UNIVERSITY OF AMSTERDAM

The Neural Dynamics of Fear Memory

R.M. Visser

Summary

The neural dynamics of fear memory

While much of what we learn will be forgotten over time, emotional memory appears to be particularly resilient to forgetting. This is in principle adaptive: the ability to store relevant information enables us to cope effectively with similar events in the future and promotes survival. Long-lasting memory for emotional experiences can, however, also become harmful and maladaptive, such as in patients suffering from post-traumatic stress disorder. Even though we know quite a lot about the processes that can weaken or strengthen memory for fearful events, our understanding of how these events are processed, consolidated and eventually expressed remains poor. Studying fear memory is complicated by the fact that we cannot directly observe memory, but have to infer it from the different ways it is expressed. Yet, only a small portion of what is concurrently happening in our brain is translated into observable actions and peripheral physiology; an even smaller portion of these processes is verbally accessible. Hence, there is a notable discrepancy between the expression of fear during *learning* and the actual formation and expression of a fear *memory*. The fact that someone is learning does not imply that a memory is being formed; and even if a memory has been formed, it may not be observable in any given context, at any moment in time, nor remain stable over time. To better understand how fearful events are transformed into durable memory and how memory influences the processing of (novel) stimuli, we need multiple indices of fear memory, including read-outs of neural processes that cannot (yet) be observed in behavior.

To explain individual differences in the formation of memories for arousing events, experimental psychology reduces complex real-life situations to relatively simple experiments that permit us to rigorously test ideas about cause and effect. The classic model to study fear learning and memory is Pavlovian fear conditioning, which is well suited for research across species. In this paradigm an initially neutral stimulus (conditioned stimulus, CS+; e.g., a picture of a face) is repeatedly paired with an intrinsically aversive stimulus (unconditioned stimulus, UCS; e.g., an electric shock), while another conditioned stimulus (CS-; e.g., a picture of another face) is never paired with the UCS. With sufficient CS+/ UCS pairings, the CS+ acquires the same aversive qualities as the UCS and will elicit a conditioned fear response (CR) on its own. After repeated presentations of the CS+ without the UCS, the fear response usually diminishes, a process that is referred to as 'extinction'.

In this thesis, we used a variety of methods to identify indices that capture the complexity of human fear memory, combining several experimental procedures (Pavlovian fear conditioning, a pharmacological manipulation, and a perceptual judgment tasks) with behavioral and neural indices of fear (BOLD-MRI, pupil dilation, fear potentiated startle, perceptual decisions and subjective reports).

To specifically gauge the dynamic nature of fear (un)learning and memory, we developed a single-trial application of multi-voxel pattern analysis (MVPA; a technique that evaluates distributed patterns of BOLD-MRI) and disentangled learning and memory processes by testing on different days, varying contextual cues to (de)activate specific memories.

In **chapter 2** we measured the dynamics of associative fear learning, to examine whether the acquisition of aversive associations would affect the neural representation of neutral stimuli and to examine how new knowledge is incorporated into existing semantic networks. By applying MVPA in a trial-by-trial manner, we quantified changes in patterns of BOLD-MRI over the course of Pavlovian fear conditioning. Our findings showed an increase in similarity of neural response patterns on consecutive trials of stimuli that were followed by an aversive stimulus (UCS, i.e., electrical stimulation), but not of unreinforced stimuli. These changes in representational similarity resulted in clear differential learning curves that indexed the formation of associative fear. Remarkably, the representations of *different* stimuli, derived from two very distinct categories (faces and houses) also became more similar over the course of conditioning, but again only when these stimuli were paired with an aversive outcome. This between-stimulus pattern similarity, induced by linking different stimuli with the same aversive outcome, putatively reflected a type of higher-order fear learning. In line with this notion, the large-scale distribution of within-stimulus pattern similarity and between-stimulus pattern similarity differed across the cortex. The entire cortex, including low-level perceptual areas, showed a tuning toward stimuli that were paired with the aversive stimulus, resulting in higher within-stimulus pattern similarity. Conversely, the representational dominance of stimuli that shared an aversive outcome, as reflected in between-stimulus pattern similarity, appeared to be restricted to frontoparietal areas. In other words, while a house was still more similar to a house in occipital and inferotemporal regions, frontoparietal regions seemed less sensitive to perceptual resemblance and more tuned toward stimulus significance. These findings suggested that trial-by-trial MVPA is a useful tool for examining how the human brain encodes relevant associations and forms new associative networks.

The question that we addressed in **chapter 3** was whether these changes in similarity structure solely reflected the sensory and behavioral processes that are active during fear conditioning, or whether they also inform about upcoming consolidation processes, thereby providing a neural signature for predicting the long-term expression of fear memory. In this study, we used trial-by-trial MVPA, as described in chapter 2, to quantify learning-dependent changes. Next, we linked these changes to a behavioral expression of long-term fear memory (i.e., differential pupil dilation responses). Replicating the findings from chapter 2, both within-stimulus and between-

stimulus pattern similarity revealed clear learning curves that indexed the formation of associative fear. This increase in pattern similarity was much weaker for stimuli that predicted a neutral outcome (a sound), indicating that the refinement in neural processing is characteristic of salient associations. After a few weeks, the reactivation of fear memory traces (through reinstatement of the same threatening context) was again reflected in differential pattern similarity. When the aversive outcome was no longer delivered, differential similarity eventually disappeared. The learning curves that were obtained with trial-by-trial similarity analysis were either absent or substantially weaker (depending on the region) for single-trial activation analysis. Crucially, our findings showed that the strength of between-stimulus pattern similarity at the time of learning - and not changes in average activation, within-stimulus similarity or the behavioral expression itself - predicted the long-term behavioral expression of fear memory. This suggests that changes in neural activation patterns at the time of encoding, specifically those that reflect the formation of higher-order associations, 'mark' the subsequent changes in synaptic structure that underlie consolidation.

In **chapter 4** we examined how changes in neural patterns could be interpreted in terms of neuromodulatory mechanisms. Given that noradrenaline is known to strengthen the consolidation of emotional memory, we aimed to investigate the effects of this neurotransmitter on neural activation patterns during fear conditioning. Replicating our previous findings, neural pattern similarity reflected the development of fear associations over time, and unlike average activation or pupil dilation, predicted the later expression of fear memory (pupil dilation 48 hours later). While no effect of yohimbine HCl (the pharmacological manipulation) was observed on markers of autonomic arousal, including salivary α-amylase (sAA), we obtained indirect evidence for the noradrenergic enhancement of fear memory consolidation: sAA levels showed a strong increase prior to fMRI scanning, irrespective of whether participants had received yohimbine, and this increase correlated with the subsequent expression of fear (48 hours later). Moreover, this noradrenergic enhancement of fear was associated with changes in neural response patterns at the time of learning. These findings provided further evidence that representational similarity analysis is a sensitive tool for studying (enhanced) memory formation. Furthermore, they indirectly support the idea that the tagging of memories for subsequent consolidation is related to noradrenaline, and - as indexed by changes in neural response patterns - already occurs during encoding.

To justify and facilitate the implementation of the methods used in chapter 2, 3 and 4, in **chapter 5** we systematically examined the effects of different design choices on single-trial pattern analysis in general, and the ability to assess the dynamics of fear learning in particular. Data from two experiments showed that designs with longer stimulus intervals showed stronger learning effects and

augmented the classification of stimulus categories (houses and faces), even if this came at the cost of the number of trial repetitions. This finding can be explained by the fact that with shorter intervals the trial-specific BOLD-responses overlap, and these overlapping responses cannot be separated unless multiple trials are combined into a single regressor. Furthermore, the trial-by-trial changes in neural response patterns were clearest when the order of stimulus presentation was fixed, such that temporal autocorrelations affected the comparisons of interest to a similar degree. Finally, the data confirmed that these slow event-related designs are inefficient for univariate analyses (where multiple trials are combined into a single regressor), emphasizing that it is important to decide on the type of data analysis before carrying out an experiment.

In **chapter 6**, we examined how a strong, previously formed fear memory influences stimulus processing. Specifically, we used support vector machine (SVM) classification on distributed patterns of activity to assess how the brain (mis)identifies ambiguous information in spider fear. In line with previous findings, individuals with high spider fear were more likely to classify ambiguous morphs as spiders than individuals with low spider fear. To our surprise, SVM classification in functional regions of interest did not reveal a clear bias in the classification of morphs in high fearful individuals. On the contrary: response patterns in visual association areas were more likely to be classified as spiders when individuals were *not* afraid of spiders. Although preliminary, these results tentatively suggest that the phobia-related overgeneralization of fear is not a perceptual phenomenon, but emerges at a later stage of information processing. Univariate analysis showed more activation in individuals with high fear of spiders in a number of areas. In contrast with the MVPA results, this heightened sensory sensitivity was independent of stimulus type and thus appeared to be rather nonspecific, fitting well with the commonly observed attentional bias in fearful individuals. These findings support the notion that univariate analysis and multi-voxel pattern analysis tell complementary stories and that the combination of these methods may be especially valuable for disentangling different neural processes involved in the formation and expression of human (fear) memory.

In sum, there are many ways to assess different levels of associative fear memory in humans, using subjective, behavioral and neuronal measures. Our findings provide new insights into the different levels at which fear associations emerge, reside and become activated, offering a framework to examine the conditions under which fear learning results in long-lasting fear memory.