The neural dynamics of fear memory

Visser, R.M.

Publication date
2016
Document Version
Final published version

Chapter 1

General introduction

Dissociating memory and response systems: the black box

While much of what we learn will be forgotten with the passage of time, emotional memory appears to be particularly resilient to forgetting (Bradley, Greenwald, Petry, & Lang, 1992; Hamann, 2001). Long-lasting memory for emotional experiences is in principle adaptive, but can become dysfunctional such as in anxiety disorders. Despite everything we know about the processes that can weaken or strengthen memory for fearful events, our understanding of how these events are represented, processed and eventually engraved into the neural architecture of the brain 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. To use the words of William James: "The only proof of there being retention is that recall actually takes place" (James, 1890). A deep-rooted debate in psychology centers on the question of how to study the mechanisms underlying observable behavior. Behaviorists guided by John B. Watson, and later Burrhus F. Skinner, formulated a clear answer to that question: we should not. Psychologists should restrict themselves to the scientific study of objectively observable behavior (Watson, 1913). The 'black box' metaphor is often used to describe the behaviorist's view on the human mind; only the input and output of that box are worth studying. However, memories are not exact copies of events (sensory input), as is evident from the way that events are recalled (behavioral output). Memories are not recorded, but constructed; shaped by prior experiences, beliefs, expectations, hopes and desires (Schacter, Gilbert, & Wegner, 2011). This unique conglomeration of private experiences begins to form in early childhood and continues to expand and change until the day we die. Memory is the entity that connects past experiences and future behavior; it is the core of our identity. If one aims to understand human behavior, 'memory' and related 'mental' concepts cannot be ignored. Hence, for the last six decades memory research has been dominated by cognitive psychology, making inferences about what happens inside the black box based on its input and output.

In memory research, a variety of behavioral measures provide different 'read-outs' of information that has been stored. In this thesis I will use the term 'behavioral responses' to refer to any output that can be directly observed - including peripheral physiology, actions or action tendencies, and (subjective) verbal reports - in order to distinguish this level of responding from the underlying neural processes that fuel these responses.

Whereas in many situations different behavioral indices converge, the following anecdote illustrates that knowledge often derives from situations in which different indices do *not* converge. The Swiss neurologist Claparède (Feinstein, Duff, & Tranel, 2010) concealed a sharp pin between his fingers while greeting one of his amnesic patients with a handshake. Even though the patient reacted with surprise and anger, she forgot the encounter within minutes. However, when the neurologist

tried to reintroduce himself shortly thereafter, the patient resolutely refused to shake his hand. She explained her reaction by stating that she was afraid that perhaps a pin was hidden in his hand, but even after repeated questioning could not remember that she herself had been stuck with it. While her brain was apparently able to form this (adaptive) association between a neutral handshake and a painful consequence, her brain was not able to consciously retrieve the event. The declarative part of the memory had not been stored.

Clinical cases like these, substantiated by abundant scientific evidence (Bradley, Miccoli, Escrig, & Lang, 2008; Hamm & Vaitl, 1996; James et al., 2015; Mauss & Robinson, 2009; Sevenster, Beckers, & Kindt, 2012a; Soeter & Kindt, 2010; Weike, Schupp, & Hamm, 2007) teach us that our mind is made up of a constellation of agents that are in principle separable and rely on semi-independent circuits. In cognitive psychology, a classic distinction is made between 'explicit' or 'declarative' memory (what we can report) and 'implicit' or 'procedural' memory (trained or reinforced skills, including conditioned responses) (Graf & Schacter, 1985, 1987; Squire, 2004). These memory systems can be independently influenced (James et al., 2015; Kindt, Soeter, & Vervliet, 2009; Sevenster et al., 2012a; Sevenster, Beckers, & Kindt, 2012b; Soeter & Kindt, 2010, 2015b) or damaged (Adolphs, Tranel, & Buchanan, 2005; Bechara et al., 1995; LaBar, LeDoux, Spencer, & Phelps, 1995; Weike et al., 2005).

Although different behavioral measures can inform us about different types of memory, they only reflect a small portion of the neural processes that are simultaneously active at a certain moment in time: The amnesic patient described by Claparède may be perfectly capable of understanding her neurologist's name, and may be able to recognize his face on the short term, as long as this information is actively rehearsed in working memory. Her behavior during the event itself does not indicate that she is not actually forming an explicit memory. Even with an intact hippocampus many people find it hard to remember names, or what they had for dinner the night before. The fact that someone is learning (i.e., actively, consciously paying attention to a stimulus or rehearsing information) does not imply that a memory is being formed. And even if a memory has been formed, this does not mean that it is observable at any given moment. To understand how fearful events are transformed into durable memory traces, we need indices that allow us to uncover processes involved in the formation of fear memory.

The formation of fear associations

A question that has been engaging clinicians for decades is why some people develop anxiety disorders and others do not. Being the victim of a violent robbery will usually result in a strong memory for that event. This is in principle adaptive: the ability to store relevant information prepares us for future events and promotes survival (LeDoux, 2003; McGaugh & Roozendaal, 2002).

Our physiological responses, our action tendencies, as well as our subjective feelings of fear remind us that we should avoid a particular situation and prevent us from getting hurt.

Given that a known threat can take many forms, generalization of stimulus-outcome associations is a critical aspect of associative learning, as it facilitates a fast response to threat without requiring learning on every separate occasion. While the formation of a generalized associative memory is in principal adaptive, it can turn into maladaptive behavior when nonthreatening stimuli or contexts are inappropriately treated as harmful. Fear of robberies should not detain anyone from going to a supermarket, while unpleasant associations with tarantulas should not transfer to house spiders, butterflies or guinea pigs. Indeed, *over*generalization of fear is thought to be a key feature of anxiety disorders (Bishop, Aguirre, Nunez-Elizalde, & Toker, 2015; Dymond, Dunsmoor, Vervliet, Roche, & Hermans, 2014; Haaker et al., 2015; Kong, Monje, Hirsch, & Pollak, 2014; Lissek et al., 2005, 2014; Mineka & Zinbarg, 2006).

To understand 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 (Pavlov, 1927), which is well suited for research across species (LeDoux, 2003; Rescorla & Holland, 1982). 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 learning'. In animals, fear responses are typically measured by assessing the amount of freezing in response to the CS, or avoidance behavior; in humans common measures include skin conductance responses, acoustic startle responses, heart rate, pupil dilation responses, action tendencies and verbal report (but see for an alternative use of the term 'fear' LeDoux, 2014).

As mentioned before, an inherent restriction of memory research, including fear conditioning, is that we can only infer fear memory from the degree to which behavior during learning (e.g., freezing in rats, physiological responding in humans) overlaps with behavior at a later retention test. Yet, much of what we learn does eventually not transform into long-term memory. Most people respond with terror during a highly aversive experience, but the fact that only a few of them develop a disorder indicates that people may form a completely different memory of a similar event. In fear conditioning, the dissociation between learning and memory has been most convincingly illustrated by experiments in which pharmacological manipulations - administered immediately after learning - induced full amnesia at long-term, while leaving short-term memory

intact (Miserendino, Sananes, Melia, & Davis, 1990; Schafe & LeDoux, 2000). Post-learning processes account for this dissociation, as they induce the synaptic changes, including long-term potentiation (LTP), underlying the stabilization of a memory trace after its acquisition (i.e., consolidation) (McGaugh, 1966; Pape & Pare, 2010). A neurobiological account of memory consolidation has proposed that the selection of information for long-term memory is orchestrated by a neuronal 'tagging' mechanism, which tags newly formed and initially unstable memories for later stabilization (Frey & Morris, 1997; Lesburguères et al., 2011; Redondo & Morris, 2011). Without postulating ideas about how this tagging occurs in humans, a neural read-out that at the time of encoding, or shortly after, could indicate whether information is being selected for subsequent consolidation would open new avenues for studying the formation of memory before it is expressed, and could offer insights into the processes that weaken and strengthen memory for salient events.

Cognitive neuroscience: modern behaviorism?

Instead of ignoring the black box as behaviorists do, or merely theorizing about the black box as cognitive psychologists do, cognitive neuroscientists try to reconcile the two traditions by glancing into the black box to actually observe the processes that build up to behavior. In that sense, cognitive neuroscience could be viewed as a type of modern behaviorism. The range of methods that is used in cognitive neuroscience is broad, including for instance positron emission tomography (PET), electroencephalography (EEG) and (functional) magnetic resonance imaging (fMRI). Especially fMRI has - due to its non-invasive nature - become incredibly popular over the last two decades for studying neural processes in humans. The primary form of fMRI uses the blood-oxygen-level dependent (BOLD) contrast to indirectly assess neural activity in the brain, by imaging local changes in blood flow (hemodynamic response) related to neuronal energy use.

Traditionally, fMRI is used to identify brain areas that are involved in a particular task. By comparing mean activation in one condition to activation in another, researchers infer that an area is preferentially involved in a particular task. Research on emotional memory has identified brain areas that are involved in the acquisition, extinction and generalization of fear (e.g., Dymond et al., 2014; Fullana et al., 2015; Sehlmeyer et al., 2009) and that are hyperactive in specific phobias (Ipser, Singh, & Stein, 2013). Though informative with regard to questions about localization of functions, many of these studies use fMRI merely as a cross-validation technique, that is, they focus on areas that parallel certain behavioral responses. The most intriguing question however is whether fMRI can also tell us something about processes that cannot be observed while they take place, but have to be retrospectively inferred from behavior. To answer this question, traditional fMRI approaches seem to fall short, as they rely on the assumption that different processes lead to differences in overall BOLD signal and can be measured along a single continuum. Much evidence indicates that this assumption is often flawed. Instead, it seems that there is more information in the relationships

between voxels, that is, in distributed patterns of activation (for an analogy see Box 1). For the last few years, the focus has therefore been shifting towards understanding patterns of activation rather than localized blobs.

Box 1: The distinction between graduation ceremonies and cross-dressing parties

Imagine we have a class of students and we are interested in how their choice of shoes (voxel tuning) is influenced by the occasion (experimental condition). We start by comparing average heel height on a regular school day with average heel height on graduation day. Say the average heel height on a regular school day is 2.5 cm, but on graduation day people suit up, more girls will put on heels and now the average height is 3.5 cm. There is a significant difference in mean heel height as a function of 'special-occasionness'. In this case, average heel height seems a valid way to measure this. Now imagine that in order to celebrate their graduation students decide to organize a cross-dressing party in the evening: guys put on heels; girls put on sneakers. Because the number of guys that put on heels roughly matched the number of girls that take off their heels, the average height is still 3.5 cm, the same as during the graduation ceremony. Yet, it is immediately apparent that choice of shoes is different on these occasions. By averaging over heel heights we lose information. Alternatively, we can preserve the information about individual heel heights and simply correlate the different heel heights between different occasions. Events that are quite similar (two consecutive school days) will lead to higher correlations than events that are very different (regular school day versus cross-dressing party). By using the variance across heel heights (or voxels), we can distinguish between different types of 'special-ocassionnes' (conditions), moving from a unidimensional scale (more or less special) to a multidimensional scale (more or less special, but also casual, formal, crazy, etcetera).

The power of patterns

Haxby and colleagues (2001) were the first to show that the evaluation of spatial patterns of activation – so called multi-voxel pattern analysis – is a sensitive and straightforward method for obtaining information about the neural representation of a stimulus or process. In their famous paper, they (Haxby et al., 2001) showed that it is possible to discriminate different stimulus categories based on their unique neural response pattern, an effect that was not driven by how much activation these stimuli evoked on average. MVPA enables us to decode the content of visual input and thought, and is therefore often referred to as a 'mind-reading' technique (Norman, Polyn, Detre, & Haxby, 2006). Over the last decade, numerous studies have underscored the superior sensitivity of MVPA compared to analysis of average activation for reading cognitive states from BOLD-MRI data (Haxby et al., 2001; Haynes & Rees, 2005; Kamitani & Tong, 2005) and quantifying

the relationships between patterns induced by different states or stimuli (Kriegeskorte, Mur, & Bandettini, 2008). Two of the most prominent applications of this technique include (binary) classification analysis and (continuous) similarity analysis. First, in classification analyses some of the response patterns are used to train a classifier (e.g., a support vector machine, SVM) and other patterns are used to test the classifier. If two classes can be discriminated in a certain brain area (classification performance is above chance), this suggests that a stimulus or process is encoded in that particular area. A limitation of classification analysis is the need to train the algorithm, which requires the repetition of stimuli during the experiment. If the representation of a stimulus is expected to be stable this is not a problem. However, if the dynamic change in representations were of specific interest, multiple stimulus repetitions would obscure the effects of interest (Chadwick, Bonnici, & Maguire, 2012). Alternatively, in representational similarity analysis (Kriegeskorte et al., 2008) similarity values (e.g., Pearson's r) are calculated between individual response patterns, resulting in matrices that display the representational (dis)similarity between single stimuli or trials. This type of analysis seems especially promising for quantifying changes over time and thus for fear learning paradigms.

Aim and outline of the present thesis

In the present thesis we aimed to disentangle different neural processes involved in the formation and expression of human fear memory. Given the potential of multi-voxel pattern analysis for quantifying detailed information about the neural representation of stimuli, we applied this technique in a series of fMRI experiments, in order to i) assess fear learning and memory, ii) assess the neuromodulatory effects of stress hormones on fear learning and memory, and iii) examine how a previously established, rigid fear memory affects stimulus processing.

The purpose of the first study was to find a way to measure the dynamic nature of associative fear learning. Up until then, most studies assessed fear learning by averaging across a number of trials and quantifying the differences between reinforced and unreinforced trials. While concealing the learning process itself, this averaging seemed necessary because of the low signal-to-noise ratio (SNR) and the sluggishness of the hemodynamic response. The question we asked was whether the reduction in SNR due to a decrease in variance of the explanatory variable (temporal domain) could be compensated by single-trial MVPA, by evaluating multiple voxels at the same time (spatial domain). The approach used in **chapter 2** was inspired by two seminal papers. First of all, previous research (Li, Howard, Parrish, & Gottfried, 2008) showed that associative fear alters the neural representation of a (previously) neutral stimulus pair, measured before and after conditioning. Second, representational similarity analysis (Kriegeskorte et al., 2008) offered a framework in which the information that is carried by a given representation, related to a particular trial, could be directly compared to the information carried by another representation. By monitoring the trial-by-

trial change in representational similarity, this framework allowed us to examine whether the acquisition of aversive associations would affect the neural representation of neutral stimuli. Moreover, it offered the potential to examine how new knowledge is incorporated into existing semantic networks, assessing how the brain categorizes stimuli according to pre-existing and emerging associations.

Building on the findings from chapter 2, in **chapter 3** we tested whether the observed changes in similarity structure solely mirrored the various behavioral responses during fear learning, or whether they would somehow reveal processes related to the formation of long-term fear memory. In that case similarity analysis would be more than a cross-validation technique for assessing transient learning-dependent changes. Indeed, neural pattern similarity during training had already been proven successful in predicting relatively short-term (1-6 h) declarative memory performance (Xue et al., 2010). However, this effect could be explained by enhanced conscious processing of the subsequently remembered items (Schurger, Pereira, Treisman, & Cohen, 2010; Xue et al., 2010) and thus did not necessarily dissociate from what could be observed in behavior. In this study we focused on a different type of memory, that is, the physiological responding to conditioned stimuli. This automatic emotional expression of fear is a major component of fear memory and, as explained, can dissociate substantially from the conscious processing of threat and the factual recollection of events. We combined fMRI with a concurrent behavioral measure of associative fear learning (i.e., pupil dilation response) during fear conditioning and a memory-test, a few weeks later. We again employed trial-by-trial representational similarity analysis in order to examine the formation, activation and extinction of fear associations, and crucially, assess whether pattern similarity predicts the long-term expression of fear memory.

In **chapter 4** we sought to replicate and extend the findings from chapter 3, by further testing what type of information can be extracted from patterns of activation during the encoding of fear associations. An important question in clinical psychology is whether the processes that lie at the root of the development of abnormal fears are already active during the initial phase of associative fear learning, or whether they are predominantly active during the post-encoding consolidation phase. To mimic abnormal fear, we aimed to modulate the strength of fear memory with an α2-adrenoceptor antagonist (yohimbine HCl) in an experiment consisting of three sessions (separated by 48 hours and 2-4 weeks respectively). We tested the strength of the fear memory in a number of ways (Bouton, 2002): the speed at which extinction occurs (stronger fear memory usually yields slower extinction), reinstatement of fear after the presentation of an unsignaled UCS, generalization of fear to stimuli that resemble the CS+, renewal of fear due to a switch of context, and finally, the speed of reacquisition of fear (stronger fear memory facilitates relearning of the CS+/UCS pairings after successful extinction). During fear learning, extinction, reinstatement and generalization of fear (session 1 and 2) we measured BOLD activation and pupil dilation responses;

during renewal and reacquisition (session 3) we measured fear-potentiated startle responses. The aim of these tests was to determine at what time point it is possible to detect the enhancing effects of noradrenaline on fear memory consolidation.

The designs used in chapter 2, 3 and 4 were optimized for single-trial similarity analysis and differed in a number of ways from regular event-related fMRI designs. Crucial features included the fact that these designs were slow event-related and that the order of stimulus repetitions was not randomized, jittered or optimized using standard algorithms (Kao, Mandal, Lazar, & Stufken, 2009), but designed in such a way that the time between consecutive presentations of a stimulus type was the same across stimulus types. We also used a partial reinforcement paradigm in order to prevent shock-related confounds in the estimation of CS-related activation patterns. The design utilized in study 2, 3 and 4 was based on data from pilot studies. While usually only the final protocol is published, the process of fine-tuning an experimental procedure can be very informative for researchers trying to replicate a published finding, especially when the attempted replication is conceptual in nature (i.e., not an exact copy of the design). Indeed, at some point we learned that other labs were conducting similar analyses on fear-conditioning data using a rapid event-related fMRI design, but had difficulties replicating our effects. Based on our pilot data, we hypothesized that the key to successful single-trial analysis would depend on the length of the inter-stimulus intervals. In **chapter 5** we systematically examined the effects of different designs on single-trial pattern analysis in general, and the ability to assess the dynamics of fear learning in particular. In Experiment 1 we employed slow event-related fMRI combined with classical fear conditioning to assess associative learning in a trial-by-trial manner in designs that differed in trial spacing and trial ordering (8.1-18.5 s). In Experiment 2 we examined the discriminability of stimulus categories (no learning), using rapid event-related fMRI in designs that varied in number and spacing of trials (2-6 s, with and without null-events, equal scan durations). Together, these experiments were intended to provide researchers with information about how to optimize designs for single-trial pattern analysis, thereby also facilitating any attempt to replicate our previous findings.

In chapters 2, 3 and 4 we examined how fear conditioning reconfigures neural circuits to form new associative networks. In **chapter 6** we took this approach one step further by examining how a previously formed fear *memory*, already consolidated into a rigid associative network, influences stimulus processing. A hallmark of anxiety disorders, including specific phobias, is the inability to recognize safety cues and the tendency to perceive ambiguous stimuli as threatening (Bishop et al., 2015; Dymond et al., 2014; Haaker et al., 2015; Kong et al., 2014; Lissek et al., 2005, 2014; Mineka & Zinbarg, 2006). An intriguing question is at what stage in the information-processing sequence an established fear memory influences ambiguity resolution: is this a process that occurs during the initial perception and categorization of a stimulus or that emerges during the subsequent interpretation of a stimulus? While behavioral indices alone could not distinguish between these

different scenarios, support for one of the scenarios could be obtained by assessing how different brain areas classify ambiguous stimuli. Translated to brain networks the question is whether overgeneralization of fear is the result of misidentification of stimuli in 'low-level' visual areas, in cortical and subcortical areas associated with salience processing, or the result of strategies employed in 'higher' cortical areas, including those that are associated with deliberate decision making. In this study, individuals with low spider fear and high spider fear underwent functional MRI scanning while viewing series of schematic flowers morphing to spiders. Participants were required to indicate for each picture whether they saw a spider, flower or none of the two. We classified neural response patterns related to the morphs using a linear support vector machine, trained on an independent set of pictures of flowers and spiders, to identify regions that code for the behavioral overgeneralization observed in spider fear. The identification of these regions could provide clues about where in the information processing stream stimuli are (mis)classified as threatening.