Chapter 6. Summary and General Discussion

6.1 Novelty detection and single-trial learning

This thesis is about single neurons in the human brain that express a fundamental piece of information: whether a stimulus is novel or familiar. By definition, a stimulus is novel only once. The second time it is seen, it is familiar. Novelty- and familiarity detecting neurons follow this by rapidly changing their firing pattern. The response of novelty- and familiarity detecting neurons is not binary. Rather, the response strength (or absence, in the case of novelty neurons) is proportional to the strength of memory. Familiarity detecting neurons, by definition, increase their firing rate for stimuli which are familiar (have been seen before). Their response is strongest for stimuli that are recognized and recollected, intermediate for recognized stimuli (but not recollected), and weak (but non-zero) for forgotten stimuli. Novelty detecting neurons, by definition, only increase firing for novel items. For familiar items, however, they tend to decrease their firing. The stronger the memory, the larger the firing rate decrease (relative to the response to novel stimuli). Thus, both novelty and familiarity detecting neurons signal the strength of memory, but with opposite polarity.

The neurons discussed here are capable of the most rapid form of plasticity: single-trial learning. Most events in life occur only once. Thus it is of fundamental importance to investigate this form of learning. Responses to novel stimuli are extremely prevalent in the brain and in behavior (Sokolov, 1963). Many neurons in many areas of the brain respond differently if a stimulus is novel or otherwise salient in some way. Responses to novelty can also be observed behaviorally: animals such as rats have a natural tendency to explore novel objects. In fact, this

effect is commonly used to test recognition memory for objects in rodents (Ennaceur and Delacour, 1988). Animals automatically orient towards novel stimuli (Sokolov, 1963; Vinogradova, 2001). Humans and non-human primates automatically move their eyes towards novel objects and they spend more time fixating novel objects (Althoff and Cohen, 1999; Smith et al., 2006; Yarbus, 1967). This preference exists even in infants (Fantz, 1964). Autonomic reactions such as skin conductance (Knight, 1996), heart rate (Weisbard and Graham, 1971), or pupil diameter-dilation (Hess and Polt, 1960) also show prominent novelty responses. Many of these novelty responses are severely reduced by lesions of parts of the medial temporal lobe (Honey et al., 1998; Kishiyama et al., 2004; Knight, 1996; Knight and Nakada, 1998; Yonelinas et al., 2002). This is particularly the case for hippocampal lesions. Being novel (or more generally, different) is a very effective modulator of memory strength. This is true even if the attribute that makes a stimulus novel is task irrelevant. This is the well known "von Restorff" effect (Hunt, 1995; Kinsbour and George, 1974; Kishiyama et al., 2004; Parker et al., 1998; von Restorff, 1933; Wallace, 1965). Thus, it is clear that novelty is an efficient modulator of memory strength. Some have proposed that novelty increases dopamine release, which is known to induce strong and long-lasting plasticity (Lisman and Grace, 2005). Given the persistence of this phenomenon, it seems warranted to speculate that the feature that novelty enhances memory constitutes an evolutionary advantage and is thus selected for.

While remembering something as best as possible (and thus detecting novelty as well as possible) is usually advantageous, there are also situations where strong memories are not advantageous for the individual. Examples are memories which can, despite best efforts, not be erased such as in post-traumatic stress disorder (PSTH), drug-induced place preference, or

memories connected to strong emotions. What these examples have in common is that the memory was established by a single experience (a single trial). Thus, the novelty advantage conveyed to memories can also be a disadvantage.

Given the importance and prevalence of novelty-dependent effects, surprisingly little is known about the neuronal mechanisms of such rapid learning. Among the many behavioral paradigms used to study learning, most require a large number of learning trials. Examples are conditioning (both classical and instrumental), the Morris water maze (learning the escape location), and maze learning. There are sophisticated models for these types of learning (see (O'Doherty et al., 2003; Schultz, 2002; Seymour et al., 2004) for examples). Such models are typically variants of reinforcement learning (Sutton and Barto, 1998). However, in everyday life most learning tasks we face are not of this nature. Rather, they are of the more rapid kind of learning where we, at best, learn from a few trials (Exceptions are acquiring new habits). Examples of behavioral paradigms for rapid learning are those described in this thesis (for humans), conditioned taste aversion as well as some forms of fear conditioning. In these paradigms, the failure to detect a stimulus as novel impairs learning severely (Welzl et al., 2001). For such rapid learning tasks we lack the formal understanding that we have for incremental learning (such as reinforcement learning models). This also applies to learning by machines. In machine learning, learning from many examples by training classifiers is well established. There is no equivalent technique to learn from just a few trials. In tasks which require many trials to learn (such as maze navigation), it has become clear that neural activity during the first few learning trials (in a novel environment) is very different to that observed when learning is completed (see (Cheng and Frank, 2008) for an example). This stresses the importance of

recording from the very initial learning trials rather then when the animal is well trained (or over-trained), as is most often done in the case of hippocampal recordings. Given that the hippocampus is thought to be most important for learning, it is likely that many important processes are never observed because they are over by the time recording starts.

6.2 The relationship between memory responses and behavior

The firing rates of single neurons in the human amygdala and hippocampus can be used to construct a simple new/old decoder that outperforms the patient. That is, if the patient made an error (either forgetting a stimulus or wrongly declaring it familiar), the neuronal responses often indicated what would have been the correct decision (which was not made by the patient). Thus, these neurons had better memory than the patient. Here, we used this fact to argue that these neurons do not represent the motor output nor the decision of the patient. This is because there is a clear dissociation between the decision (and thus motor output) and the neural response. For both types of trials (forgot and new trial), the behavioral response is the same: a press of the "New" button. The neuronal response, however, is very different. Thus, these neurons do not represent the patient's decision, rather they may represent the input to the decision-making process.

Research in decision making typically focuses on decisions about external stimuli, i.e., which cue is most likely to indicate a reward. Memory retrieval, however, involves a different kind of decision making: decisions about internal states. Deciding whether a stimulus is novel requires deciding that there is no trace or representation of the presented stimulus in the system and thus the stimulus is novel. Similarly, deciding that a stimulus is familiar requires judgment of

whether there is enough evidence of previous occurence. Both types of decisions are about neuronal firing, which represents an internal state rather than an external stimulus. Little is known about how such decisions are made. The paradigms and neuronal responses presented in this thesis lend themselfs well to investigating this process. A first step would be to identify an area of the brain that contains novelty/familiarity detectors (such as the ones documented here) that follow the decision rather then the memory (and that are not trivially related to motor output). Candidates for such area(s) are frontal areas such as the anterior cingulate, medial prefrontal, or orbitofrontal areas (Badre and Wagner, 2007; Koechlin and Hyafil, 2007; Lepage et al., 2000; Wagner et al., 2001). For saccadic decisions, it is known that the frontal eye fields (FEF) represent the decision rather then the visual input (Hanes and Schall, 1996). Simultaneous recordings from the hippocampus/amygdala and this yet to be identified area (preferably in humans, which is possible for several candidate areas) would be a powerful system to investigate how decisions are made about the presence or absence of memories. Asking humans to judge their confidence would be particularly useful in this setting. While identifying the location of such neurons itself only tells us where the decision is represented, simultaneous recordings from both areas will allow detailed investigation into how the decision itself is made (e.g., by looking at their interactions in time during errors).

6.3 Novelty and familiarity responses in the amygdala

Most of the data reported in this thesis are pooled across the amygdala and the hippocampus. We also analyzed the data separately, however. Surprisingly, the differences in

terms of novelty/familiarity responses are (on average) subtle. This is in agreement with previous human record data (Fried et al., 1997; Kreiman et al., 2000a). Clearly, both the amygdala and the hippocampus contain neurons which respond as described in detail in this thesis. What is remarkable, however, is that the difference between recollected and not-recollected familiar items is much more pronounced in the amygdala (Figure 4-11). The response for stimuli which are only familiar but not recollected is much smaller in the amygdala than in the hippocampus. At first, this seems surprising, as it suggests that the amygdala is more involved in recollective memory than the hippocampus is. However, an alternative interpretation is that the amygdala is proportionally more active for memories that have an emotional component. Since the emotional component is only attributed to an object if it is recollected, it seems reasonable that the response to objects which are not recollected is rather weak in the amygdala. Also note that this comparison is based on the average response to all stimuli. The stimuli we used could have an emotional value for some patients and not for others. Averaging would erase these effects. In a trial-by-trial comparison it is possible that there are stimuli which evoke a stronger amygdala familiarity/novelty response due to some stimulus property such as emotional content. This suggests a further experiment, using trial-by-trial correlations with image rankings along different dimensions (saliency, emotional content). Stimuli could either be rated by an independent subject population or a standardized dataset can be used (such as the International Affective Picture System dataset, (Lang and Cuthbert, 1993)).

The amygdala is well known to have a strong influence on memory. Emotional stimuli are remembered better than non-emotional stimuli (Heuer and Reisberg, 1990). Patients with amygdala lesions have good memory but lack the enhanced memory for emotional stimuli

(Adolphs et al., 1997; Adolphs et al., 2000; Phelps et al., 1997). Thus, the role of the amygdala in memory formation seems to be modulatory (Mcgaugh et al., 1990; Phelps, 2004). However, in some situations the amygdala is necessary for rapid (and often novelty-dependent) learning as, for example, in conditioned taste aversion (Lamprecht and Dudai, 2000) or in fear conditioning (Wilensky et al., 2006). It thus seems reasonable that neurons in the amygdala are novelty sensitive as well as plastic.

6.4 Differential response strength in epileptic tissue

All data reported in this thesis has been recorded from patients with a long history of epilepsy. Based on careful neuropsychological measures, we have argued that our patient population is not different from the normal population in their ability to learn, remember, or reason (Table 4-1). It is thus reasonable to conclude that, in the absence of seizures, their brains function comparable to normals since they achieve the same behavior. One of the primary clinical aims of intracranial electrode implantation is to determine whether seizures have a clear unilateral origin. If this is the case, unilateral resection of parts of the MTL is a possible treatment (see Introduction). Excluding all patients who did not receive a clear unilateral MTL diagnosis, we used knowledge of the laterality of the epileptic focus to compare neural responses between the epileptic and the presumably non-epileptic side. As a measure of response strength we used a response index that is equal to the absolute difference between the response to novelty and familiarity of neurons that are novelty sensitive. Using this index we find (Figure 4-11) that the response index for neurons in the epileptic hemisphere was much weaker when compared to the non-epileptic side (For recollected trials, 88% vs. 36%). Also, neurons in the epileptic side did

not fire significantly differently for recollected vs. not-recollected trials (for healthy neurons there was a 20–30% difference). While we have not verified this with a predictive study, this finding nevertheless suggests potential value for this paradigm as a useful diagnostic. Due to the large difference it is imaginable that similar differences in novelty/familiarity responses also exist in multi-unit data or even LFP, which would make it easier to use this diagnostic clinically.

6.5 Predictors of successful learning

Transforming a new experience into a long-term memory is a complex process that is poorly understood. It starts with the neural activity during the initial acquisition and continues at least for hours (but probably for much longer) after initial acquisition (consolidation). At the time of retrieval, which can be many years after initial acquisition, these changes are sufficient to evoke the feeling of familiarity, sometimes together with other attributes that were part of the learning experience (an episode). While it is clear that the initial acquisition is clearly necessary for successful retrieval, it is unclear how much of the retrieval variability can be attributed directly to it rather then all the other events that contribute to a memory (Paller and Wagner, 2002). Representing a robust memory likely requires changes (plasticity) in a large number of neurons. Inducing the cellular changes thought to underlie these changes requires tightly coordinated neuronal activity. It is thus thought that one of the important contributors to successful learning is synchrony (Axmacher et al., 2006). In this thesis I show that specific components of the LFP, measured during learning, are predictive of whether a stimulus will be remembered or not. This supports the hypothesis that increased synchrony is crucial for successful learning.

6.6 The value of studying single-unit responses in humans

The observation of spike trains emitted by single neurons in the human brain while the subject is awake and engaged in a task is a tremendous opportunity. Spikes are arguably the common currency of communication of brains, and thus the appropriate units that we would like to observe and study. There are a multitude of functions that are extremely difficult (or sometimes impossible) to study in animal models (see Introduction for details). The object of study in this thesis, episodic memory, is one example of this. Recording from humans in a clinical setting has many disadvantages over animal models. For example, the experimental conditions are relatively poorly controlled. Head and eye movements can not be constrained, nor can patients be over-trained to do a task perfectly (owing to human subjects concerns). Single-unit isolation quality is often not as good as with animals (due to single-wire recordings rather then tetrodes, and to electrode movement issues). Electrode location is known only approximately and cannot be confirmed with histology. No cellular or molecular manipulations are possible. With these caveats in mind, human recordings nevertheless offer a tremendous opportunity that should be utilized as much as possible. Great care should be taken to only use this rare opportunity to address problems that are well-suited to this technique, and not better addressed in other systems. It seems of dubious value to me to simply reproduce standard experiments done in monkeys or rats to conclude that it is the same in humans. Also, there are clearly questions which are better addressed with other techniques such as surface EEG or fMRI. Examples of such experiments are questions related to which brain area responds to some particular condition. Given the restricted (and fixed) implantation sites of depth electrodes, the questions best approached with this question are distinctively different. Examples are: What subclasses of neurons respond to a given

stimulus? What is the latency of the response? How does the response, trial-by-trial, relate to behavior (particularly during errors, or different confidence levels, or awareness)? How selective are the responses of the same neurons to different stimuli? What are the dynamics of interactions between neurons in the same population? These are questions which cannot be addressed using other techniques. This also stresses the importance of robust behavior. Many neuronal responses only make sense if studied in the context of an appropriately designed task where all behavioral variables are properly controlled. The true power of human recordings is combining behavior with the observation of single neurons. In the absence of behavior (such as passive viewing), many of the benefits of awake human recordings are not taken advantage of.

6.7 Note on visual tuning of MTL neurons

Many neurons in the MTL respond selectively to certain aspects of the visual input, such as its category (i.e., animal, person, house) or its identity (see Introduction for details). One curious aspect of these studies is that in almost every patient recorded, one finds such cells. This is remarkable, since in a typical recording session there are at very best several tens of neurons (and often fewer). Out of these few neurons, which are sampled entirely randomly from implanted fixed electrodes, invariably a few are tuned to the task variable (such as visual category) at hand. This indicates that the tuning of these neurons is probably not static. Rather it seems to be the case that MTL neurons are automatically tuned to all relevant attributes of a particular task. One of the earliest single-unit studies already remarked on this aspect by stating "These data suggest that MTL stimulus-specific responses represent a temporary allocation of a subset of MTL neurons to the ensemble encoding of distinct events within a given context" (Heit

et al., 1988). While this aspect of sensitivity to the task has not been studied systematically, it nevertheless suggests that this aspect of MTL neuron function is distinctively different from neurons found in sensory areas, where tuning is typically thought to be static (such as receptive fields in early visual areas or even object-selective neurons in IT cortex). Alternatively, if tuning is static, this would imply that each neuron responds to many different categories (Waydo et al., 2006). However, this would make it difficult to reconcile the finding that one finds tuning to almost everything that is task relevant, whatever the task. This puzzling finding suggests further experiments such as changing the task-relevant categories or similar manipulations.