimuli many times during learning, so that the resulting neuronal changes cannot be related to any specific stimulus presentation. Similar studies of associative memory in the monkey hippocampus (Wirth et al., 2003; Yanike et al., 2004) are also complicated by this issue: stimuli were presented a large (10–30) number of times in order for the monkey to achieve behavioral criterion. These studies generally find that hippocampal neurons only change their response after many learning trials and thus seem to represent some form of "well learned" information. In contrast, in our study of human MTL neurons we use a single-trial learning paradigm that reveals that neurons are capable of rapid, single-trial plasticity.

3.3.2 Neurons that remember better than subjects

The finding that the neuronal activity during a majority of the error trials predicts the correct response represents an interesting disassociation between behavior and neuronal activity. In theory, an error could occur because the subject did not pay attention (not see the stimulus), accidentally pressed the wrong button, or because the subject did not remember the image correctly. Since the population decoder was not at chance levels for error trials, the first possibility can be excluded. Whether the subject accidentally pressed the wrong button or did not remember the image correctly cannot be determined from the available data. However, given the generally very high performance in the task and the absence of pressure to respond fast, it is unlikely that a majority of the error trials are caused by accidental wrong responses. If one examines the successful recognition trials exclusively, one might conclude that the neuronal responses represent the outcome of the decision taken (Old or New) or a consequence of that decision, e.g., planning and/or pre- or post-motor activity. If this were the case, however, activity during error trials would have to predict the response that was actually observed. However, we observed the opposite: activity during error trials predicts the correct response. We thus conclude that the neurons reported here represent some form of memory. In addition, the proportion of trials correctly identified by the neuronal responses is higher than what we observed behaviorally. Our data do not address at what point in the circuit the accurate neuronal responses on error trials fail to translate into correct behavioral responses. However, it is likely that information from multiple brain areas must be integrated to decide about the novelty of a stimulus. Any system of this nature requires an internal threshold for what is considered sufficient cumulative evidence for a stimulus to be classified as familiar. One could thus imagine situations where some brain areas provide input indicating familiarity but the cumulative evidence does not pass this threshold. Such a system would be maximally robust because it integrates multiple sources of information, perhaps trusting some more than others (Pouget et al., 2003). While it seems puzzling to have neurons that have better memory than is behaviorally observable, it makes sense in light of resistance to noise and erroneous transmission.

It has previously been observed that the average firing rate of some MTL neurons differs for successful vs. non-successful retrieval (Fried et al., 2002; Fried et al., 1997). However, in these studies, activity of the same neuron was not recorded during learning and it has thus remained impossible to determine whether these neurons changed their firing as a function of previous stimulus exposure or as a function of the task. In contrast, here we demonstrate that these changes result from a single stimulus exposure.

3.3.3 Relationship to fMRI and ERP findings

It has proven difficult to find human MTL fMRI activity correlated with behavioral success in recognition memory tasks (Manns et al., 2003; Stark and Squire, 2000). Using single-unit recordings we find evidence for the coexistence of novelty and familiarity cells recorded at the same time in the same brain region. On half of all macroelectrodes (18 of 36), we detected both novelty and familiarity neurons. On 2 of 6 microwires with more than one novelty/familiarity neuron both types were found. Since fMRI methods have limited spatial and temporal resolution and often rely on subtractive techniques, it is likely that the presence of both classes of neurons prevented their detection (Logothetis et al., 2001). The coexistence of MTL neurons that signal novelty or familiarity is likely an important feature used in establishing the significance of environmental events during learning.

Scalp and intracranial event-related potentials (ERP) recorded during serial recognition tasks have revealed a prominent potential (P300) to novel as well as target stimulus items (McCarthy et al., 1989; Sutton et al., 1965). That is, there is a potential to both novel as well as familiar (task relevant) items, but not to distractors. In hippocampal lesion patients it has been observed that the P3a component of the P300 is reduced (Knight, 1996). While we did not record ERPs in this study, the P300 response has been observed previously with intracranial electrodes in similar locations (McCarthy et al., 1989). It is thus of interest to note that the identified subpopulations of novelty and familiarity neurons we identified here could contribute to the P300.

3.3.4 Interaction with other brain systems

What is driving the response of these neurons? Neurons from multiple other brain areas can signal novelty or, more generally, the behavioral relevance of stimuli encountered in the environment. These include noradrenergic neurons in the locus coeruleus, cholinergic neurons in the basal forebrain as well as dopaminergic neurons in the midbrain (see (Schultz and Dickinson, 2000) for a review). Their response to novel events habituates with brief delays, evidence for short-term memory. Common to all these areas is the modulatory nature of their output — it is thus unlikely that their output is sufficient to account for the MTL responses we observe. These modulatory systems are known to regulate the strength of hippocampal-dependent learning, however (Frey et al., 1990; Neuman and Harley, 1983; Williams and Johnston, 1988), raising the possibility that the rapid plasticity we describe is related to the simultaneous release of neuromodulators that help induce long-lasting memories.

It is well known that animal behavior can be modified by a single exposure to a relevant stimulus (Sokolov, 1963). One instance of such memory is episodic memory, which is, by definition, memory of a single experience (Tulving et al., 1996). Other instances of single-trial learning include object recognition (Standing et al., 1970), spatial learning, and food caching (Clayton et al., 2001). In contrast, other forms of learning, like classical conditioning or rule learning (Wirth et al., 2003), require many learning trials. The neurons that underlie or participate in the rapid behavioral plasticity have, for the most part, evaded detection. Here we find that MTL neurons exhibit remarkable plasticity: a single exposure to a stimulus was sufficient to induce a dramatic and significant change in the spiking pattern. The observation of single-trial learning in MTL neurons indicates that, at least in principle, the rapid learning that human subjects exhibit has an electrophysiological correlate that occurs at the level of individual neurons.

3.4 Experimental procedures

3.4.1 Subjects and electrophysiology

Subjects were 6 patients (3 male, 3 female; mean age 37.5 ± 5.5 years; all native English speakers) diagnosed with drug-resistant temporal lobe epilepsy and implanted with intracranial depth electrodes to record intracranial EEG and single-unit activity. Patients underwent stereotactic placement of hybrid depth electrodes containing both clinical field potential contacts and microwire (50 µm) single-unit contacts, as described by (Fried et al., 1999). Briefly, electrodes were placed using orthogonal trajectories through the dorsolateral cortex, with the tip of the electrode targeting the amygdala, anterior hippocampus, orbitofrontal region, supplementary motor area, or anterior cingulate gyrus. The commercially available electrodes (Behnke hybrid depth electrode, Adtech Inc, Racine, MN), contain 4–6 platinumiridium 5 mm long circular electrodes, with a hollow center. After insertion of the electrode in the target, the inner cannula was removed and a bundle of microwires was passed through the center of the electrode, extending 5 mm beyond the tip of the electrode in a “flower spray” design. The electrodes were secured in place via a skull anchor bolt. All electrodes were placed based on clinical criteria alone. Patients were recruited for the research study after surgery was completed and EEG monitoring was initiated. Participation was voluntary and patients could withdraw from the study at any time. Informed consent was obtained and the protocol was approved by the Institutional Review Boards of the Huntington Memorial Hospital and the California Institute of Technology. For further details regarding the electrophysiological recordings, please see the supplemental material.

3.4.2 Data analysis

Spikes were sorted with a template-matching method (Rutishauser et al., 2006b). Only well-separated single neurons were used (see supplemental methods for details). We used a nonparametric bootstrap statistical test (Efron and Tibshirani, 1993) to assess significance at $p < 0.05$ (see supplement for discussion why not a t-test). To determine whether a neuron responds to new or old stimuli we compared the number of spikes fired for old vs. new stimuli during the stimulus on (4 s) and the post stimulus (2 s) period. For bootstrapping, 10,000 randomly re-sampled (with replacement) sets of spike counts were generated and tested for equality of means (Efron and Tibshirani, 1993). A second statistical test was performed to determine whether the firing of a neuron between old stimuli during recognition and all stimuli during learning (which are, by definition, new) was different. Only if both statistical tests were passed with $p < 0.05$ was the neuron determined to function as a novelty or familiarity detector. We randomly shuffled the start/endpoints of trials (in time) while keeping everything else the same to establish chance performance for this statistical procedure. We repeated this procedure 10 times and found a chance performance of 4.4% of all neurons (Figure 3-6D). Error trials during learning (incorrect position) and recognition (New/Old wrong) were excluded from this analysis.

All errors are standard error (s.e.), unless noted otherwise.

3.4.3 Population analysis

To quantify how well we were able to decode information about the novelty of the stimulus for a single trial, we used a population decoder. This also allowed us to analyze whether and how the decoding performance depends on the number of simultaneously recorded neurons. We used a simple weighted sum classifier of the form $y = a_0 + a_1 s_1 + ... + a_n s_n$, where s_x represents the number of spikes in the 6 s period following stimulus onset for neuron x, and a_x is the weight of this neuron. The weights are determined from labeled training data using multiple linear regressions (Johnson and Wichern, 2002). The label y is either set to 1 (New) or -1 (Old). Only neurons which were previously found to be signaling novelty/familiarity were considered for this analysis.

For verification purposes, we trained the classifier on behaviorally correct trials using leave-one-out cross validation. The performance of this classifier was then verified by evaluating its prediction for the left-out trial. Repeating this procedure many times gives an accurate estimate of the true performance of the estimator. We repeated the same analysis by restricting the number of neurons the classifier had access to. In cases where more neurons were available than the classifier could consider, a random subset of the available neurons was chosen and the procedure was repeated multiple times so that all possible combinations were explored. All error bars in the population analysis are given as s.e., with n being the number of sessions, to demonstrate the variance over multiple patients and recording sessions rather than over multiple neurons.

3.5 Supplementary material

3.5.1 Electrophysiology

Recordings were conducted using a commercial (Neuralynx Inc, Arizona) acquisition system with specially designed, head-mounted pre-amplifiers. Signals were filtered and amplified by hardware amplifiers before acquisition. The frequency band acquired was either 1–9000Hz or 300–9000Hz, depending on the noise levels. Great care was taken to eliminate noise sources. This included using batteries to power the amplifiers, experimental computers, IV machines and heartbeat monitors. Recordings commenced the second day after surgery and continued for 2–4 days for about 1 hour per day. The experiments reported in this paper were done on two consecutive days for all 6 patients (12 sessions in total).

The amplifier gain settings, set individually for each channel, were typically in the range of 20000–35000 with an additional A/D gain of 4 (2 in some cases). The raw data was sampled at 25 kHz and written to disk for later filtering (300–3000Hz bandpass), spike detection, and spike sorting. Spikes were detected using a local energy method (Bankman et al., 1993) and sorted by a template-matching method (Rutishauser et al., 2006b). Great care was taken to ensure that the single units used passed stringent statistical tests (projection test (Pouzat et al., 2002)) . It is thus likely that we underestimate the number of single units present. Only neurons with mean firing rates ≥ 0.25 Hz were included in the analysis.

3.5.2 Electrodes

In each macroelectrode, 8 microwires were inserted (Fried et al., 1999). One microwire was used as local ground and the other 7 were used for recordings. The impedance of a total of 56 microwires in 2 patients was, on average, 135 ± 62kOhm (± s.d.) with a range of 38–245 kOhm.

Electrode position was determined by an experienced neurosurgeon (ANM) from structural MRIs taken 1 day after electrode implantation on a clinical 1.5 Tesla MRI system (Toshiba, Inc). We always recorded from 3 macroelectrodes simultaneously: left/right hippocampus and either left or right amygdala (total of 24 channels, 8 channels for each macroelectrode with 1 channel used as local ground).

3.5.3 Localization of electrodes

We localized the position of each macroelectrode in a standardized stereotactic coordinate system (Talairach) in a subset of 4 patients for which high-resolution structural MRIs were available (Table 3-1). We transformed each structural 1.5 T MRI scan to Talairach space by manually identifying the anterior and posterior commisure as well as the anterior, posterior, superior, and inferior points of the cortex. We used BrainVoyager (Brain Innovation B.V.) for this procedure. After co-registration we identified the Talairach coordinates by finding a consensus from the different structural scans. For each patient, we performed 4 different scans with 1x1 mm resolution in the following plane: coronal, sagittal, and 2 axial with different pulse sequences (2TW and FLAIR).

Patient	Amygdala (r/l)	Hippocampus (r/l)
P2	-20,1,-19 26,-2,-20	-26,-9,-11 28,-11,-20
P3	-20,-3,-15 18,-4,-15	-23,-13,-12 33,-12,-16
P4	-19,4,-26 28,7,-26	-21,-9,-25 27,-7,-26
P6	-23,-2,-14 23,-6,-13	-25,-13,-12 29,-18,-12

Table 3-1. Electrode position in stereotactic coordinates (Talairach)

3.5.4 Implementation of behavioral task

The task was implemented using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) in Matlab (Mathworks Inc) and ran on a notebook PC placed directly in front of the patient. Distance to the screen was approximately 50 cm and the screen was approximately 30 by 23 degrees of visual angle. The pictures used were approximately 9 by 9 degrees. Specially marked keys ("New", "Old") on the keyboard were used to acquire subject responses. We chose to use natural pictures as stimuli rather than words or faces because it has been shown that pictures reliably result in bilateral fMRI activation of the MTL, whereas words and faces result in primarily unilateral (left) activation (Kelley et al., 1998).

3.5.5 Data analysis

We conducted all statistical analysis using bootstrap tests (see Methods of main text). To be thorough, we repeated the same analysis using a two-tailed t-test ($p < 0.05$) and found reasonable overlap with the pool of neurons determined to signal novelty or familiarity using the above bootstrap method. We found, however, that using the t-test more neurons were classified as novelty/familiarity detectors, some of which (by visual inspection) were likely false positives. Also, the chance performance determined by random shuffling was high (~ 10%). We thus decided to exclusively use the bootstrap method since it yielded the most consistent and conservative results. Post-stimulus histograms (PSTH) were created by binning the number of spikes into 250 ms bins. To convert the PSTH to an instantaneous firing rate, a Gaussian kernel with standard deviation = 300 ms was used to smooth the binned representation. Population averages (Figure 3-3C and D) were constructed by averaging the normalized firing rate of each neuron. Firing rates were normalized to the mean firing rate of the neuron during the particular part of the experiment (learning block or recognition block). We averaged the raw normalized PSTH of each neuron (above PSTH smoothing is not applied to normalized PSTH of each neuron, nor to the population average).

3.5.6 Spatial recollection analysis

To investigate whether the response observed during familiarity/novelty recognition required later successful spatial recollection we conducted additional data analyses. Based on several pieces of evidence we find that successful spatial recollection is not required for emergence of novelty/familiarity cells: i) In 4/12 sessions spatial recollection performance was at chance levels (mean 21.7 ± 7.9%) and yet we found that 14.8% of the recorded neurons in these sessions signaled novelty/familiarity during recognition and showed single-trial learning. This percentage is remarkably similar to the percentage of all neurons that signal novelty or familiarity (Figure 3-6). Thus despite the fact that these patients weren't able to correctly recollect the spatial location in any of the trials the same percentage of cells signaled novelty as in the other sessions. ii) In the 8 sessions with above-chance spatial recollection performance (mean 63.91±7.02%), 28 neurons were found (17.2% of all recorded neurons). Repeating the analysis as described above, but only including trials with successful recollection, results in 26 of those 30 neurons remained significant. The number of selective neurons is thus decreased if only trials with successful spatial recollection are included and error trials are thus contributing valuable information. iii) In 9 sessions there were at least 4 spatial recollection error trials (correctly recognized as Old, but location wrong). Considering only these error trials (disregarding trials with correctly remembered locations), 20 out of originally 26 (77%) neurons remain significant. A high proportion of all originally identified neurons thus signal novelty/familiarity even in the absence of successful spatial recollection.

3.5.7 Single-neuron ROC analysis

To determine how well the response of a single neuron during recognition predicts whether the patient is currently viewing a familiar or novel stimulus we conducted an ROC (receiver-operator characteristic) analysis (Britten et al., 1996; Green and Swets, 1966). This analysis assumes that an ideal observer, who only has access to the number of spikes fired by a single neuron during the presentation of the stimulus and the post-stimulus period (6 s period), should be able to correctly classify individual neurons as signifying novelty vs. familiarity. Only trials where the subject correctly replied with "Old" or "New" were used for this analysis (this was 88.5% of all trials). We quantify the ROC for each neuron recorded by integrating the area under the curve (AUC) of the ROC. This number equals the probability of correctly predicting, on a single-trial basis, whether the “subject” has viewed a novel or familiar stimulus. An AUC of 0.5 equals chance. We confirmed the validity of our analysis by randomly shuffling the labels "New" and "Old" while leaving the spike trains intact. Repeating this procedure 50 times for each neuron resulted in AUC values clustered around 0.5 (Figure 3-7A,B).

We conducted this ROC analysis without preclassifying neurons into novelty/familiarity detectors. This results in a cluster of neurons with a prediction probability significantly below 0.5 and one significantly above 0.5. Since Old/New is a binary state, this contributes equal information and we thus subtracted 1-x for all ROC values x < 0.5 to get an unimodel distribution, as shown in Figure 3-7A.

We repeated the analysis above for different time bins following stimulus onsets (step size 500 ms), e.g. counting spikes in bins 2000–2500 ms, 2000–3000 ms, 2000–3500 ms, etc. Using this analysis we defined for each neuron when its ROC value became significantly above chance the first time (Figure 3-7C).

3.5.8 Epileptic vs. non-epileptic tissue

One concern regarding the neurons described in this paper is that they were recorded from epilepsy patients. To confirm that our findings are also valid for "healthy" tissue, we repeated our analysis but excluded all electrodes which were in tissue that was later resected

(Table 3-2). Of the total 244 recorded neurons, 138 were in tissue which was not resected. Of these 138 neurons, 22 signalled novelty or familiarity (15.9%).

Patient	**Side of temporal lobe lobectomy**
P1	left
P2	left
P3	right
P4	left
P5	left
P6	right

Table 3-2. Location of resected tissue (temporal lobe lobectomy in each case).

3.6 Supplementary figures

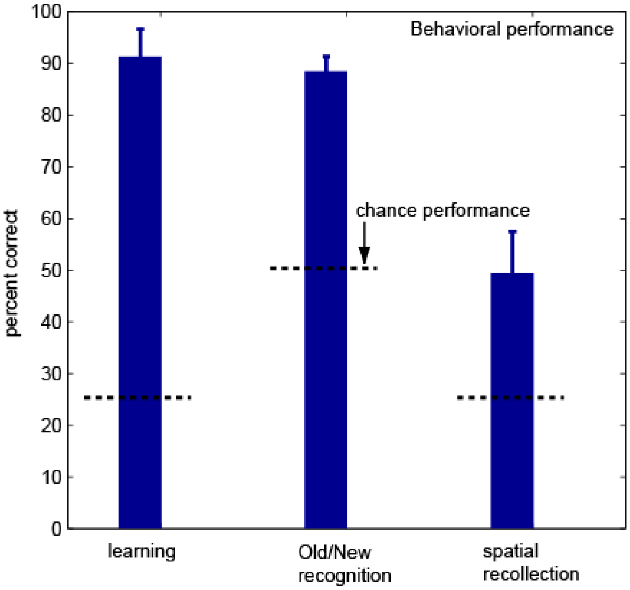

Figure 3-5. Behavioral performance of all subjects.
Recognition performance (Old/New) was close to 90% (chance 50%) whereas spatial recollection, in which the subject reports the quadrant in which the images was presented for all images classified as "Old", was 49%. All performance levels are significantly different from chance ($p < 0.05$).

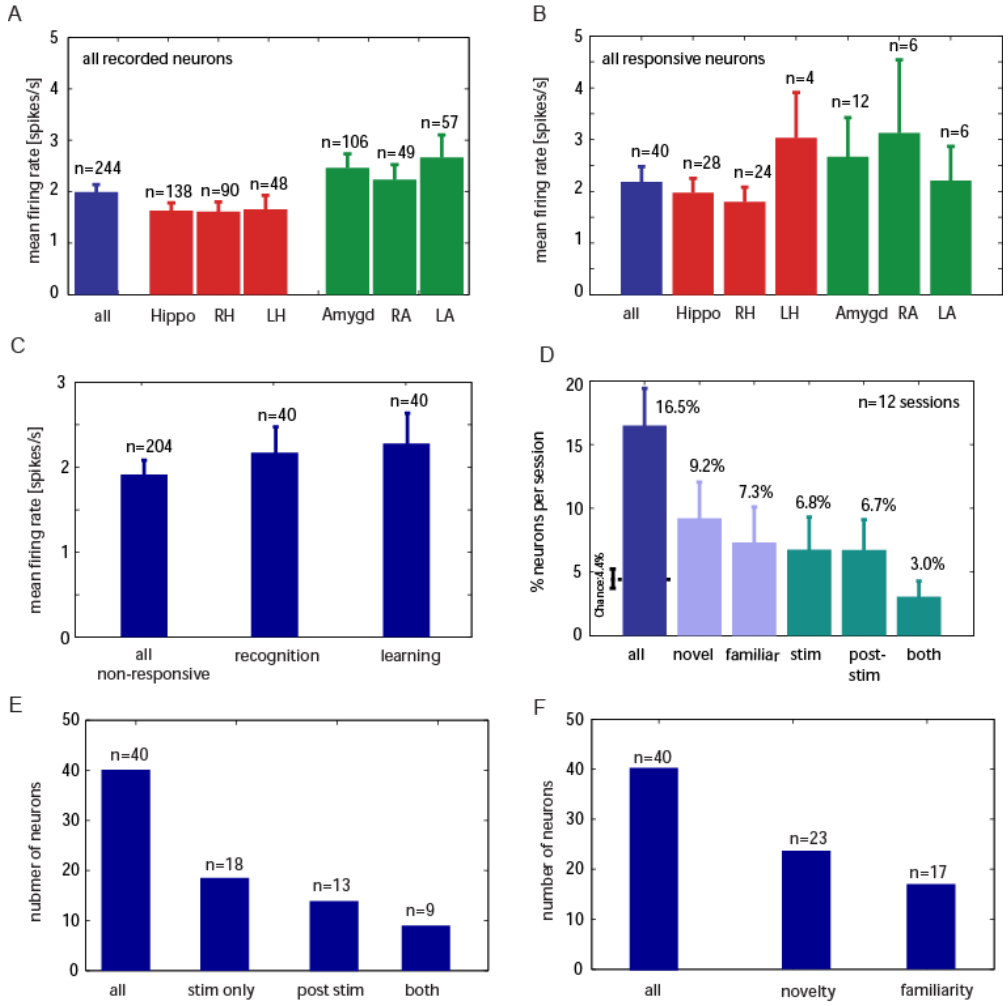

Figure 3-6. Population statistics for all neurons.

(**A**) as well as the subset of significantly responsive neurons (**B-F**). (**A**) The mean firing rates of all neurons recorded (n = 244) was 1.96 ± 0.14 Hz. The mean firing rate was not significantly different among different brain areas (1-way ANOVA, $p < 0.05$). (**B**) The mean firing rate of all responsive neurons (n = 40) was 2.17 ± 0.30 Hz, with no significant difference amongst different brain areas. (**C**) The mean firing rate for novelty and familiarity neurons was not statistically different from all other neurons recorded (1-way ANOVA, $p < 0.05$) during either learning or recognition. (**D**) Considering all sessions, 16.5% of all recorded neurons indicated novelty or familiarity in every session (2 sessions each in 6 patients). There were slightly more novelty neurons (9.2%/per session) than familiarity neurons (7.3%/per session). (**E**) We found a total of 40 significant neurons, 18 of which signaled during the stimulus period, 13 during the post-stimulus period, and 9 during both; (**F**) There were 24 novelty and 18 familiarity neurons.

Abbreviations: RH, right hippocampus; RA, right amygdala, LH, left hippocampus; LA, left amygdala; hippo, hippocampus; amygd, amygdala. All error bars are ±s.e and n always specifies number of neurons.

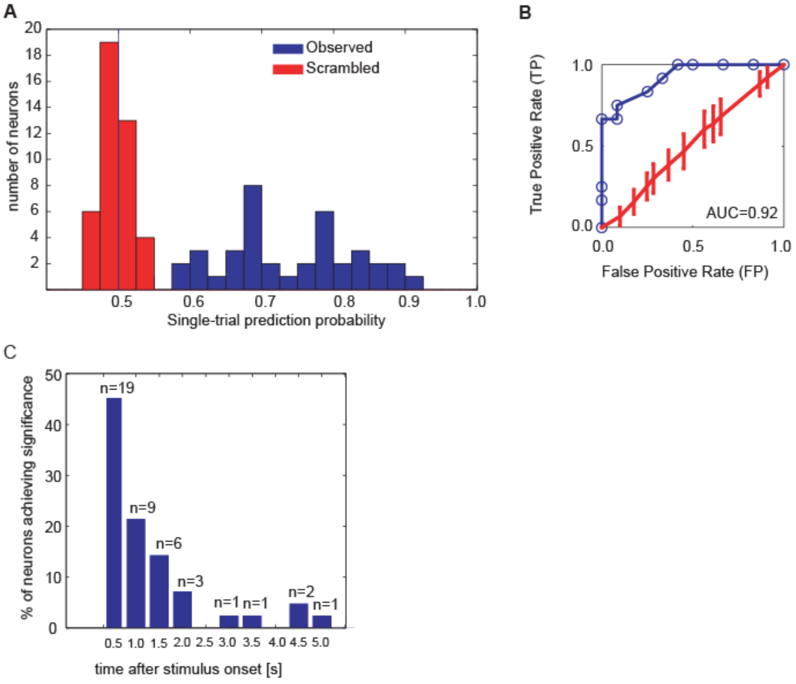

Figure 3-7. Single-neuron prediction probabilities.
(A) Histogram of the single-trial prediction probabilities for all 40 significant neurons. The mean probability was 0.72±0.02. The prediction probability is equal to the area under the curve of the ROC of each neuron and specifies the ratio of recognition trials in which novelty or familiarity is successfully predicted on a trial-by-trial basis by observing a single neuron. Randomly shuffling (scrambled) the spike counts of new and old trials results in a mean of 0.5 (red in A, error bars are s.d.). The ROC for the same neuron as shown in figure 2 is shown in (**B**) (blue = real trials, red = randomly shuffled). (**C**) Latency of response for all neurons. Shown are, for each time following stimulus onset, the percentage of neurons which became significant for the first time in this time bin.

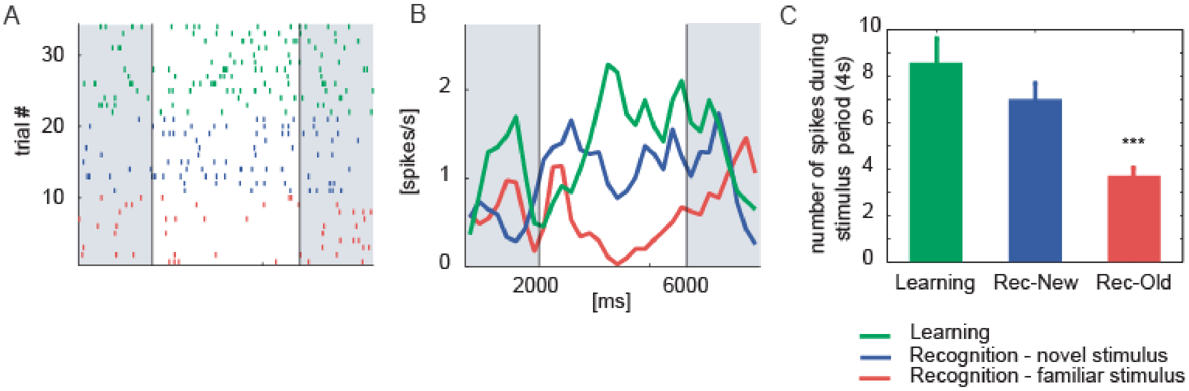

Figure 3-8. Example of a novelty-sensitive neuron.
Neuron which increases firing to novel stimuli during both learning and recognition. (A) Raster for all spikes during learning (green), recognition old (red), and recognition new (blue). (B) Histogram summarizing the response. Note the decrease to familiarity. (C) Comparison of the number of spikes fired during the 4 s stimulus period (white in B). The number of spikes fired for familiar items is significantly different from the number of spikes fired during learning and recognition of new items. ($p < .001$ for both comparisons, 1-way ANOVA with posthoc multiple comparison. n = 12 (number of trials)).

Chapter 4. Activity of human hippocampal and amygdala neurons during retrieval of declarative memories

4.1 Introduction[4]

Episodic memories allow us to remember not only whether we have seen something before but also where and when (contextual information). One of the defining features of an episodic memory is the combination of multiple pieces of experienced information into one unit of memory. An episodic memory is, by definition, an event that happened only once. Thus, the encoding of an episodic memory must be successful after a single experience. When we recall such a memory, we are vividly aware of the fact that we have personally experienced the facts (where, when) associated with it. This is in contrast to pure familiarity memory, which includes recognition, but not the "where" and "when" features. The MTL, which receives input from a wide variety of sensory and prefrontal areas, plays a crucial role in the acquisition and retrieval of recent episodic memories. Neurons in the primate MTL respond to a wide variety of stimulus attributes such as object identity (Heit et al., 1988; Kreiman et al., 2000a) and spatial location (Rolls, 1999). Similarly, the MTL is involved in the detection of novel stimuli (Knight, 1996; Xiang and Brown, 1998). Some neurons carry information about the familiarity or novelty of a stimulus (Rutishauser et al., 2006a; Viskontas et al., 2006) and are capable of changing that response after a single learning trial (Rutishauser et al., 2006a). The MTL, and in particular the

[4] The material in this chapter is based on Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2008). Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. Proc Natl Acad Sci U S A *105*, 329-334.

hippocampus, are thus ideally suited to combine information about the familiarity/novelty of a stimulus with other attributes such as the place and time of occurrence.

The successful recall of an experience depends on neuronal activity during acquisition, maintenance, and retrieval. The MTL plays a role in all three components. Here, we focus on the neuronal activity of individual neurons during retrieval. The MTL is crucially involved in the retrieval of previously acquired memories: brief local electrical stimulation of the human MTL during retrieval leads to severe retrieval deficits (Halgren et al., 1985). Two fundamental components of an episodic memory are whether the stimulus is familiar and if it is, whether information is available as to when and where the stimulus was previously experienced (e.g., recollection). How these components interact, however, is not clear. A key question is whether there are distinct anatomical structures involved in these two processes (familiarity vs. recollection).

Some have argued that the hippocampus is exclusively involved in the process of recollection but not familiarity (Eldridge et al., 2000; Yonelinas, 2001). Evidence from behavioral studies with lesion patients, however, seems to argue against this view (Manns et al., 2003; Stark et al., 2002; Wais et al., 2006). Rather than removing the capability of recollection while leaving recognition (familiarity) intact, hippocampal lesions cause a decrease in overall memory capacity rather than the loss of a specific function. Lesion studies, however, do not allow one to distinguish between acquisition vs. retrieval deficits.

Recollection of episodic memories is difficult to study in animals (but see (Hampton, 2001)) but can easily be assessed in humans. Recordings from humans offer the unique opportunity to observe neurons engaged in the acquisition and retrieval of episodic memories. We recorded from single neurons in the human hippocampus and amygdala during retrieval of episodic memories. We used a memory task that enabled us to determine whether a stimulus was only recognized as familiar or whether an attribute associated with the stimulus (the spatial location) could also be recollected. We hypothesized that the neuronal activity evoked by the presentation of a familiar stimulus would differ depending on whether the location of the stimulus would later be recollected successfully or not. We found that the neuronal activity contains information about both the familiarity and the recollective component of the memory.

4.2 Results

4.2.1 Behavior

During learning, subjects (see Table 4-1 for neuropsychological data) were shown 12 different pictures presented for 4 seconds each (Figure 4-1A). Subjects were asked to remember the pictures they had seen (recognition) and where they had seen them (position on the screen). After a delay of 30 min or 24 h, subjects were shown a sequence of 12 previously seen ("Old") and 12 entirely different ("New") pictures (Figure 4-1B). Subjects indicated whether they had seen the picture before and where the stimulus was when they saw it the first time. We refer to the true status of the stimulus as *Old* or *New* and the subject’s response as *Familiar* or *Novel*. With the exception of error trials the two terms are equivalent. Subjects remembered 90 ± 3% of all old stimuli and for 60 ± 5% of those they remembered the correct location (Figure 4-1C). Some subjects were not able to recollect the spatial location of the stimuli whereas others remembered the location of almost all stimuli. For each 30 min retrieval session, we determined whether the patient exhibited, on average, above chance (R^+) or at chance (R^-) spatial recollection and then calculated the behavioral performance separately (Figure 4-1D,E). Patients with good same-day spatial recollection performance (30 min R^+) remembered the spatial location of on average 77±6% (significantly different from 25% chance, $p < 0.05$, z-test) of stimuli they correctly recognized as familiar whereas at-chance patients (30 min R^-) recollected only 35±4% of stimuli (approaching but not achieving statistical significance, $p = 0.07$). There were thus two behavioral groups for the 30 min delay: one with good and one with poor recollection performance.

We also tested a subset of the subjects that had good recollection performance on the first day with an additional test 24 h later (4 subjects). Subjects saw a new set of pictures and were asked to remember them overnight. Overnight memory for the spatial location was good (66±1%, $p < 0.05$). All 3 behavioral groups (30 min R^+, 30 min R^-, 24 hr R^+) had good recognition performance (Figure 4-1E) that did not differ significantly between groups (ANOVA, $p = 0.24$). The FP rate was on average 7±3% and did not differ significantly between groups (ANOVA,$p = 0.37$).

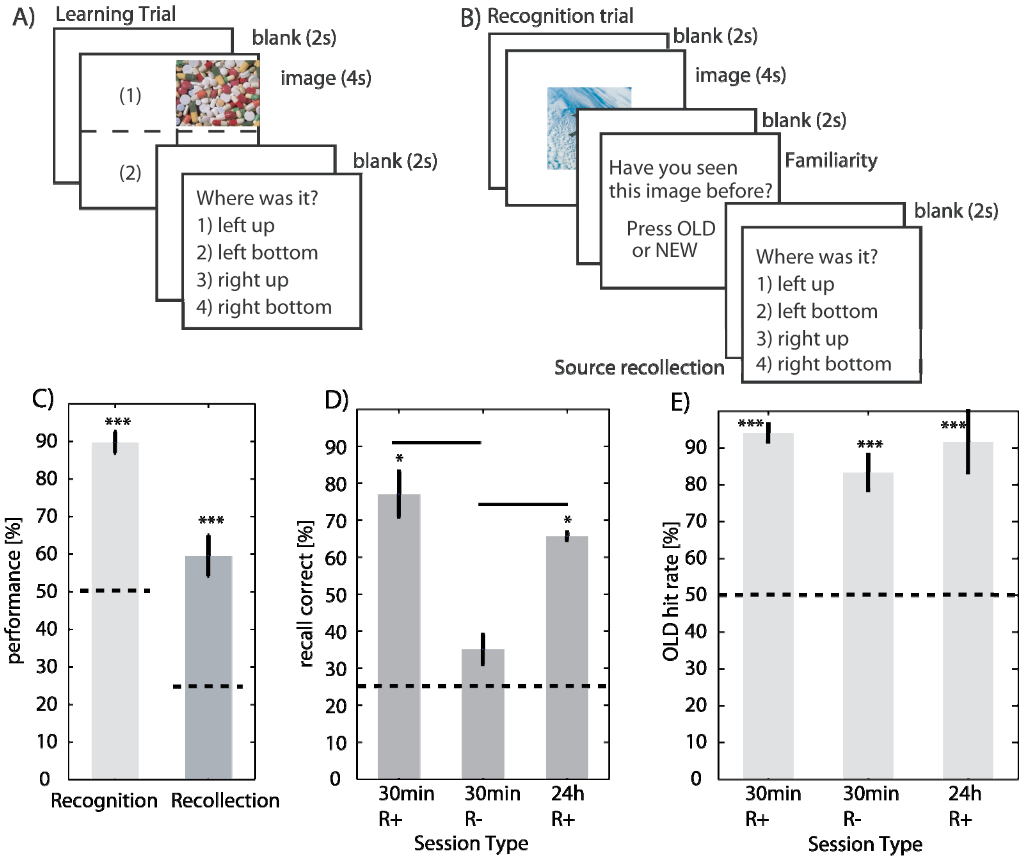

Figure 4-1. Experimental setup and behavioral performance.
The experiment consists of a learning (A) and retrieval (B) block. (C) Patients exhibited memory for both the pictures they had seen (recognition) as well as where they had seen them (recollection). n = 17 sessions. (D) Two different time delays were used: 30 min and 24 h. 30min delay sessions were separated into two groups according to whether recollection performance was above chance or not. (E) For all groups, patients had good recognition performance for old stimuli, regardless of whether they were able to successfully recollect the source. n = 7,5,4 sessions, respectively. Errors are ± s.e.m. Horizontal lines indicate chance performance. R^+ = above chance recollection, R^- at chance recollection.

4.2.2 Single-unit responses during retrieval

We recorded the activity of 412 well separated units in the hippocampus (n = 218) and amygdala (n = 194) in 17 recording sessions from 8 patients (24.24±11.51 neurons (±s.d.) per session). The mean firing rate of all neurons was 1.45±0.10 Hz and was not significantly different between the amygdala and the hippocampus (Figure 4-5A). For each neuron we determined whether its firing differed significantly in response to correctly recognized old vs. new stimuli. Note that "old" indicates that the subject has seen the image previously during the learning part of the experiment. Thus, the difference between a novel and old stimulus is only a single stimulus presentation (single-trial learning). We found a subset of neurons (114, 6.7±4.7 per session, see Table 4-2) that contained significant information about whether the stimulus was old or new. Because error trials were excluded for this analysis, the physical status (old or new) is equal to the perceived status (familiar or novel) of the stimulus. Neurons were classified as either familiarity (n = 37) or novelty detectors (n = 77) depending on the stimulus category for which their firing rate was higher (see methods). The analysis presented here is based on this subset of neurons. The mean firing rate of all significant neurons (1.6±0.2Hz, n=114) did not differ significantly from the neurons not classified as such (1.4±0.1Hz, n = 298). Similarly, the mean firing rate of neurons that increase firing in response to novel stimuli was not different from neurons that increase firing in response to old stimuli (Figure 4-5C,D).

The response of a neuron that increased firing for new stimuli is illustrated in Figure 4-2A–C. This neuron fired on average 1.1±0.2 spikes/s when a new stimulus was presented and only 0.6±0.1 spikes/s when a correctly recognized, old stimulus was presented (Figure 4-2C). Of the 10 old stimuli (2 were wrongly classified as novel and are excluded), 8 were later recollected whereas 2 were not. For the 8 later recollected items (R+) the neuron fired significantly less spikes than for the not recollected items (0.5±0.1 v. 0.9±0.3, $p < 0.05$, Figure 4-2C). Thus, this neuron fired fewer spikes for items which were both recollected and recognized than for items which were not recollected. We found a similar, but opposite pattern for neurons that increase their firing in response to old stimuli (see below). We thus hypothesized that these neurons represent a continuous gradient of memory strength: the stronger the memory, the more spikes that are fired by familiarity-detecting neurons (Figure 4-2D). Similarly, we hypothesized that the opposite relation would hold for novelty neurons: the fewer spikes, the stronger the memory.

We analyzed 3 groups of sessions separately: Same day with good recollection performance (30 min R^+), same day with at chance recollection performance (30 min R^-) and overnight with above-chance recollection (24 h R^+). Sessions were assigned to the 30 min R^+ or 30 min R^- groups based on behavioral performance. We hypothesized that if the neuronal firing evoked by the presentation of an old stimulus is purely determined by its familiarity, the neuronal firing should not differ between stimuli which were only recognized and stimuli which were also recollected. On the other hand, if there is a recollective component, then a difference in firing rate should only be observed for recording sessions in which the subject exhibited good recollection performance.

First we examined the novelty (Figure 4-2E) and familiarity neurons (Figure 4-2F) in the 30 min R+ group. The pre-stimulus baseline was on average 1.7±0.4 Hz (range 0.06–9.5) and 2.6±1.0 Hz (range 0.2–12.9) for novelty and familiarity neurons, respectively, and was not significantly different. Units responding to novel stimuli increased their firing rate on average by 58±5% relative to baseline. Similarly, units responding to old stimuli increased their firing by 41±8% during the second stimulus presentation. We divided the trials for repeated stimuli into two classes: stimuli that were later recollected (R+) and not recollected (R-). A within-neuron repeated measures ANOVA (factor trial type: new, R- or R+) revealed a significant effect of trial type for both novelty ($p < 1e-12$) as well as familiarity units ($p < 1e-6$). This test assumes that neurons respond independently from each other. For both types of units we performed two planned comparisons: i) New vs. R- and ii) R- vs. R+. For novelty neurons, the hypothesis was that the amount of neural activity would have the following relation: New > R- and R- > R+. For familiarity, the hypothesis was the opposite: New < R- and R- < R+ (Figure 4-2D). For novelty as well as familiarity neurons, each prediction proved to be significant (one-tailed t-test. Novelty: New vs. R- $t = 4.3$, $p < 1e-4$ and R- vs. R+ $t = 2.2$, $p = 0.01$. Familiarity: New vs. R- $t = -1.7$, $p = 0.05$ and R- vs. R+ $t = -2.0$, $p = 0.02$). Thus both novelty- and familiarity-detecting neurons signaled that a stimulus is repeated even in the absence of recollection (New vs. R-) and whether a stimulus was recollected or not (R- vs. R+).

The same analysis applied to the remaining groups (30 min R- and 24 h R+) revealed a significant main effect of trial type for novelty ($p < 1e-4$ and $p < 1e-5$, respectively) as well as familiarity neurons ($p < 0.001$ and $p < 0.001$, respectively). However, only the New vs. R- planned comparison was significant (Novelty: $p < 0.001$ and $p < 0.001$; Familiarity: $p < 0.001$ and $p < 0.001$) whereas the R- vs. R+ comparison was not significant for either group (Novelty: $p = 0.6$ and $p = 0.7$; Familiarity: $p = 0.68$ and 0.49). Thus, the activity of these units was different for new vs. old stimuli but the response to old items was indistinguishable for recollected vs. not recollected stimuli.

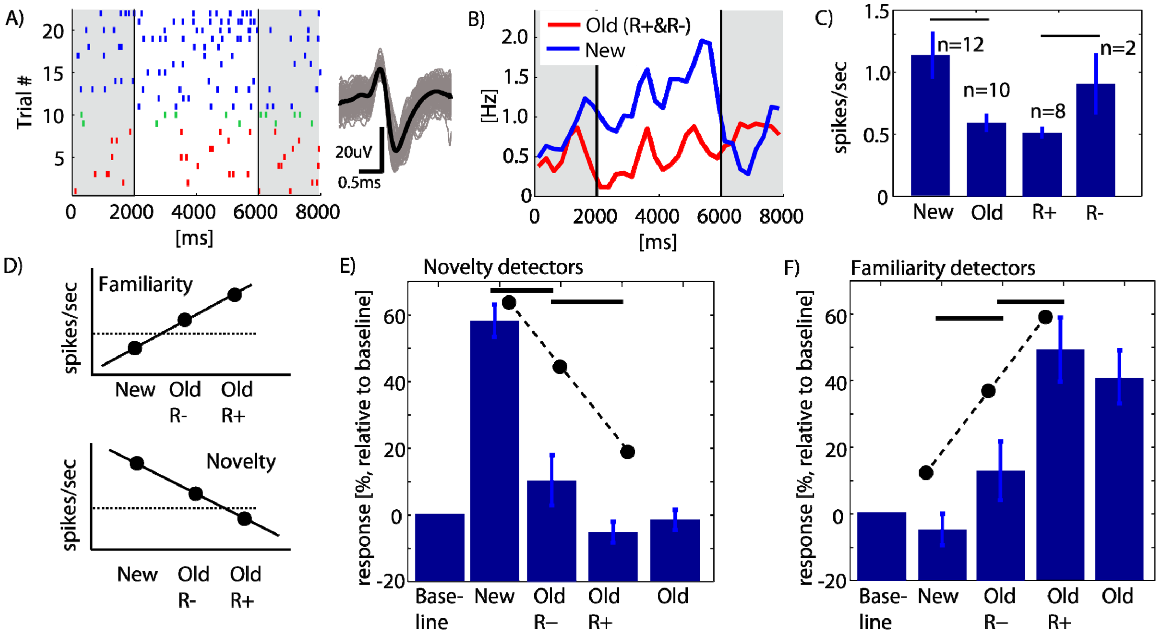

Figure 4-2. Single cell response during retrieval.
(A–C) Firing of a unit in the right hippocampus that increases its firing in response to new stimuli that were correctly recognized (*novelty detector*). (A) Raster of all trials during retrieval and the waveforms associated with every spike. Trials: New (blue), old and recollected (red, R+) and old and not recollected (green, R-). (B) PSTH. (C) Mean number of spikes after stimulus onset. Firing was significantly larger in response to new stimuli and the neuron fired more spikes in response to stimuli which were later not recollected compared to stimuli which were recollected. (D) The hypothesis: the less novelty neurons fire, the more likely it is that a stimulus will be recollected. The more familiarity-detecting neurons fire, the more likely it is that a stimulus will be recollected. The dashed line indicates the baseline. (E–F) Normalized firing rate (baseline = 0) of all novelty (E) and familiarity-detecting (F) neurons during above-chance sessions (30 min R+). Novelty neurons fired more in response to not recollected items (R-) whereas familiarity neurons fired more in response to recollected items (R+). Errors are ±s.e.m. nr of trials, from left to right, 388, 79, 259, 338 (E) and 132, 31, 96, 127 (F).

4.2.3 Quantification of the single-trial responses

Both groups of neurons distinguished recollected from not recollected stimuli, but the difference was of opposite sign. In the novelty case, neurons fire less for recollected items (Figure 4-2E) whereas in the familiarity case neurons fire more (Figure 4-2F). We thus hypothesized that both neuron classes represent a continuous gradient of memory strength. In one case, firing increases with the strength of memory (familiarity detectors) whereas in the other case firing decreases with the strength of memory (novelty detectors). Thus, a strong memory (R+) is signaled both by strong firing of familiarity units as well as weak firing of novelty neurons. Weak memory (R-) is signaled by moderate firing of familiarity and novelty neurons. No memory (a new item) is signaled by strong firing of novelty detectors and weak firing of familiarity detectors. Another feature of the response is that it is often bimodal (see also Figure 4-6). For example, familiarity neurons do not only increase their firing for old items but also decrease firing to new items (Figure 4-2F). This pattern can also be observed in the firing pattern shown in Figure 4-2A: Immediately after stimulus onset, this neuron reduces its firing if the stimulus is old.

We developed a response index R(i) that takes into account the opposite sign of the gradient for the two neuron types, the bimodal response as well as different baseline firing rates. This index makes use of the entire dynamic range of each neuron's response. R(i) is equal to the number of spikes fired during a particular trial i, minus the mean number of spikes fired to all new stimuli divided by the baseline (Eq 1). For example, if a neuron doubles its firing rate for an old stimulus and remains at baseline for a novel stimulus the response index would equal 100%.

By definition, R(i) is negative for novelty units and we thus multiplied R(i) by -1 if the unit was previously classified as a novelty unit.

First, we describe the response of the 30 min R^+ group. In terms of the response index, the average response was significantly stronger to presentation of old stimuli that were later recollected when compared to stimuli which were later not recollected. This was true for a pairwise comparison for every neuron (Figure 4-3A, 68% vs. 50%, n = 45 neurons from 4 subjects) as well as for a trial-by-trial comparison (Figure 4-3B, 67% vs. 45%, $p < 0.01$, n = number of trials). Note that the same difference exists if neurons from the hippocampus (n = 30, R+ vs. R-, $p < 0.05$) or the amygdala (n = 15, R+ vs. R-, $p < 0.05$) are considered separately (see Figure 4-7A and Table 4-2). The difference in response (of 22%) is entirely due to recollection of the source. Re-plotting the data as a cumulative distribution function (cdf) shows a shift of the entire distribution due to recollection (Figure 4-3C, green vs. red line; $p \leq 0.01$). The cdf shows the proportion of all trials that are smaller than a given value of the response index. It illustrates the entire distribution of the data rather than just its mean. We also calculated the response index for correctly identified new items. By definition the mean response to novel stimuli is 0, but it varies trial-by-trial (blue line). The shift in response induced by familiarity alone (blue vs. green, $p \leq 10\text{-}5$) lies in between the shift induced by comparing novel stimuli with old stimuli that were successfully recollected (Figure 4-3C, blue vs. red, $p \leq 10\text{-}19$). The response index is thus a continuous measure of memory strength. From the point of view of this measure, novel items are distractors and old items are targets. We fitted normal density functions to the three populations (distractors, R- and R+ targets). R+ targets showed a greater difference from the distractors than R- targets (Figure 4-3D).

Is there a significant difference between recollected and not recollected stimuli for patients whose behavioral performance was near chance levels? We found that the mean response to recollected and not recollected stimuli did not differ (Figure 4-3E,F. 45% vs. 46%, p = 0.93). This is further illustrated by the complete overlap of the distribution of responses to R+ and R- stimuli (Figure 4-3F, p = 0.53). (This is also true if hippocampal neurons are evaluated separately, Figure 4-7). Thus, the difference (22%) associated with good recollection performance was entirely abolished in the subjects with poor recollection memory.

Was the neuronal response still enhanced by good recollection performance after the 24 h time delay? Subjects in the 24 h delay group had good recollection performance (66%) that was not significantly different from their performance on the 30 min delay period. Thus, information about the source of the stimulus was available to the subject. Surprisingly, however, we found that the firing difference between recollected and not recollected items was no longer present (Figure 4-3G,H). Firing differed by 59% for recollected items compared to 61% for not recollected items (Figure 4-3G,H. p = 0.81). (This is also true if hippocampal neurons are evaluated separately; Figure 4-7C). This lack of difference between R+ and R- items is in contrast to the 30 min R+ delay sessions, where a difference of 22% was observed.

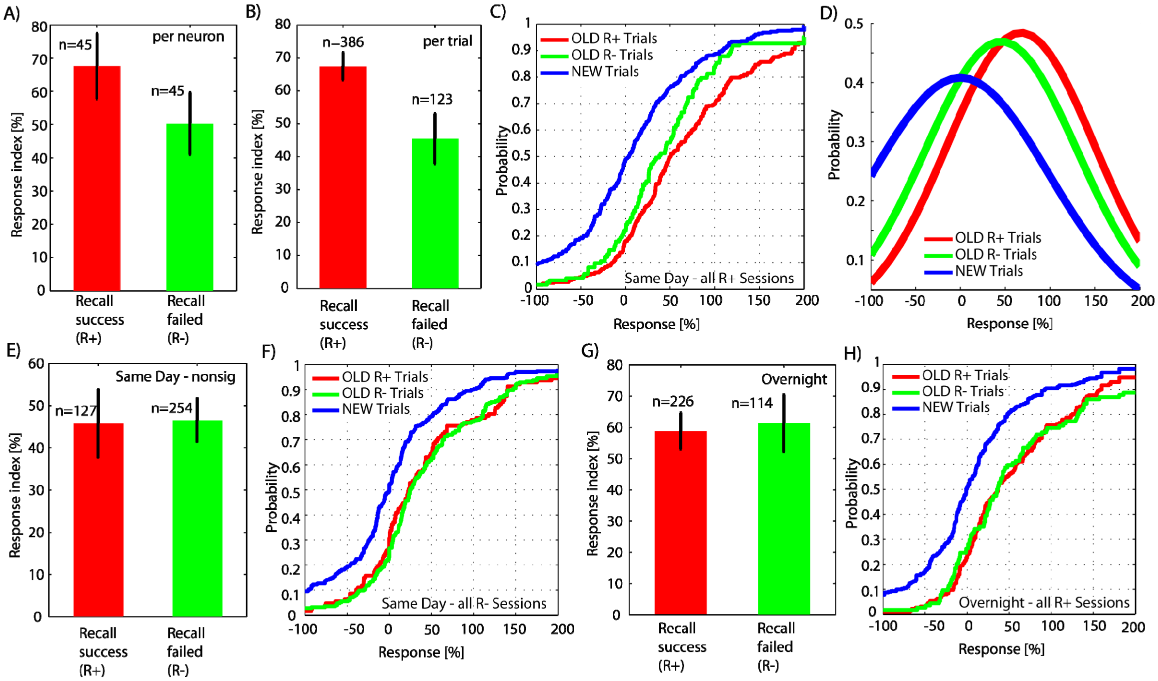

Figure 4-3. Neuronal activity distinguishes stimuli that are only recognized (R-) from stimuli that are also recollected (R+).

(A–E) Same day sessions with above-chance recollection performance (30 min R+). (A) Pairwise comparison of the mean response for all 45 neurons (paired t-test). (B) Trial-by-trial comparison. The response was significantly higher for stimuli which were recalled (R^+, n = 386) compared to the response to stimuli which were not recalled (R^-, n = 123). n is number of trials. (C) Cumulative distribution function (cdf) of the data shown in (B). The response to new stimuli is shown in blue (median is 0). The shift from new to R^- (blue to green) is induced by familiarity only. (D) Normal density functions showing a shift of R^+/ R^- relative to new stimuli. (E–F) Same plots for sessions with chance level performance. There is no significant difference. The cdfs of R^+ (n = 127) and R^- (n = 254) overlap completely but are different from the cdf of new trials (blue v. red/green, p < 10-9). (G–H) activity during retrieval 24h later did not distinguish successful (n = 226) from failed (n=114) recollection. Errors are ±s.e.m.

4.2.4 Neural activity during recognition errors

What was the neural response evoked by stimuli that were incorrectly recognized by the subject? Patients could make two different types of recognition errors: i) not remembering an item (false negative, FN) and ii) identifying a new picture as an old picture (FP). Here, we pooled all same-day sessions (13 sessions from 8 patients) regardless of recollection performance. First, we focused on the FNs. We hypothesized that if the neuronal activity truly reflects the behavior, the response should be equal to the response to correctly identified novel stimuli. On the other hand, if the neurons we recorded from represent a general representation of memory strength, we expect to see a response that is smaller than that observed for correctly recognized items. Indeed, we found that the mean response during "forgot" error trials was 14±3% (Figure 4-4A, yellow), significantly different from the response to novel stimuli (Figure 4-4B, blue vs. yellow; p < 10-4, ks-test). It was also significantly weaker when compared to all correctly recognized items (Figure 4-4B, yellow v. green and red, $p \leq 0.05$, ks-test, Bonferonni corrected). What was the response to stimuli which were incorrectly identified as familiar? We hypothesized that if the FPs represent responses that were truly wrongly identified as old (rather than an accidental button press) we would observe a neuronal response that was significantly different from that observed for novel items. Indeed we found that the response to FPs was significantly different from 0 as well as from the response to novel stimuli (Figure 4-4B, blue v. gray; ks-test p = 0.007). The response to FPs and FNs was not significantly different (Figure 4-4B, gray vs. yellow; ks-test, p = 0.14). (For the previous analysis we pooled neurons recorded from the hippocampus as well as the amygdala. The same response pattern holds, however, if hippocampal units are evaluated separately; Figure 4-7D). This pattern of activity during behavioral errors is consistent with the idea that the neurons represent memory strength on a continuum.

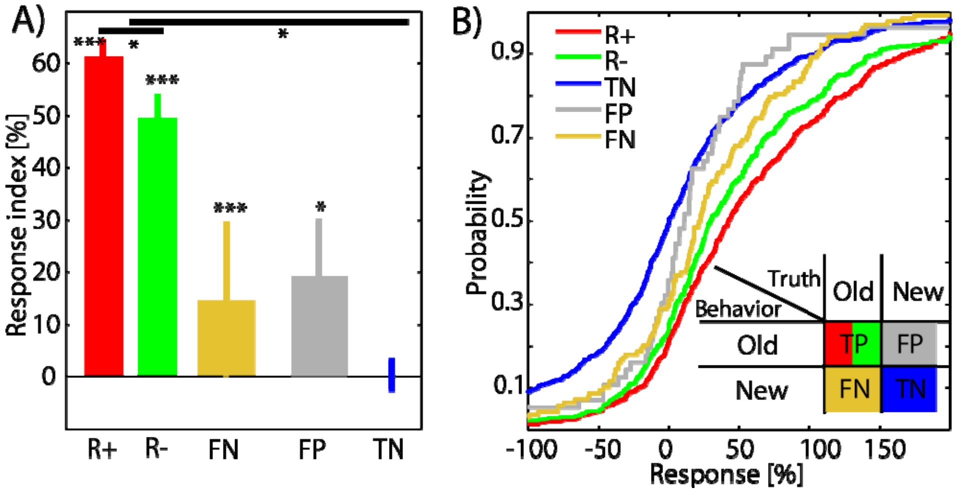

Figure 4-4. Activity during errors reflects true memory rather than behavior. All 30 min sessions are included for this analysis. (A) Neural response. (B) Response plotted as a cdf. Notice the shift from novel to false negatives (p < 10-4): the same behavioral response (novel) leads to a different neural response still differed significantly when compared to real novel pictures. The inset shows the different possible trial types. Errors are ±s.e.m, n is nr of trials (759, 521, 1372, 148, and 56, respectively; 13 sessions, 8 patients).

4.3 Discussion

We analyzed the spiking activity of neurons in the human MTL during retrieval of declarative memories. We found that the neural activity differentiated between stimuli that were only recognized as familiar and stimuli for which (in addition) the spatial location could be recollected. Further, we found that the same neural activity was also present during behavioral errors, but with reduced amplitude. This data is compatible with a continuous signal of memory strength: the stronger the neuronal response, the better the memory. Forgotten stimuli have the weakest memory strength and stimuli which are only recognized but not recollected have medium strength. The strongest memory (and thus neuronal response) is associated with stimuli which are both recognized and recollected.

We used the spatial location of the stimuli during learning as an objective measure of recollection. An alternative measure is the “remember/know” paradigm (Eldridge et al., 2000). However, this measure suffers from subjectivity and response bias. Alternative theories hold that remember/know judgments reflect differences in memory strength rather then different recognition processes (Donaldson, 1996). Thus we chose to use an explicit measure of recollection instead.

We tested 2 different time delays: same day (30 min) and overnight (24 h). Despite good behavioral performance on both days, the neuronal firing only distinguished between R+ and R- trials on the same day. Thus, while the information was accessible to the patient, it was not present anymore in the form of spike counts — at least in the neurons from which we recorded. In contrast, information about the familiarity of the stimulus was still present at 24 hrs and distinguished equally well between familiar and novel pictures (Figure 4-8). While the lack of recordings from cortical areas prevents us from making any definitive claims about this phenomena, it is nevertheless interesting to note that these two components of memory (familiarity and recollection) may be transferred from the MTL to other brain areas with different time courses. Indeed, recent data investigating the replay of spatial sequences by hippocampal units suggest that episodic memories could be transferred to the cortex very quickly. Replay starts in quiet (but awake) periods shortly after encoding and continues during sleep (Foster and Wilson, 2006).

We found that the responses described here can be found both in the hippocampus and the amygdala. Previous human studies have similarly found that visual responses can be found in both areas with little difference (Fried et al., 1997; Kreiman et al., 2000a). Similarly, recordings from monkeys have also identified amygdala neurons which (i) respond to novelty and (ii) habituate rapidly (Wilson and Rolls, 1993). It has long been recognized that the amygdala plays an important role in rapid learning. This is exemplified by its role in conditioned taste aversion (CTA), which is acquired in a single trial, is strongly novelty-dependent, and requires the amygdala (Lamprecht and Dudai, 2000).

The subset of neurons that we selected for analysis exhibited a significant firing difference between old and new stimuli during the stimulus presentation period. This selection criteria allows for a wide variety of response patterns. The simplest case is when a neuron increases firing to one category and remains at baseline for the other. But more complex patterns are possible: the neuron could *decrease* firing for one category and remain at baseline for the other. Or the response could be bimodal, e.g., increase to one category and decrease to the other. To further investigate this, we compared firing during the stimulus period to the pre-stimulus baseline (see supplementary discussion and Table 4-2). 54% of the neurons changed activity significantly for the trial type for which the unit was classified (i.e., old trials for familiarity neurons). 92% of the neurons change their firing rate relative to baseline for either type of trial (e.g., decrease in firing rate of familiarity neurons for new trials). Thus, 38% of the neurons signal information by a significant firing decrease and 8% of the neurons have a bimodal response which individually is not significantly different from baseline. We maintain that the firing behavior of this 8% group contains information about the novelty of the stimulus, even though the responses are not significantly different from baseline. Below we describe several scenarios by which this 8% population might contain decodable information. We repeated our analysis with only the remaining 92% of neurons to assess whether our previous conclusions, based on the entire data-set, still hold true. We found that all results remain valid: The within-repeated ANOVA for the 30 min R+ group revealed a significant difference of New vs. R- as well as R+ vs. R- for both novelty ($p < 1e-4$ and $p = 0.03$, respectively) as well as familiarity units ($p = 0.05$ and $p = 0.02$, respectively). Similarly, the per-neuron ($N = 42$ neurons, $p = 0.03$) as well as the per-trial comparison ($p = 0.01$) remained significant (compare to Figure 4-3A-C). Considering only hippocampal neurons that fire significantly different from baseline, the difference between R+ and R- ($p = 0.04$), R- and New ($p < 0.001$) and New vs. FNs ($p = 0.003$) remained significant (all are tailed ks-tests; compare to Figure 4-7A). All R+ vs. R- comparisons for the 30 min R- and 24 h sessions remained insignificant.

How might a neural network decode the information about a stimulus if it is signaled with no change or a decrease in firing rate? One obvious possibility is by altering excitatory-inhibitory network transmission: if the neuron that signals with a decrease in firing is connected to an inhibitory unit that in turn inhibits an excitatory unit, the excitatory neuron would only fire if the input neuron decreases its firing rate. A similar network could be used to decode information that is present in an unchanged firing rate. How can a network decode information from units that are significantly different new vs. old but not relative to baseline? One possibility is that the network gets an additional input that signals the onset of the stimulus. Thus, it knows which time period to extract. Also, while we can only listen to one single neuron, a readout mechanism gets input from many neurons and can thus read signals with much lower signal-to-noise ratios.

4.3.1 Models of memory retrieval

It is generally accepted that recognition judgments are based on information from (at least) the two processes of familiarity and recollection. How these two processes interact, however, is unclear. Here we have shown that both components of memory are represented in the firing of neurons in the hippocampus and amgydala. Clearly, the neuronal firing described here can not be attributed to one of the two processes exclusively. Rather, the neuronal firing is consistent with both components summing in an additive fashion.

This result has implications for models of memory retrieval. There are two fundamentally different models of how familiarity and recollection interact. The first (i) model proposes that recognition judgments are either based on an all-or-nothing recollection process ("high threshold") or on a continuous familiarity process. Only if recollection fails is the familiarity signal considered (Mandler, 1980; Yonelinas, 2001). An alternative (ii) model is that both recollection as well as familiarity are continuous signals that are combined additively to form a continuous signal of memory strength that is used for forming the recognition judgment (Wixted, 2007). Our data is more compatible with the latter model (ii). We found that the stronger the firing of familiarity neurons, the more likely that recollection will be successful. However, the ability to correctly decode the familiarity of the stimulus does not depend on whether recollection will be successful. This is demonstrated by the single-trial decoding (Figure 4-8): recognition performance only marginally depends on whether the stimulus will be recollected or not. Also, the familiarity of the stimulus can be decoded equally well in patients that lack the ability to recollect the source entirely. Thus, the firing increase caused by recollection is additive and uncorrelated with the familiarity signal. This is incompatible with the high-threshold model, which proposes that either the familiarity *or* the recollective process is engaged. The neurons described here distinguished novel from familiar stimuli regardless of whether recollection was successful. Thus the information carried by these neurons does not exclusively present either index. Rather, the signal represents a combination of both.

4.3.2 Neuronal firing during behavioral errors

What determines whether a previously encountered stimulus is remembered or forgotten? We found that stimuli which were wrongly identified as novel (forgotten old stimuli) still elicited a significant response. Previously we found that this response allows single-trial decoding with performance significantly better than the patient's behavior (Rutishauser et al., 2006a). Thus, information about the stimulus is present at the time of retrieval. This implies the stimuli were (at least to some degree) properly encoded and maintained. However, the neural activity associated with false negative recognition responses was weaker than the responses to correctly recognized but not recollected stimuli (about 60% reduced, Figure 4-4A). The response to false negatives fell approximately in between the response to novel and correctly recognized familiar stimuli (Figure 4-4B). The neuronal response can thus be regarded as an indicator of memory strength. The memory strength for not remembered items is less than for remembered items but it is still larger than zero. However, the memory strength was not strong enough to elicit a "familiar" response. Others (Messinger et al., 2005) have also found neurons that indicate, regardless of behavior, the "true memory" associated with a stimulus. Thus, the neurons considered here likely signal the strength of memory that is used for decision making rather than the decision itself.

False recognition is the mistaken identification of a new stimulus as familiar. The false recognition rate in a particular experiment is determined by many factors, including the individual bias of the subject as well as the perceptual similarity of the stimuli (gist) or their meaning (for words). Here, we found that neurons responded similarly (but with reduced amplitude) to stimuli that were wrongly identified as familiar when compared to truly familiar stimuli. Thus, from the point of view of the neuronal response, the stimuli were coded as somewhat familiar. As such, it seems that the behavioral error possesses a neuronal origin in the very same memory neurons that respond during a correct response — and can thus not be exclusively attributed to simple errors such as pressing the wrong button. MTL lesions result in severe amnesia, measured by a reduction in the TP rate and an increased FP rate relative to controls. However, in paradigms where normal subjects have high FP rates due to semantic relatedness to studied words, amnesics have lower FP rates than controls (Schacter and Dodson, 2001). Thus, in some situations, a functional MTL can lead to more false memory. Similarly, activation of the MTL (and particularly the hippocampus) during false memory has also been observed with neuroimaging (Schacter et al., 1996). This and our finding that neuronal activity does consider such stimuli as familiar suggests that FPs are not due to errors in decision making.

4.4 Methods

4.4.1 Subjects and electrophysiology

Subjects were 10 patients (6 male, mean age 33.7). Informed consent was obtained and the protocol was approved by the Institutional Review Board. Activity was recorded from microwires embedded in the depth electrodes (Rutishauser et al., 2006a). Single units were identified using a template-matching method (Rutishauser et al., 2006b).

4.4.2 Experiment

An experiment consisted of a learning and retrieval block with a delay of either 30 min or 24 h in between. During learning, 12 unique pictures were presented in random order. Each picture was presented for 4 s in one of the 4 quadrants of a computer screen. We asked patients to remember both which pictures they had seen and where on the screen they had seen them. To ensure alertness, patients were asked to indicate where the picture was after each presentation during learning.

In each retrieval session, 24 pictures (12 New, 12 Old, randomly intermixed) were presented at the center of the screen. Afterwards, the patient was asked whether he/she had seen the picture before or not. If the answer was "Old", the question "Where was it?" was asked (see Figure 4-1A). During the task no feedback was given.

4.4.3 Data analysis

A neuron was considered responsive if the firing rate in response to correctly recognized old vs. new stimuli was significantly different. We tested in 2 sec bins (0–2, 2–4, 4–6 s relative to stimulus onset). A neuron was included if its activity was significantly different in at least one of these 3 bins. We used a bootstrap test (p <= 0.05, B = 10000, two-tailed) of the number of spikes fired to New vs. Old stimuli. We assumed that each trial is independent, i.e. the order of trials does not matter. Neurons with more spikes in response to new stimuli were novelty neurons whereas neurons with more spikes in response to Old stimuli were familiarity neurons.

We also used an aggregate measure of activity that pools across neurons. For each trial we counted the number of spikes during the entire 6 s post stimulus period. The response index (Eq 1) quantifies the response during trial i relative to the mean response to novel stimuli.

$$R_i = \frac{nrSpikes_i - mean(NEW)}{mean(baseline)} 100\% \qquad (1)$$

R(i) is negative for novelty detectors and positive for familiarity detectors (on average). R(i) was multiplied by -1 if the neuron is classified as a novelty neuron. Notice that the factor -1 depends only on the unit type. Thus, negative R(i) values are still possible.

The cdf was constructed by calculating for each possible value x of the response index how many examples are smaller than x. That is, $F(x) = P(X \leq x)$ where X is a vector of all response index values.

All statistical tests are t-tests unless stated otherwise. Trial-by-trial comparisons of the response index are Kolmogorov-Smirnov tests (abbreviated as ks-test). All errors are ± s.e. unless indicated otherwise.

4.5 Supplementary results

4.5.1 Behavior quantified with d'

d' was 3.11±0.08, 2.40±0.28 and 2.67±0.68 for the 30 min R^+, 30 min R^- and 24 h groups, respectively. Pairwise tests revealed a significant difference between the 30 min R^+ and R^- group (t-test, $p \leq 0.05$). Thus, in terms of d', patients that exhibited no recollection had significantly lower recognition performance.

4.5.2 Neuronal ROCs

Based on the response values as summarized in Figure 4-3 we constructed two neuronal ROCs (Macmillan and Creelman, 2005): one for trials with spatial recollection and one without (Figure 4-9). The z-transformed ROC was fit well by a straight line (R = 0.997 and R = 0.988 for R+ and R-, respectively). The slope for both curves was significantly different from 1, indicating that the variance of the targets and distractors was different (for a 95% confidence interval the slope was 1.11±0.03 and 1.16±0.07, respectively). The d' for recognized and recollected targets was 0.81 and for targets that were only recognized it was 0.55. Thus, the d' was increased by the addition of recollective information. This is in analogy to the behavioral recognition performance, which was also increased (Figure 4-1E, see above).

Interestingly, the slopes of the neuronal z-ROCs are bigger than 1 (see above). This indicates greater variability for distractors (here new items) compared to familiar items. z-ROC slopes derived from behavioral data are found to be smaller than 1 (Ratcliff et al., 1992).

This has been used as evidence that the target distribution has higher variance compared to the distractor distribution. Intriguingly, we found that the slopes of our z-ROCs are bigger than 1. This further indicates that the neuronal signals in the medial temporal lobe (which we analyze here) represents a memory signal that should be regarded as the input to the decision process, not its output. What is measured behaviorally is the decision itself and it is thus conceivable that the decision process adds sufficient variance to change the slope of the z-ROC.

4.5.3 Responses of novelty and familiarity neurons compared to baseline

The neurons used for our analysis were selected based on a significant difference in firing in response to new vs. old stimuli. This is the most sensitive test because it detects many different patterns in which activity could differ. Example patterns that are detected by this way of classifying units are: i) increase of firing only for one category (new or old) whereas the other remains at baseline, ii) decrease of firing only for one category, with the other remaining at baseline, iii) a bimodal response with an increase to one category and a decrease to the other category. One concern with this analysis is that the response itself might not be significantly different from baseline. This would primarily be the case if the response is bimodal, i.e., a slight increase to one category and a slight decrease to the other. To investigate this possibility we performed additional analysis by comparing the activity of neurons which are classified as novelty or familiarity detecting units against baseline (Table 4-2). We used two different methods: the first ("method 1") tests whether the unit increases its firing rate significantly for either the old (familiarity neurons) or the new trials (novelty neurons). However, there are several classes of units which this method misses. For example, a unit which remains at baseline for old trials and reduces its firing rate for new trials would be classified as a familiarity unit. However, it would not pass the baseline test since the response for old trials remains at baseline. To include such units we used a second method ("method 2"): for a unit to be considered responsive, the activity of either the new or the old trials needs to be significantly different from baseline. The unit in the above example would pass this test.

Using method 2, we found that 92% of all units which were classified as signalling a difference between new and old were in addition also firing significantly different relative to baseline (see Table 4-2 for details). Using method 1, 54% of all units pass this additional test. Thus approximately 40% of the units signal information by a decrease in firing rate rather then an increase.

4.5.4 Population activity

So far we have analyzed the spiking of single neurons which fired significantly different for new vs. old stimuli. However, the majority of neurons (72% of neurons; 298 of 412) did not pass this test and thus were not considered in our first set of analyses. Was there a difference in mean firing between new and old stimuli if neurons were not pre-selected? To address this, we calculated a mean normalized activity for all recorded neurons in all sessions, separately for new and old trials (Figure 4-10A). This signal reflects the overall mean spiking activity of all neurons and is thus similar to what might be measured by the fMRI signal (see discussion). Only trials where the stimulus was correctly recognized were included. The mean firing activity of the entire population was significantly different in the time period from 2–4 s relative to stimulus onset ($p \leq 0.05$, t-test, Bonferroni corrected for $n = 8$ comparisons). Thus, a difference in overall mean activity for novel vs. familiar stimuli can be observed even without pre-selecting neurons. However, the initial response (first 1 s, Figure 4-10A) did not differentiate between the two types of stimuli. Rather, a sharp onset in the response could be observed for both classes of stimuli. Did the population only differentiate because the novelty and familiarity detectors were included in the average? We also calculated the population average (as in Figure 4-10A) using only the units which were not classified as either novelty or familiarity detectors. The average population activity still exhibited a sharp peak for both types of stimuli after stimulus onset and significantly differentiated between novel and familiar items in subsequent time bins ($p \leq 0.05$, t-test, Bonferroni corrected for $n = 8$ comparisons).

Is the population response different for stimuli which are recollected compared to stimuli which are only recognized? The previous average included all old trials, regardless of whether the stimulus was recollected or not. Next, we averaged all trials from all neurons recorded for the 30 min delay sessions with good recollection performance (30 min R^+). We found a similar pattern of population activity (Figure 4-10B). Crucially, however, the neuronal activity in response to familiar stimuli which were later not recollected peaked earlier. Measured in time bins of 500 ms, the only significant difference between familiar stimuli that were recollected or not was in the first 500 ms after stimulus onset ($p \leq 0.05$, t-test, Bonferroni-corrected for $n = 16$ comparisons). Thus, the population activity peaks first for stimuli that are not recollected, followed by novel and recollected stimuli.

4.5.5 Decoding of recognition memory

Is the ability to determine whether a stimulus is old influenced by whether the stimulus was recollected or not? In the main text we have shown that the responses to recollected stimuli are stronger compared to items which are not recollected. Here, we investigate whether this increased response leads to an improvement in the ability to determine (based on the neuronal firing only) whether a stimulus is new or old. If the two types of information (familiarity and recollection) interact, one would expect that the ability to recollect would increase the ability to determine whether a stimulus has been seen before. Alternatively, recollection could be a process that is only triggered after the familiarity is already determined and these two types of information would thus be independent. Thus, one would expect no difference in the ability to determine the familiarity from the spiking of single neurons in cases of successful vs. failed spatial recollection. To answer this question, we used a simple decoder. It used the weighted linear sum of the number of spikes fired after the onset of the stimulus. The weights were determined using regularized least squares, a method very similar to multiple linear regression (see methods). The decoder had access to the number of spikes in the 3 consecutive 2 s bins following stimulus onset (3 numbers per trial).

First, we used the decoder to determine for how many trials we could correctly predict whether the stimulus was new or old, based only on the firing of a single neuron. For all sessions (n = 17), the decoder was able to predict the correct identity for 63 ± 1% of all trials. We repeated this analysis for each of the 3 behavioral groups (R^+ 30 min, R^- 30 min, and R^+ 24 hr). We found (Figure 4-8A) that the recognition decoding accuracy (chance 50%) did not depend on whether the subject was able to recollect the source of the stimulus or not (1-way ANOVA, p =

0.35). Thus, decoding of familiarity is equally effective, even in the group where patients were not able to recollect at all (Figure 4-8A, 30 min R- sessions).

Was there a difference in decoding performance in the same-day group where subjects had good recollection performance? We selectively evaluated the performance of the decoder for two groups of trials: trials with correct recollection and trials with failed recollection. We find that firing during trials with failed recollection does carry information about the familiarity of the stimulus (Figure 4-8B, R-). The ability to predict the familiarity of the stimulus was slightly improved for the behavioral group with good recollection performance on the first day (Figure 4-8B, right. p = 0.03, paired t-test).

4.6 Supplementary discussion

4.6.1 Differences between amygdala and hippocampal neurons

So far, we have analysed neurons recorded from the amygdala and the hippocampus as a single group. We pooled the responses from both groups because we previously found that both structures contain units which respond to novel and familiar items in a very similar fashion (Rutishauser et al., 2006a). Nevertheless we also analyzed the activity separately for both brain structures. We find that the previous finding still holds — while the response magnitude differs, the overall response pattern is very similar. In particular, all primary findings of our paper hold independently for the hippocampus as well as the amygdala (see below).

We found that the increased response to old stimuli which are recollected (R+) compared to stimuli which are not recollected (R-) is present in both hippocampal as well as amygdala neurons (Figure 4-11; 74.8±5.3% v. 61.3±8.6% for the hippocampus and 52.2±6.8% vs. 13.7±14.2% for the amygdala). The response magnitude (comparing all old trials, regardless of whether they are R+ or R-), however, is larger in the hippocampus (71.6±4.5% v. 42.8±6.3%, $p < 0.001$). While the amplitude of the response is different there is nevertheless a significant difference between R+ and R- trials in both areas.

This is further illustrated in Figure 4-7, where we replotted the response to old R+, old R, new, and false negatives (forgotten items) for all 3 behavioral groups only considering hippocampal units (Figure 4-7A–C). The relevant differences (R+ vs. R-, New vs. false negative) are the same as for the pooled responses (see Figure 4-7 legend for statistics). Similarly, the responses during the error trials (false negatives and false positives) are the same (compare Figure 4-7D to Figure 4-4B).

We also repeated the within-group ANOVA for only the hippocampal units of the 30min R+ session. The ANOVA was significant for novelty ($p = 4.1e-6$) as well as familiarity ($p = 1.3e-19$) units. The planned contrasts of R- v.s New and R+ vs. R- revealed a robust difference for novelty ($p = 5.1e-5$ and $p = 0.04$, respectively) units. For familiarity units, the R- vs. New contrast was significant ($p = 0.002$) whereas the R+ vs. R- contrast was only approaching significance ($p = 0.17$). This is because there were only 7 familiarity units that contribute to this comparison. Repeating the same comparisons while excluding all units that do not fire significantly different from baseline (see Table 4-2) reveals a similar pattern: the ANOVA for familiarity units remains unchanged (all units different from baseline) whereas the novelty units ANOVA still shows a significant difference between R- vs. New (p = 2.7e-5) as well as R+ vs. R- (p = 0.016).

4.6.2 Differences between epileptic and non-epileptic tissue

Was the neuronal response reported here influenced by changes induced by disease? All subjects for this study have been diagnosed with epilepsy and as such some of the effects may not extend to the normal population. Behaviorally, our subjects were comparable to the normal population (see Table 4-1). Also, we separately analyzed a subset of neurons which were in a non-epileptic region of the subject's brain. We found a comparable (but stronger) response to old stimuli in this "healthy" neuron population (Figure 4-11D). Similarly, we find that neurons from the "to be resected" tissue still exhibited a response to old stimuli (Figure 4-11E). This response was, however, weaker and there was no significant difference between recollected and not recollected stimuli. Thus, it is possible that the average difference between recollected and not recollected items in normal subjects will be larger than that observed in the epileptic patients in our study.

4.6.3 Relationship to previous single-cell studies

A previous human single-cell study (Cameron et al., 2001) concluded that the neuronal activity observed during retrieval is due to recollection. The task used was the repeated presentation of word pairs with later free recall and thus included no recognition component. Due to the choice of words and the repeated presentation of the same word pairs, the novelty/familiarity of the stimuli was not controlled for. It is thus not clear whether the activity observed was related to recollection or to the recognition of the familiarity of the stimuli. Here, we combine both components in the same task and thus demonstrate that the same neurons represent information about both aspects of memory simultaneously. Similar paired associates tasks have been used with monkeys (Sakai and Miyashita, 1991; Wirth et al., 2003). Changes in neuronal firing were, however, only observed after many learning trials (> 10). A neuronal correlate of episodic memory requires changes after a single learning trial. It thus seems possible that this study documented the gradual acquisition of well-learned associations rather than episodic memories.

4.6.4 Relationship to evoked potentials

Both surface and intracranial evoked potentials show prominent peaks in response to new stimuli. Scalp EEG recordings during recognition of previously seen items show an early frontal potential (~ 300 ms) which distinguishes old from new items, as well as a late potential (~ 500–600 ms) that is thought to reflect the recollective aspect of retrieval (Rugg et al., 1998). However, the signal origin of these scalp recordings is not known. These differences between evoked potentials in response to new and old items are reduced or absent in patients with hippocampal sclerosis (Grunwald et al., 1998). Intracranial EEG recordings from within the hippocampus as well as the amygdala show prominent differences between new and old items (around 400–800 ms) (Grunwald et al., 1998; Mormann et al., 2005; Smith et al., 1986), further suggesting the MTL as a potential source for the scalp signal. The latencies and nature of these potentials are also in agreement with the average population activity that we have analyzed (Figure 4-10). We find that the peak activity is within the 500–1000 ms timeframe (Figure

4-10B). Remarkably, the activity peaks first (within the first 500 ms) if recollection fails. If recollection is successful, the peak is in the second bin (500–1000 ms). This suggests that a recognition judgment based purely on familiarity occurs quicker. In addition, it is worth noting that the average population activity we recorded is compatible with the previous intracranial EEG findings but conflicts with BOLD signals obtained by others (Eldridge et al., 2000; Yonelinas et al., 2005) .

4.6.5 Relationship to fMRI studies

This is also in apparent conflict with previous functional magnetic resonance imaging (fMRI) findings (Eldridge et al., 2000; Yonelinas et al., 2005) that identified regions within the MTL that are selectively activated only for memories that are recollected. Crucially, however, these studies assumed *a priori* that model (*i)* above is correct by searching for brain regions which correlate with the components identified by that model. If model (*i)* is not correct, however, these results are subject to alternative interpretation. Also, these studies used the "remember/know" paradigm to identify memories which were recollected by the subjects. However, this paradigm requires a subjective decision (yes/no) as to whether the memory was recollected or not (as discussed above). It is thus possible that the brain areas identified using these paradigms reflect the decision taken about the memory rather than the retrieval process itself. In our study, no decision as to whether or not recollection succeeded was necessary. Also, our data analysis makes no assumptions about the validity of any particular model.

What is the appropriate baseline activity to consider in the MTL? The MTL is highly active during quiet rest. In fact it is often more active during rest than during memory retrieval

(Stark and Squire, 2001). Imaging studies can suffer from this undefined baseline and results may vary owing to different choices of representative baseline activity (Stark and Squire, 2001). This may also contribute to the apparently disparate findings regarding the involvement of the MTL in recognition memory.

To further investigate the discrepancy between fMRI and single-cell studies, we averaged the neuronal activity of all neurons recorded regardless of their behavioral significance, to approximate a signal that might be similar to an fMRI signal (Figure 4-10, see Results). We found that even under this condition, the overall population activity successfully distinguished between new and old items. The response to old items was not selective for recollected items and was clearly present even if the failed recollected trials were considered separately (Figure 4-10B). Clearly these data differ from previously measured hippocampal BOLD signals (e.g. (Eldridge et al., 2000)).

4.7 Supplementary methods

4.7.1 Electrophysiology

All patients were diagnosed with drug-resistant temporal lobe epilepsy and implanted with intracranial depth electrodes to record intracranial EEG and single units. Electrodes were placed based on clinical criteria. Electrodes were implanted bilaterally in the amygdala and hippocampus (4 electrodes in total). Each electrode contained 8 identical microwires, one of which we used as ground. We were able to identify single neurons in the hippocampus and/or amygdala in 9 of the 10 patients. One additional patient was excluded because he had no recognition memory (performance was at chance). Thus, this study is based on

8 patients (6 of which overlap with a previous study; (Rutishauser et al., 2006a)). We recorded a total of 21 retrieval sessions from these 8 patients. 4 of these sessions (from 4 different patients) were excluded due to insufficient recognition performance (see below). Thus, this study is based on 17 retrieval sessions from 8 different patients. The 17 retrieval sessions were distributed over 16 different days (on one day, 2 retrieval sessions were conducted). We recorded from 24–32 channels simultaneously (3 or 4 electrodes) and found, on average, 11.9±4.4 (±s.d.) active microwires (counting only microwires with at least one well-separated unit). The average number of identified units per wire was 2.0±1.0 (± s.d.). Inactive wires (no units identified) are excluded from this calculation (77 of 280). There were 130 wires with more than one unit (on average 2.6±0.8 for all wires with > 1 unit). For those wires, we quantified the goodness of separation by applying the projection test (Rutishauser et al., 2006b) for each possible pair of neurons. The projection test measures the number of standard deviations the two clusters are separated after normalizing the data such that each cluster is normally distributed with a standard deviation of 1 (see (Rutishauser et al., 2006b) for details). We found that the mean separation of all possible pairs (n=315) is 13.68±6.98 (± s.d.) (Figure 4-12A). We identified, in total, 412 well-separated single units. We quantified the quality of the unit isolation by the percentage of all interspike intervals (ISI) which are shorter than 3 ms. We found that, on average, 0.3±0.4 percent of all ISIs were below 3ms (Figure 4-12B). The signal-to-noise ratio (SNR) of the mean waveforms of each cluster relative to the background noise was on average 2.4±1.2 (Figure 4-12C).

For the purpose of comparing only neurons from the "healthy" brain side (left or right), we excluded all neurons from either the left or right side of the patient if the patient's diagnosis (Table 4-1) included temporal lobe damage (Figure 4-11). No neurons were excluded if the diagnosis indicated that the seizure focus was outside the temporal lobe.

4.7.2 Behavior

Each session consisted of a learning and retrieval block. We quantified, for each session, the recognition rate (percentage of old stimuli correctly recognized), the false positive rate (percentage of new stimuli identified as old), and the recollection rate. The recollection rate was the percentage of stimuli identified as old for which the spatial location was correctly identified. Sessions with a recognition rate of $\leq$ 50% were excluded (3 sessions). Each session was assigned to either the 24 h or 30 min delay group.

For each session, we estimated whether spatial recollection rate was significantly different from chance (25%). Due to the small number of trials (maximally 12), the significance was estimated using a bootstrap procedure (see below). Based on this significance value, we further divided each of these two groups into a group with good spatial recollection performance ($p \leq 0.05$, above chance, R^+) and one with poor spatial recollection performance (not significantly different from chance, $p > 0.05$, R^-). For the 24 h group there was only one session with poor recollection performance and thus this analysis was not conducted. Thus, there were 3 behavioral groups which were used for the neuronal analysis: 30 min R^+ ($n = 7$), 30 min R^- ($n = 6$) and 24 h R^+ ($n = 4$). The assignment of sessions to groups was based entirely on behavioral performance. Neuronal activity was not considered.

4.7.3 Data analysis — behavioral

We labeled each retrieval trial during which a correctly recognized old stimulus was presented as either correctly or incorrectly recollected. For each session we then tested (bootstrap, $p \leq 0.05$, one-tailed, $B = 20000$) whether recollection performance was above chance level. We used the bootstrap test instead of the z-test because of the small number of samples. The resulting p values were more conservative (larger) compared to the p values obtained with the z-test. Only sessions which passed this test were considered to have "above chance" recollection performance. Trials which failed this test were considered as "at chance". This was to ensure that only neurons from patients that had a clearly demonstrated capability for source memory were included. Also, recording sessions with less than a 50% hit rate for old stimuli were excluded to ensure that only sessions with sufficient recognition performance were included. We verified for each group of sessions (Figure 4-1) whether performance was significantly above chance using a z-test. For this, we pooled all trials of a particular group and labeled each as either correct or incorrect. Then we used one z-test to test whether the ratio correct:incorrect was above chance. We used this instead of individual tests for each session to avoid artificially boosting performance due to the small sample size (e.g., 4 out of 12 correct) in each particular session.

4.7.4 Data analysis — response index

We compared, trial-by-trial, the response (quantified by the response index) to old stimuli which were successfully recollected (R^+) to old stimuli which were not recollected (R^-). For this comparison, trials with recognition errors were excluded (thus, all trials are familiar).

The error trials were analysed separately. There was one data point for every trial for every neuron (e.g., if there are 10 trials and 10 neurons, there are 100 data points). There were 1368 old stimulus trials (12 retrieval sessions with total 114 neurons), with 1230 trials with a correct recognition response (familiar, TP), and 138 trials which were errors (misses). We analyzed the error trials separately.

We compared the responses of the R^+ and R^- trials with a two-tailed t-test, as well as using a Kolmogorov-Smirnov test. Both were significant at $p \leq 0.05$. Paired comparisons were made with a t-test. Normal density functions were constructed by estimating the mean and standard deviation from the data (using maximum likelihood).

4.7.5 Data analysis — baseline comparison

To determine whether a unit was responsive relative to baseline we compared the firing during the 2 s period in which the new vs. old comparison is significant to the 2 s period before the stimulus onset. These comparisons were performed using a boostrap test as described in the main methods.

4.7.6 Neuronal ROCs

Neuronal ROCs (Figure 4-9) were constructed by considering all trials as old if the response R(i) was above a threshold T. The threshold T was varied in variable steps (see below) from the smallest to the largest value of R(i). Thresholds were varied such that each increase accounted for a 5% quantile of all available datapoints (the 0% and 100% quantiles were excluded). This procedure assured that the same number of datapoints was used for the calculation of each point in the ROC. The hit/false positive rate was calculated for each threshold value. d' was calculated for each pair of hit/false positive rates and averaged. We z-transformed the ROC and fit a line through all points using linear regression to find the slope of the curve. A slope of 1.0 indicates that the two distributions (distractors and targets) are of equal variance whereas a slope of unequal 1.0 indicates a difference in variance. The z transformed ROC was fit well by a straight line for both R^+ and R^- trials (Macmillan and Creelman, 2005).

4.7.7 Population averages

Population averages (Figure 4-6, Figure 4-10) were constructed by normalizing each trial to the mean baseline firing in the 2 s before stimulus onset. The number of spikes were binned into 1 s bins (non-overlapping) and averaged for all neurons. No smoothing was applied. To avoid normalization artifacts, only neurons with a baseline rate of at least 0.25Hz were considered for the population averages (346 of 412 neurons for Figure 4-5). Also, for Figure 4-10 only neurons with a significant response in the stimulus period (first two of the 2 s bins) were considered (this does not apply for the trial-by-trial analysis).

4.7.8 Decoding

We used a linear classifier to estimate how well the firing of a single neuron during a single trial can signal the identity (new or old) of the presented stimulus. The classifier was provided with the number of spikes fired in 3 consecutive 2 s bins after stimulus onset (0–2 s, 2–4 s, 4–6 s). The classifier consisted of a weighted sum of these 3 numbers. The weights were estimated using regularized least squares (RLSC) (Evgeniou et al., 2000; Rifkin et al., 2003). This method is equal to multiple linear regression with the exception of an added regularizer term λ (see below; we used λ = 0.01 throughout). The decoding accuracy of the classifier was estimated using leave-one-out crossvalidation for all training samples available. The estimated prediction error was equal to the percentage of correct leave-one-out trials. There were maximally 12 samples in each class (old or new). However, due to behavioral errors, fewer trials were sometimes available for analysis. Error rates for false positives and false negatives were approximately equal and the number of samples was thus approximately balanced in both classes. Of concern was whether a slight imbalance of the number of samples in one class could bias the results. We performed two controls to assess whether this was the case: we performed leave-one-out cross-validation with the label of the test sample randomly re-assigned with 50% probability. If the classifier was biased, the resulting error would be different from 50%. We found that this was not the case (Figure 4-8A). Also, we re-ran all analysis that used the decoder with a balanced number of samples (that is, equal number of samples in either class) and found no difference in the results.

The weights were determined by regularized least squares. Regularized least squares are very similar to multiple linear regression. In the following we would like to point out these differences because in a previous study we used a multiple linear regression (Rutishauser et al., 2006a).

With multiple linear regression (Eq S1), the weights w are determined by multiplying the inverse of data samples Z with the trainig labels y (Johnson and Wichern, 2002).

$$w = \left[Z'Z\right]^{-1} Z' y \qquad \text{(S1)}$$

In contrast, in regularized least squares (Evgeniou et al., 2000; Hung et al., 2005; Rifkin et al., 2003), an additional term is added to the data samples (Eq S2). Here, I is the identity matrix and λ is a scalar parameter (the regularizer).

$$w = \left[Z'Z + \lambda I\right]^{-1} Z'y \qquad \text{(S2)}$$

The value of the regularizer is arbitrary. The bigger it is, the more constraints are placed on the solution (the less the solution is determined by the data samples). A small value of the regularizer, on the other hand, makes the solution close to the multiple linear regression solution. Importantly, however, even a small value of the regularizer punishes unrealistically large weights and also guarantees full rank of the data matrix. Regularization becomes particularly important when there are a large number of input variables relative to the number of training samples. This is the case in our study because each neuron contributed 3 variables (3x 2 s time periods) and the number of training samples was small (on the order of 10). Thus, regularization was necessary. We found that performance was maximal for a small (but non-zero) regularizer and used $\lambda = 0.01$ throughout.

4.8 Supplementary figures

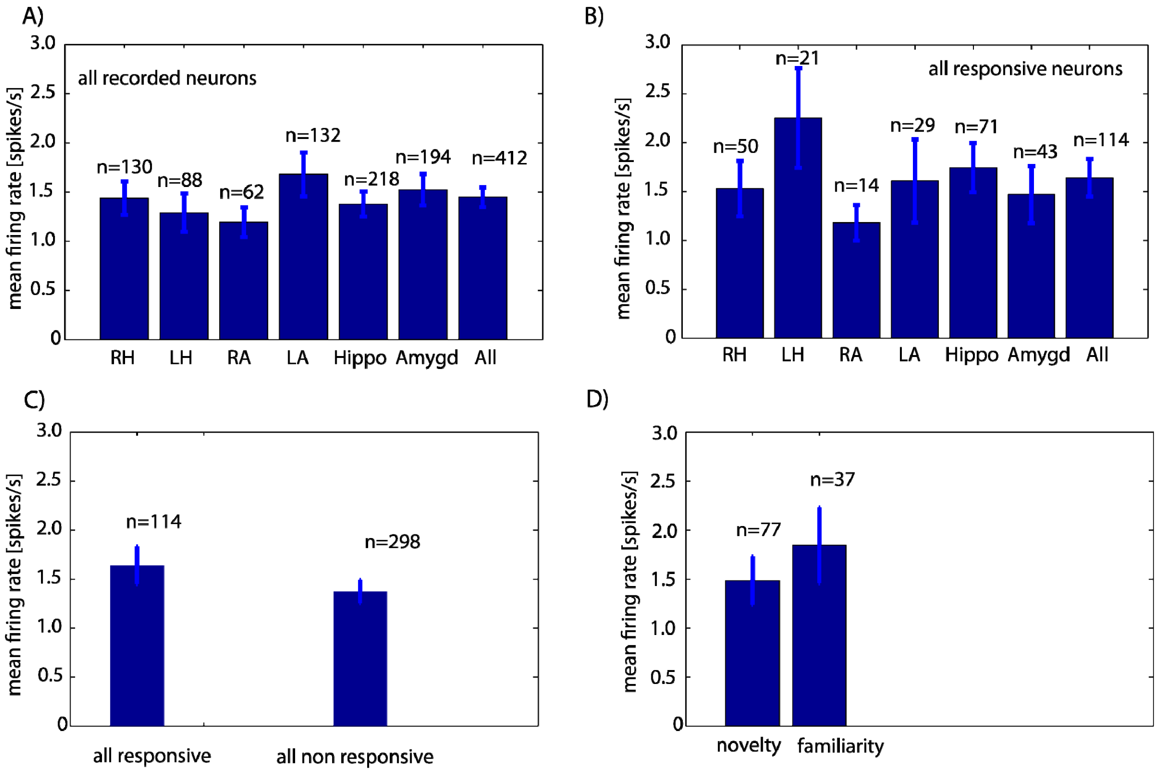

Figure 4-5. Population average of all recorded neurons.
(A) Population average of all recorded neurons that have a baseline firing rate of >0.25Hz (n = 346). While the firing of most neurons was not significantly different between new vs. old, a significant difference between new and old stimuli could still be observed in the population average. Errors are ±s.e.m and ** indicates significance of a one-tailed t-test at $p \leq 0.006$ ($p \leq 0.05$ Bonferonni-corrected for 8 multiple comparisons). (B) Population average of all neurons with recollected and not recollected familiarity trials shown separately. (C) Population average of all neurons recorded in the 30 min delay sessions with above chance recollection performance. The signal for the not recollected items peaked earlier than the signal for recollected items. ** indicates a significant difference between recollect (R^+) and not recollected (R^-) items at $p \leq 0.003$ ($p \leq 0.05$ Bonferonni-corrected for 16 multiple comparisons). The only difference was for the first time bin (0–500 ms after stimulus onset). n = 134 neurons.

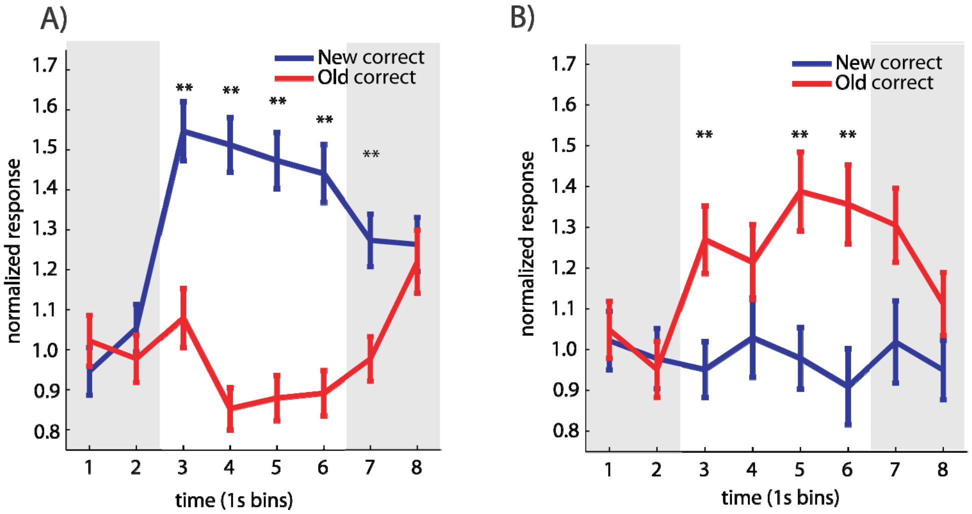

Figure 4-6. Population response.

(A-B) Population average of all neurons that responded significantly during the stimulus period. The stimulus was on the screen during the 4 s period marked in white. **(A)** Average of all neurons that increased firing to correctly recognized new items (“novelty detectors”) (*n* = 48). **(B)** Average of all neurons that increased firing to correctly recognized old items (“familiarity detectors”) (*n* = 26). Errors are ± SEM and ** indicates significance of a one-tailed *t* test at $P \leq 0.006$ ($P \leq 0.05$ Bonferroni corrected for multiple comparisons). Firing was normalized to the 2 s baseline firing before stimulus onset marked in gray. Note that this does not mean all neurons fired during the entire period; but rather represents the population average.

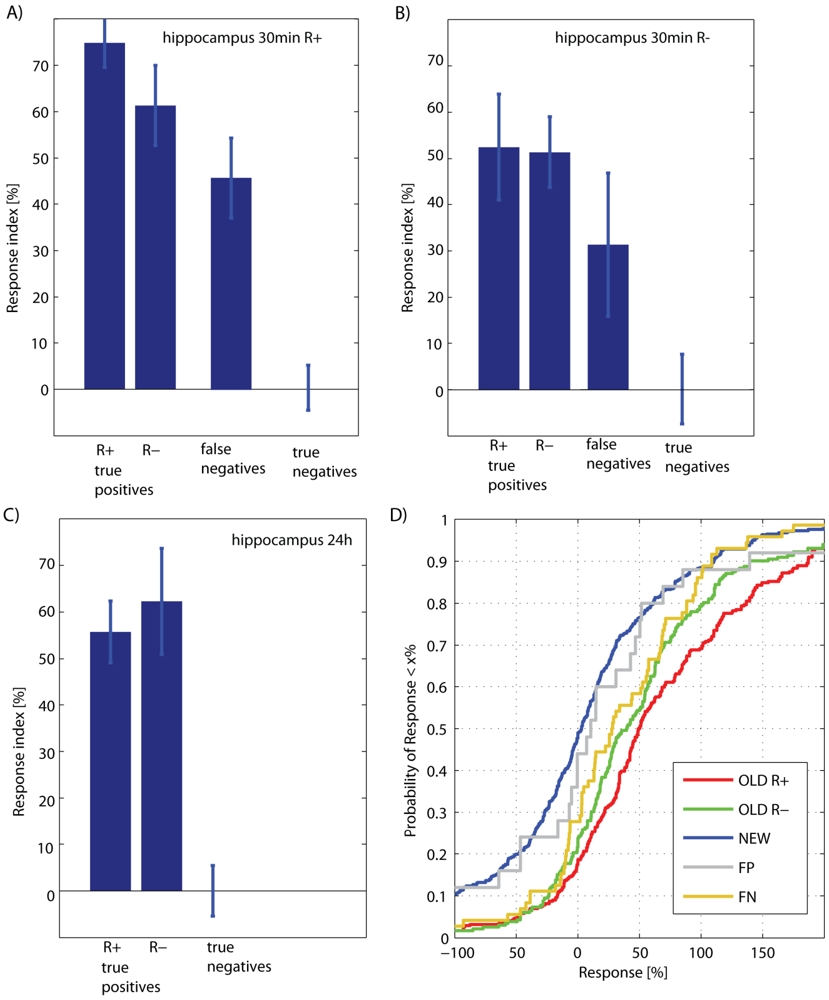

Figure 4-7. A continuous strength of memory gradient exists when the hippocampal neuronal population is considered in isolation.

In this figure, the same measures are replotted, but all units recorded from the amygdala are excluded. All findings remain valid. (A) Trials from the 30 min R+ sessions. There is a significant difference between R+ and R- trials ($P = 0.03$) as well as between new and false negatives ($P = 0.001$). Compare to Figure 4-3C. (B) Trials from the 30 min R- session. There is no significant difference between R+ and R- trials ($P = 0.93$) but false negatives are still significantly different from new trials ($P = 0.07$). Compare to Figure 4-3F. (C) Trials from the 24 h sessions. There is no significant difference between R+ and R- trials. Error trials are not shown (not enough for 24 h sessions). Compare to Fig. 4-3H. (D) cdf of response index of all hippocampal neurons recorded in all 30 min sessions. R+ and R- trials are significantly different (red v. green, $P = 0.01$) as are new and false negatives (blue vs. yellow, $P < 0.001$). Not enough false positive trials are available to allow statistical analysis of false positives. Compare to Fig. 4-4. All errorbars are ± SE.

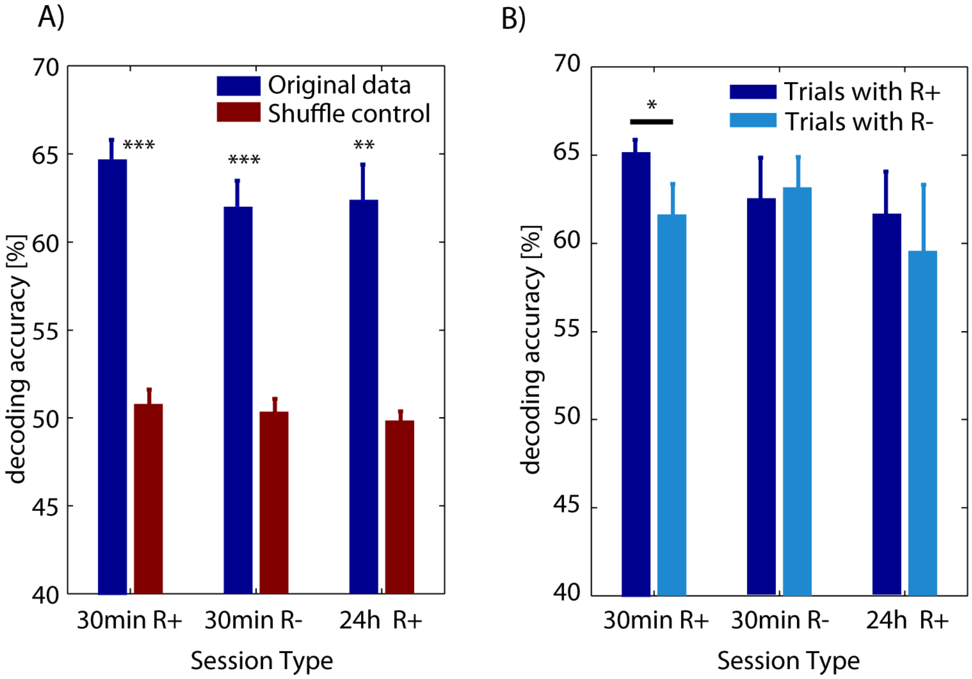

Figure 4-8. Whether a stimulus is new or old can be predicted regardless of whether recall was successful or not.

The decoder had access to the number of spikes fired in the 3 consecutive 2 s bins following stimulus onset (3 numbers total). (A) Session-by-session differences. The performance of the decoder did not change for all 3 groups (ANOVA, $P = 0.35$). n = 7,6,4 sessions, respectively. (B) Trial-by-Trial differences. Here, the decoder was trained on the complete set of trials but its performance was evaluated separately either for failed (R^-) or successful (R^+) recall trials. Clearly, the familiarity of the stimulus could be decoded for trials with failed recall (R^-). In the 30 min delay sessions with successful recall (30 min R^+), firing during successful recall trials contained significantly more information about the familiarity of the stimulus ($P = 0.037$, paired t test, $n = 7$ sessions). All errorbars are ± SE.

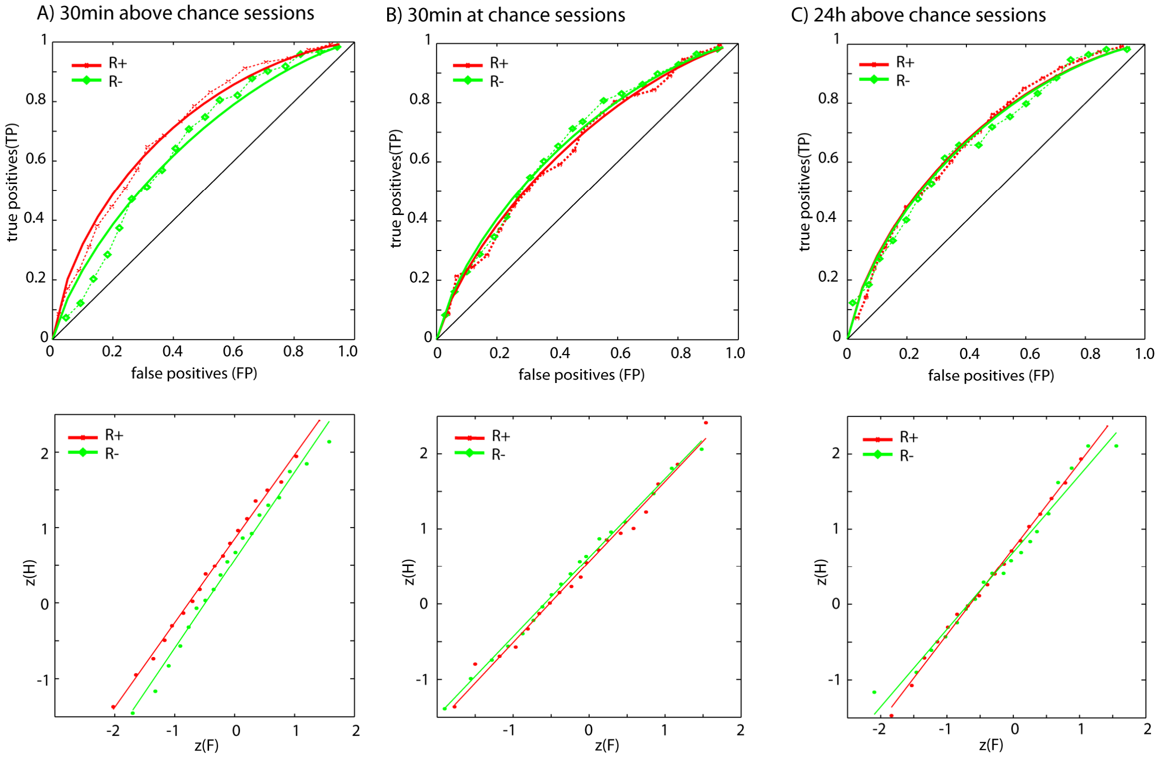

Figure 4-9. ROC analysis of the neuronal data for all 3 behavioral groups. (A: 30 min above chance, B: 30 min at chance, C: 24 h above chance). The top row shows the raw datapoints as well as fits computed from d’. The bottom row shows the same but z-transformed. R^2 is > 0.97 for all straight line fits. See the supplementary methods for how the ROC was computed. A) d’ for R^+ and R^- groups was 0.81 and 0.55, respectively. The slope (s) of the z-transformed line was 1.11 ± 0.03 and 1.16 ± 0.07, respectively. ± are 95% confidence intervals. B) d’ was 0.55 and 0.61 and s was 1.07 ± 0.06 and 1.05 ± 0.04, respectively. C) d’ was 0.73 and 0.69 and, was 1.14 ± 0.04 and 1.02 ± 0.08, respectively.

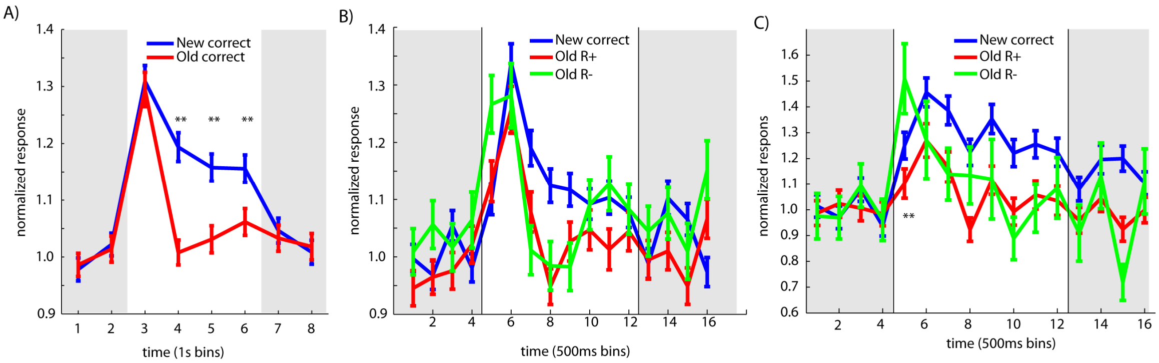

Figure 4-10. Population average of all recorded neurons.

(A) Population average of all recorded neurons that have a baseline firing rate of > 0.25 Hz (*n* = 346). While the firing of most neurons was not significantly different between new vs. old, a significant difference between new and old stimuli could still be observed in the population average. Errors are ± SEM and ** indicates significance of a one-tailed *t* test at $P \leq 0.006$ ($P \leq 0.05$ Bonferonni-corrected for 8 multiple comparisons). (B) Population average of all neurons with recollected and not recollected familiarity trials shown separately. (C) Population average of all neurons recorded in the 30 min delay sessions with above chance recollection performance. The signal for the not recollected items peaked earlier than the signal for recollected items. ** indicates a significant difference between recollect (R^+) and not recollected (R^-) items at $P \leq 0.003$ ($P \leq 0.05$ Bonferonni-corrected for 16 multiple comparisons). The only difference was for the first time bin (0–500 ms after stimulus onset). *n* = 134 neurons.

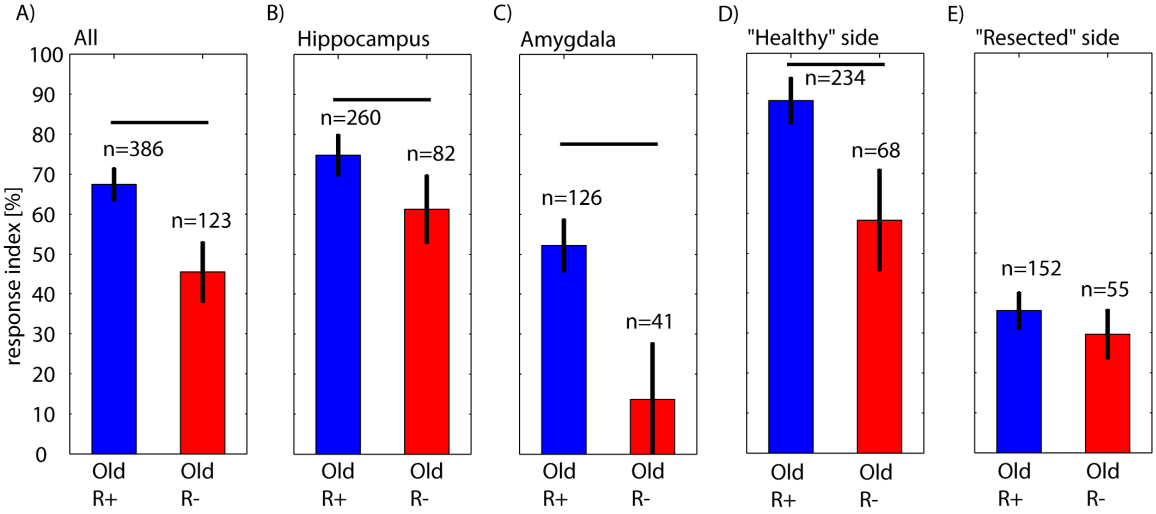

Figure 4-11. Comparison of trial-by-trial response strength for different subcategories of neurons.

In this figure, only neurons from 30 min delay with successful recollection (30 min R+) are included. **(A)** All trials from all areas (same as Figure 3B). **(B)** Only trials from hippocampal neurons. **(C)** Only trials from amygdala neurons. **(D)** Only trials from the "healthy" hemisphere. **(E)** Only trials from neurons in the eventually resected hemisphere. In **(A-D)**, the response to R+ compared to R- trials is significantly different (*P* < 0.05, two-tailed Kolmogorov-Smirnov test, compare to Figure 3B). The response in (E) is not significantly different.

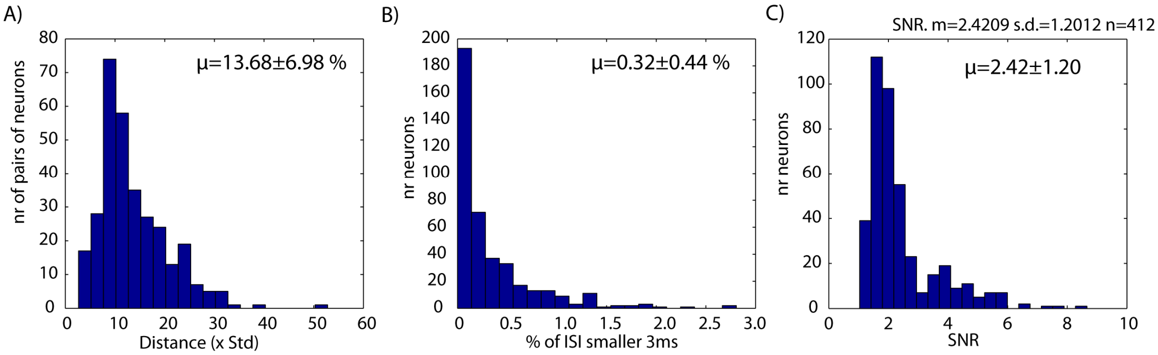

Figure 4-12. Sorting quality for the 412 recorded units.
(A) Histogram of the distance, in standard deviations, between all pairs of clusters. Only channels on which more than one unit was detected are included (315 pairs from 130 channels). The mean distance was 13.68 ± 6.98 (± s.d.) (B) Histogram of the percentage of interspike intervals (ISI) that were shorter than 3 ms. On average 0.32 ± 0.44% of all ISIs were shorter than 3 ms (n = 412). (C) Histogram of the SNR of all 412 units.

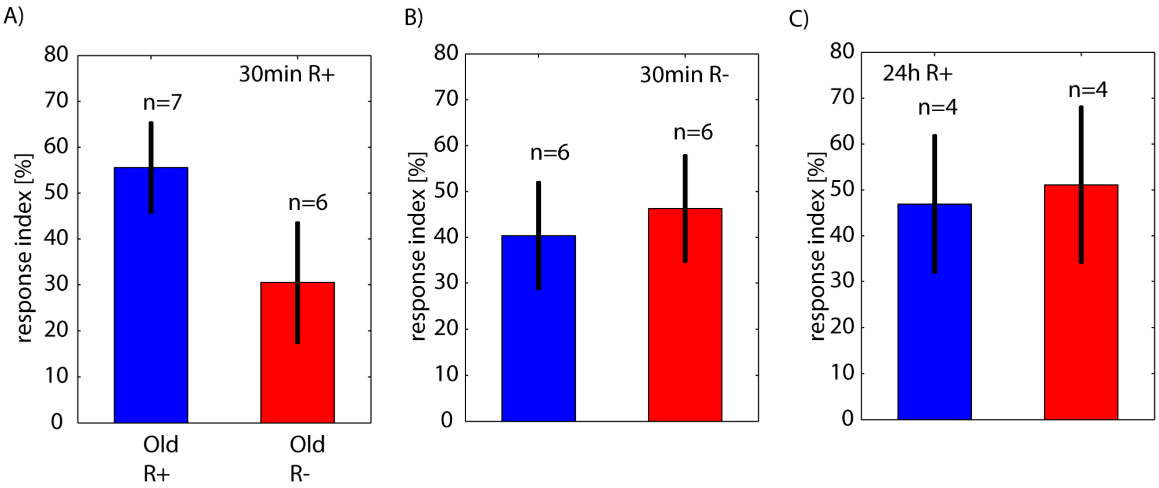

Figure 4-13. Comparison of response strength across different recording sessions (days).

The difference is only significant for the 30 min R+ sessions. The data displayed here is the same as detailed in Figure 4-3. However, here the mean response index for R+ and R- trials is compared between recording sessions. **(A)** The response index for all recording sessions that had above chance recollection. The difference approaches significance (*P* = 0.07). Number of sessions is 7 and 6, respectively (from 4 patients; one session had no R- trials). **(B)** Same as (A) but for all recording sessions with at chance recollection. Number of sessions is 6 for both groups (from 5 patients). There was no significant difference (*P* = 0.63). **(C)** Same as (A) but for all recording sessions with 24 h delay and above chance recollection. Number of sessions is 4 from 3 patients. There was no significant difference (*P* = 0.57). Errorbars are ± SEM with n as specified. p values are from a *t* test.

4.9 Supplementary tables

				WAIS-III			WMS-R					
Patient	**Age**	**Sex**	**Diagnosis**	**PIQ**	**VIQ**	**FSIQ**	**Verbal Mem**	**Mental control**	**VPA 2**	**LM 2**	**Vis Rep 1**	**Vis Rep 2**
1	28	m	left temporal	125	98	110	114	6	4	24	37	39
2	41	f	left temporal	92	91	91	91	5	8	18	37	29
3	20	f	left temporal	92	93	93	83	6	8	16	34	28
4	58	f	left temporal	85	83	83	83	6	4	10	22	7
5	23	m	left temporal & frontal pole	144	111	126	122	6	8	26	39	39
6	44	m	right temporal	76	92	84	83	6	5	10	29	14
7	51	f	left temporal	90	95	93	89	6	4	23	34	34
8	16	m	right lateral frontal	84	91	88	n/a	n/a	8	n/a	31	29
av	*35.1*	-	-	*98.5*	*94.3*	*96.0*	*95.0*	*5.9*	*6.1*	*18.1*	*32.9*	*27.5*
mean raw								5.0±1.2	7.6±0.7	21.9±9.2	32.5±5.3	29.5±7.1

Table 4-1. Neuropsychological evaluation of patients.
Intelligence was measured using the Wechsler Intelligence Scale (WAIS-III) measures of performance IQ (PIQ), verbal IQ (VIQ), and full scale IQ (FSIQ). All IQ scores have an average of 100 (by design). Memory measures are from the Wechsler Memory Scale Revised (WMS-R). Verbal memory is an WMS-R index score with a mean of 100 of the normal population (by definition). The remaining WMS-R scores are raw (unnormalized) scores. For the raw scores, the mean and standard deviation of the normal population (from WMS-R) is shown in the last row for the average age of our population. Abbreviations: Verbal paired associates 2 (VPA 2), Logical Memory 2 (LM 2), Visual Reproduction 1 (Vis Rep 1), Visual Reproduction 2 (Vis Rep 2).

	Group	**Hippocampus**			**Amygdala**			**All**		
Recorded	*30min R+*	77			103			180		
	30min R-	96			47			143		
	24h R+	45			44			89		
	all	218			194			412		
		Nov	**Fam**	**All**	**Nov**	**Fam**	**All**	**Nov**	**Fam**	**All**
New v. old	*30min R+*	25	7	32	10	5	15	35	12	47
	30min R-	11	11	22	13	3	16	24	14	38
	24h R+	11	6	17	7	5	12	18	11	29
	all			71			43	77	37	114
New v. old & baseline 1	*30min R+*	14	5	19	6	3	9	20	8	28
	30min R-	5	6	11	6	1	7	11	7	18
	24h R+	5	4	9	5	2	7	10	6	16
	all			39 (55%)			23 (53%)			62 (54%)
New v. old & baseline 2	*30min R+*	22	7	29	10	5	15	32	12	44
	30min R-	10	10	20	11	3	14	21	13	34
	24h R+	9	6	15	7	5	12	16	11	27
	all			64 (90%)			41 (95%)			105 (92%)

Table 4-2. Number of neurons recorded.

Number of neurons recorded in each area (first row) and number of neurons that responded in each behavioral group(2nd, 3rd, 4th row). The second row shows the number of neurons which had a significantly different firing rate for old vs. new trials during the post-stimulus period (6s). The last two rows show the number of neurons which are, in addition, also significantly different for two different baseline comparisons (1 and 2). The two baseline comparisons are: i) The trials associated with the type of unit are significant from baseline. (That is, if the neuron is classified as a familiarity neuron, the old trials were significantly different from baseline. The same applies for the novelty neurons, but for the new trials). ii) Either the new or the old trials are significantly different from baseline. Note that the first (i) baseline condition is the most restrictive: for example, a familiarity unit that decreases firing to novel items but remains at baseline for familiar items would not pass this test. For the second baseline condition, 92% of units (105 of 114) remain significant. Thus, almost all units fired significantly different from baseline for either the new or old condition. Note that some of the n’s reported in the main analysis are slightly lower than the numbers reported in this table. This is because additional constraints were applied (for example, at least one R+ and one R- trial for each included unit).