



UvA-DARE (Digital Academic Repository)

The neural dynamics of fear memory

Visser, R.M.

Publication date

2016

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Visser, R. M. (2016). *The neural dynamics of fear memory*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Chapter 7

General discussion

Our brain is a busy organ, capable of simultaneously perceiving, processing, acting and reflecting upon its environment. Only a small portion of what is concurrently happening in our brain is translated into actions and peripheral physiology; an even smaller portion of these processes is verbally accessible. The discrepancy between the expression of fear during learning and the actual formation of a fear memory has thus far complicated research on emotional memory. While neuroimaging techniques seem intuitively promising in revealing the parallel processes that take place during memory formation, traditional applications of these techniques often fail to quantify the moment-to-moment changes in associative networks as a function of learning. Conversely, while behavioral and peripheral measures of fear often do capture these moment-to-moment changes, these measures only reflect the fear *at that moment* and do not predict long-term expression of fear. In this thesis we used multi-voxel pattern analysis to disentangle different neural processes involved in the formation and expression of human fear memory, combining several experimental procedures (Pavlovian fear conditioning, a pharmacological manipulation, and a perceptual judgment task) with different behavioral and neural indices of fear (BOLD-MRI, pupil dilation, fear potentiated startle, perceptual decisions and subjective reports). This general discussion will start with a summary of the findings for each of the individual studies presented in this thesis. Hereafter, the findings are interpreted from different perspectives. Possible implications, limitations, and future directions that have not been addressed in the previous chapters will be discussed.

Summary of findings

In **chapter 2** we aimed to measure the dynamic nature of associative fear learning. By applying MVPA in a trial-by-trial manner, we quantified changes in patterns of BOLD-MRI over the course of fear conditioning. Our findings showed an increase in similarity of neural response patterns on consecutive trials of stimuli that were followed by a nocuous stimulus (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, comparable to the curves that are observed using peripheral measures of fear learning (Soeter & Kindt, 2010). 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 stimuli were paired with an aversive consequence. 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 outcome, 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 tuned more towards 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.

Even though it is tempting to interpret the changes in representational similarity as a sign of fear memory (Bach et al., 2011; Dunsmoor et al., 2014), they were not examined in relation to a later, independent test of fear. 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. This would indicate that - in the case of fear memory research - MVPA is more than a cross-validation technique for assessing transient learning-dependent changes. In this study, we used trial-by-trial MPVA, as described in chapter 2, to quantify learning-dependent changes. Next, we linked these changes to a behavioral expression of long-term fear memory, which was measured by 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 functional 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. 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). To our surprise, we did not observe an effect of yohimbine HCl (an α_2 -adrenergic auto-receptor antagonist that increases central noradrenergic activity) on markers of autonomic arousal, such as salivary α -amylase (sAA). However, 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 (Frey & Morris, 1997; Lesburguères et al., 2011; Redondo & Morris, 2011) is related to noradrenaline (Johansen et al., 2014) 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. Representational similarity analysis on data from Experiment 1 revealed clear learning curves in all conditions, but showed the strongest effects when trial order was counterbalanced, such that temporal autocorrelations affected the comparisons of interest to a similar degree. Furthermore, support vector machine (SVM) classification on data from Experiment 1 and 2 showed that classification of stimulus category was above chance in all conditions and comparable for different pattern estimation techniques (Least Squares Single and Least Squares All). Yet, the data indicated that - given a fixed amount of time - longer intervals are preferred over more stimulus repetitions for single-trial pattern analysis, while at the same time confirming that these designs are inefficient for univariate analyses. These findings were in line with our hypothesis that the key to successful single-trial analysis depends on the length of the inter-stimulus intervals, since overlapping BOLD signals cannot be decorrelated unless multiple trials are combined into a single regressor. The results furthermore emphasize the importance of deciding on the type of data analysis before carrying out an experiment.

In **chapter 6**, we examined how a strong, previously formed fear memory influenced stimulus processing. Specifically, we used SVM classification on distributed patterns of activity to assess how the brain (mis)interprets ambiguous information in spider fear. In line with previous findings (Kolassa et al., 2007), 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 ROIs 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 overgeneralization of fear in phobic individuals 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 (Bar-Haim et al., 2007). Together, these findings support the notion that univariate analysis and multi-voxel pattern analysis tell complementary stories (Jimura & Poldrack, 2012). The combination of these methods may be especially valuable for disentangling different neural processes involved in the formation and expression of human (fear) memory, although it does seem to require special design considerations.

What is driving the change in neural pattern similarity?

In the previous chapters we showed that patterns of BOLD-MRI carry information about stimulus identity and stimulus significance. This information can be used to decode stimuli or predict future behavior. However, the human brain consists of neurons, not of voxels. Eventually, our aim is to understand how neuronal signaling and neuromodulatory mechanisms give rise to cognition, not changes in MR signal. At this point, a straightforward question is what constitutes (the changes in) patterns of BOLD-MRI. Despite the predictive value of patterns of voxels, which makes them a useful tool in memory research, they may by themselves not inform us about how something is encoded in the brain (Davis et al., 2014). Differences between results found for MVPA and other analysis techniques often merely indicate that one is more sensitive than the other, while the signal that is being detected may be only indirectly related to the process of interest (Chadwick et al., 2012; Davis et al., 2014; Davis & Poldrack, 2013). My interpretation of the data will undoubtedly be outdated in a few years from now, given that this is a relatively young and fast developing field of research (Haxby, 2012). There are also quite a few levels at which this issue could be addressed. In the next sections, I will first present an idea of what may be occurring at the voxel level, before

focusing on the significance of the findings in terms of cognition (perception, attention and memory) and emotion. I will use the rather arbitrary term ‘neural’ (responses or tuning) to refer to the signal measured with BOLD-MRI, while the term ‘neuronal’ (responses or tuning) refers to the cellular level. Then, I will briefly discuss the findings at a more conceptual level, touching upon their clinical implications, followed by suggestions for future research and some concluding remarks.

Pattern metamorphosis: the voxel level

In this section, the term ‘voxel tuning’ refers to the preferential response of a voxel to a particular stimulus or stimulus category (Çukur, Nishimoto, Huth, & Gallant, 2013). In short, I propose four possible scenarios of how neural response patterns might change as a function of learning (Figure 1).

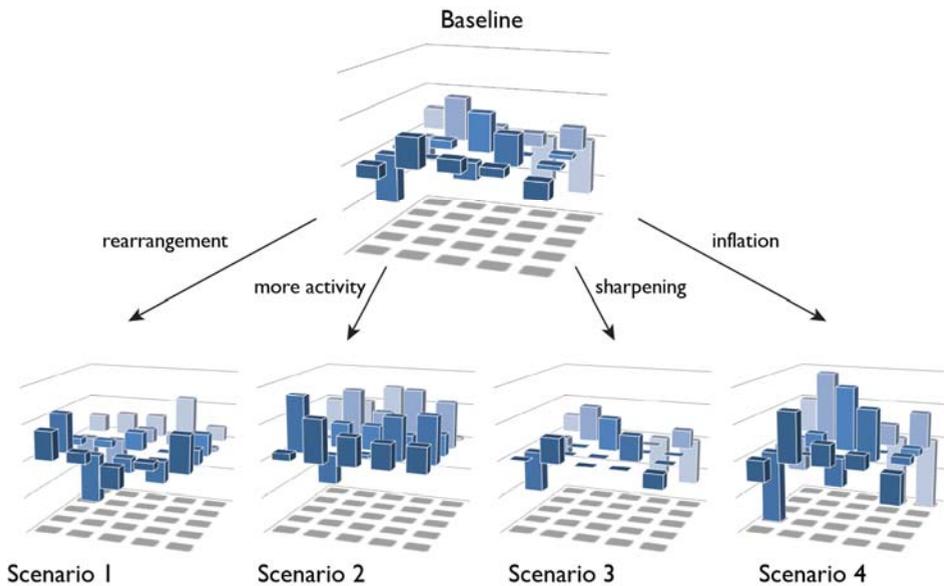


Figure 1. Four possible scenarios of how neural response patterns might change as a function of learning

First, voxels that had a certain tuning completely shift their tuning as a function of learning (i.e., voxels that were activated compared to baseline now become deactivated and vice versa: a rearrangement of the pattern). Second, more voxels become involved in a certain task (i.e., global tuning towards the salient stimuli: more concerted activity). Third, voxels do not change their tuning, but noisy voxels are silenced (global reduction of noise: sharpening of the pattern). Four, voxels that show a certain tuning, will enhance their tuning (activated voxels become more activated

and deactivated voxels become more deactivated: inflation of the pattern). Which of these scenarios best explains our findings may depend on the brain area, but each scenario yields specific testable predictions. Although we have not formally tested all of these scenarios, examination of the data from chapter 2, 3, 4 and 5 provided some clues as to what could have been happening at a voxel level during associative fear learning.

To start with the second scenario, more concerted activity would necessarily predict a change in average signal amplitude. In at least some of the brain areas where changes in pattern similarity were observed (e.g., the superior frontal gyrus) there was no such change in average signal, ruling this scenario out (at least for those areas). It is noteworthy that a global increase in blood flow that is related to the experimental manipulation, yet unrelated to neuronal activity (e.g., changes in heart rate in response to emotional stimuli), would also lead to a net change in average signal. This suggests that the effect that we observe is not merely a physiological artifact caused by emotional arousal. The distinction between the third and the fourth scenario can be made based on noise levels and signal strength. In the third scenario, higher pattern similarity is driven by a reduction of noise and hence more consistent pattern reinstatement. In the fourth scenario, higher similarity would be driven by a stronger signal. In our data we observed an increase in variance for stimuli that were followed by an aversive consequence compared to the control stimuli (data not reported in this thesis), indicating that within a particular region some voxels showed more activation and other voxels showed more deactivation than before conditioning. This renders scenario 3 (no change in signal, solely a reduction in noise) unlikely. Whether the increase in signal additionally coincided with a reduction of noise remains to be explored.

Clearly, the creation of a completely different pattern (scenario 1) would lead to lower correlations between consecutive trials, instead of higher correlations. Nevertheless, it is possible that very early in the experiment a new pattern emerged (scenario 1), which would then be amplified (scenario 4). In this case, there should be a point where two patterns (e.g., before the first UCS administration and after the first UCS administration) are very dissimilar, leading to a low or negative correlation. If there is only an inflation of a pattern that was already present, then there does not have to be such an initial dip. To distinguish between the two scenarios, we searched (per individual) for an initial dip in correlation values calculated over consecutive trials, but - in line with the fourth scenario - similarity seemed to increase steadily over trials (data not reported in this thesis). A preliminary conclusion to draw from this is that aversive conditioning 'inflates' the pattern that is already present.

A general inflation of existing patterns (i.e., amplification in both directions) would also explain the distribution of the effects across the cortex that we described in chapter 2: if a representation is inflated in areas that are retinotopically organized, then this will increase within-stimulus pattern similarity, as response patterns reflect the same (perceptual features of a) stimulus and even more so after conditioning. However, amplification of patterns related to perceptually *dissimilar* stimuli would not increase the (between-stimulus) correlations. In higher cortical areas, response patterns hardly represent specific perceptual features (the identity of the stimulus), but may represent the fact that a stimulus - regardless which - is presented. Amplifying a response pattern will result in higher correlations with other amplified response patterns (between-stimulus pattern similarity). Merely the fact that these stimulus patterns are amplified, reflecting a type of stimulus significance, makes them more alike. This touches upon the next question: How do (changes in) neural response patterns relate to cognitive processes such as perception, attention, memory and emotion, and their underlying neural-endocrinological circuits?

Pattern metamorphosis: attention

In the previous chapters we showed that stimuli that predict threat elicit more stable response patterns than safe stimuli. Several lines of research have demonstrated a close link between pattern stability and behavioral performance, suggesting that greater pattern stability may be a sign of more refined and efficient processing. For example, motor training increases the stability of activation patterns in motor cortices (Huang et al., 2013; Wiestler & Diedrichsen, 2013) and higher pattern stability is associated with conscious (versus unconscious) experiences (Schurger et al., 2010), task-relevant object detection (Li et al., 2009; Zhang et al., 2010) and face recognition (Zhang, Li, Song, & Liu, 2012) as well as better explicit memory (Kuhl, Bainbridge, & Chun, 2012; Xue et al., 2010).

The fact that pattern stability has been nonspecifically associated with different types of performance raises the question whether the effects described in this thesis could be explained by attention. Insofar as the classic neurocognitive interpretation of attention equates attention with more activation - not deactivation - in BOLD and in spike activity (Corbetta & Shulman, 2002; Desimone & Duncan, 1995; Hillyard & Anllo-Vento, 1998), the multivariate results presented in this thesis cannot all be explained by attention, as they were independent of global increases in BOLD-activation. However, recent accounts of attention acknowledge that both activation and deactivation are important for the efficient processing of relevant stimuli. For example, attention during natural vision has been shown to tune voxels in occipitotemporal as well as frontoparietal regions toward certain perceptual categories, while suppressing other (Çukur et al., 2013; Sprague & Serences,

2013). This tuning was observed even when the attended category was not present (in a movie), indicating that it was not a detection-artifact and suggesting that attention alters visual representations to optimize processing of behaviorally relevant objects (Çukur et al., 2013). Since aversive conditioning is a powerful method to make neutral stimuli behaviorally relevant, it is conceivable that the changes in pattern similarity simply indicate that some stimuli are better attended to than others. In retinotopically organized areas this would manifest itself as a sharpening of individual stimulus presentations, whereas in higher cortical areas (where there is no such organization) this would result in a more general arousal pattern. In line with this, our findings in higher cortical areas do not show a clear categorical similarity structure before any learning takes place. In other words, these areas only seem to distinguish between safety and (learned) saliency, without first representing one semantic category and then switching over to another. This suggests that differential pattern similarity in these areas reflects a state, rather than a mnemonic engram.

Pattern metamorphosis: memory

The question is whether there are other areas, for example in the visual cortex, where pattern similarity *does* reflect a memory trace. Obviously, insofar as ‘engram’ refers to the physical entity in the brain that stores information (Golet, Castellucci, Schacher, & Kandel, 1986), engrams cannot be *directly* studied with fMRI. However, given that some memories involve distributed associative storage systems, encompassing multiple regions (Alvarez & Squire, 1994; McClelland, McNaughton, & O’Reilly, 1995; Norman & O’Reilly, 2003), memory engrams may - when activated - elicit subtle changes in regional blood flow that could be detected with functional imaging, enabling their *indirect* investigation. For example, Li and colleagues (2008) used MVPA to show that the neural patterns related to a pair of indiscriminable odor cues became discriminable after conditioning. This behavioral change was paralleled by changes in the representation of these stimuli: while patterns in the olfactory cortex were virtually identical before conditioning, they became distinct after conditioning. Similarly, Bach and colleagues (2011) found a stable sparse representation of conditioned stimuli in the human amygdala, which - as in our studies - was unrelated to average activation.

These effects are reminiscent of findings from animal studies, showing that only a subset of the neurons involved in learning is involved in the encoding of a memory (Rogerson et al., 2014). In a number of brain regions, including amygdala, hippocampus and sensory cortices, the allocation of memory to specific neurons and synapses seems dependent on the presence of proteins such as cyclic AMP-responsive element-binding protein (Sano et al., 2014; Zhou et al., 2009). As only

neurons containing high concentrations of this protein continue to respond during and after learning, the total number of neurons responding to a stimulus decreases, suggesting a refinement of neuronal activation (sparse population coding), which may reflect a memory trace, and possibly more efficient processing (Rogerson et al., 2014). Because this 'refinement of patterns' is observed both at the neuronal level as sparse population coding, and at the systems level as changes in patterns of BOLD-MRI, it is appealing to assume that the two levels are somehow connected. In line with this, many terms derived from computational neuroscience have been used to describe processes at both levels. Pattern separation, for example, refers to the transformation of overlapping memory traces into discrete (orthogonal) traces to enhance their discriminability and limit interference among them, while pattern completion refers to the reinstatement of a complete memory, based on a partial input pattern (Marr, Willshaw, & McNaughton, 1991; McClelland et al., 1995; O'reilly & McClelland, 1994). Using electrophysiology in rodents, evidence for these processes has been obtained in hippocampal areas CA3 and the dentate gyrus (Clelland et al., 2009; Guzowski, Knierim, & Moser, 2004; Lee, Yoganarasimha, Rao, & Knierim, 2004; Leutgeb, Leutgeb, Moser, & Moser, 2007; Leutgeb, Leutgeb, Treves, Moser, & Moser, 2004; Vazdarjanova & Guzowski, 2004). Recently, pattern separation has been used to describe the decorrelation of patterns of BOLD-MRI in the human hippocampus (Bonnici et al., 2012; Lacy, Yassa, Stark, Muftuler, & Stark, 2011). With regard to the results presented in this thesis, we could use the term pattern separation to label the decrease in similarity within an original category, with the simultaneous increase in similarity (over consecutive presentations) of the same stimulus, an effect mainly observed in occipital and inferotemporal regions. Pattern completion could at the voxel level be used to describe the similarity between patterns related to distinct categories, through their pairing with a shock (chapter 2-5), given that two input patterns seem to activate a common representation. Likewise, the individual representations of different exemplars within a certain category (e.g., tools) can become more similar as a function of conditioning (Dunsmoor et al., 2014), or, as shown in chapter 6, may lead to a better classification of distinct (unambiguous) spider pictures in individuals with spider fear.

At the voxel level and the cellular level, terminology derived from computational neuroscience can be insightful. However, given the widespread changes in similarity structure that we observed at the voxel level, it seems implausible that these effects reflect underlying structural changes related to memory formation. The formation and activation of a memory automatically initiates and affects a variety of other processes, such as perception, attention, subjective feelings and action tendencies, all of which we may be sampling with MVPA. For example, it has been shown that it is possible to decode subjective mnemonic states (e.g., a feeling that a face is new or old)

from whole-brain patterns of voxel activity (Rissman, Greely, & Wagner, 2010). The pattern of activation did not seem specific to any particular stimulus, but instead reflected the *cognitive states related to* recognition memory (Chadwick et al., 2012). Especially in studies that involve salient stimuli (Bach et al., 2011; Dunsmoor et al., 2014; Li et al., 2009; Visser et al., 2015, 2013, 2011) changes in patterns may be secondary to mnemonic processes happening elsewhere, at a much smaller scale, and mainly reflect a type of threat-related attention. Corroborating this notion is the fact that the effects that we observed in hippocampus and amygdala were relatively small, despite their generally accepted (modulating) role in the formation of emotional memory (Hermans et al., 2014; Johansen, Cain, Ostroff, & LeDoux, 2011; McGaugh, 2004; McGaugh & Roozendaal, 2002; McIntyre, McGaugh, & Williams, 2012). It is likely that the effects observed in the cortex are driven or modulated by amygdala activity (as suggested in chapter 2), and the release of noradrenaline (as suggested in chapter 4). However, it should be noted that the role of the amygdala in human fear conditioning is still debated as it is rarely confirmed in fMRI studies (Fullana et al., 2015; Mechias et al., 2010; Sehlmeier et al., 2009). This is possibly due to the difficulties in imaging this structure (Mechias et al., 2010; Sehlmeier et al., 2009), or because the amygdala tends to habituate fast if no new information is presented (Büchel et al., 1998; LaBar et al., 1998). Alternatively, decades of convincing animal work may have resulted in an overestimation of the role of the amygdala in associative fear learning in humans.

Another indication for the idea that changes in patterns do *not* map closely onto the physical memory engram is that in chapter 3 we showed that differential pattern similarity was absent when the shock electrodes were not attached, despite the fact that participants previously learned - and had explicit knowledge about - the contingencies. Apparently, the conscious anticipation of an aversive outcome, not just the presentation of a conditioned stimulus, was necessary to reinstate the tuned neural representations. Finally, memory formation is not unique to threatening stimuli. Unreinforced stimuli become safety cues over learning and pairing stimuli with a neutral tone also induces a form of associative learning. Yet, nowhere in the brain were these types of learning expressed in higher pattern similarity. Although these stimuli could be decoded in a recognition test using other types of MVPA (classification), this would probably not directly reflect the memory trace, but would point at a general 'recollective' state (Chadwick et al., 2012; Rissman et al., 2010). It remains to be elucidated whether we could ever use MVPA to discriminate between the different states indirectly triggered by activation of the engram, and local changes in blood flow specifically related to activation of the engram itself.

Pattern metamorphosis: fear

As argued above, it seems that changes in neural patterns primarily reflect 'a state'. Whether this state is purely attentional, or specifically reflects fear, is another question. Obviously, the previously cited studies demonstrate that higher pattern similarity *per se* is not specific for fear (Huang et al., 2013; Kuhl et al., 2012; Schurger et al., 2010; Wiestler & Diedrichsen, 2013; Xue et al., 2010; Zhang et al., 2012). However, pattern similarity may reflect multiple systems, and thus provide distinct read-outs of fear memory, depending on which stimuli are compared and where in the brain the comparison is made. In chapter 2-5 we examined three types of pattern similarity: within-stimulus and between-stimulus correlations, the latter including the similarity between adjacent trials of original-related stimuli (original associations: faces and houses) and the similarity between adjacent target trials that shared (non)reinforcement (learned associations: CS+ face with CS+ house and CS-face with CS- house). To what degree do these different types of pattern similarity relate to other measures of fear?

Behaviorally, the degree to which response systems reflect certain components of fear has been examined by systematically varying dimensions such as valence, arousal and conscious expectations (Bradley & Lang, 2000). For example, evidence shows that the pupil dilation response and the skin conductance response - clearly the most popular measure of fear, especially in neuroimaging research (Fullana et al., 2015; Sehlmeier et al., 2009) - are elicited by non-aversive but arousing events (i.e., positive pictures and reaction time tasks) as well (Bos, Jentgens, Beckers, & Kindt, 2013; Bradley, Codispoti, Cuthbert, & Lang, 2001; Bradley et al., 2008; Critchley, Melmed, Featherstone, Mathias, & Dolan, 2002; Hamm & Vaitl, 1996; Reinhard & Lachnit, 2002; Reinhard et al., 2006). Furthermore, several studies show that skin conductance requires the conscious anticipation of threat and is in that sense more an index of contingency learning than a specific measure of fear (Sevenster et al., 2012a; Sevenster, Beckers, & Kindt, 2014; Weike et al., 2005, 2007). The startle response, on the other hand, seems to be an automatic defensive reflex that is specifically elicited by aversive experiences and is therefore regarded as a reliable index of fear (Hamm & Vaitl, 1996; Sevenster et al., 2012a, 2014; Weike et al., 2005, 2007). As shown in chapter 3 and as explained above, the conscious notion of a threat was necessary to reinstate differential pattern similarity, suggesting that pattern similarity is not purely a measure of valence (Sevenster et al., 2012a, 2014), but like skin conductance and pupil dilation responses reflects anticipatory arousal or attention. However, this seems only partly true as in chapter 3 and to a lesser extent in chapter 4, we observed a striking dissociation between within-stimulus and between-stimulus pattern similarity. While both types of similarity showed learning-dependent changes (acquisition and

extinction), only between-stimulus pattern similarity predicted the long-term expression of fear. More importantly, between-stimulus pattern similarity proved to be a better predictor of later pupil dilation responses than pupil dilation responses themselves during learning. Even though the process that is reflected in between-stimulus similarity may not be specific to fear, the fact that it reflects memory formation has important implications for the study of emotional memory.

Neural markers for emotional memory: conceptual and clinical implications

Neuroimaging has been frequently used for studying memory processes, not in the least because of the intriguing possibility to predict which elements of a learning-experience will later be remembered and which will later be forgotten. The approach that we used in chapter 3 and 4 can be regarded as a 'subsequent memory paradigm' (Paller & Wagner, 2002), which has been combined with both univariate and multivariate analysis techniques to study memory formation (Rissman & Wagner, 2012). Furthermore, consolidation processes have also been directly studied during post-encoding rest-periods (Lewis, Baldassarre, Committeri, Romani, & Corbetta, 2009; Tambini, Ketz, & Davachi, 2010; Tompary, Duncan, & Davachi, 2015; van Kesteren, Fernández, Norris, & Hermans, 2010), yielding several markers that predicted the later conscious recollection of (emotional) items. What makes our application unique is that we demonstrated that neural pattern similarity is not restricted to the arena of declarative memory, but can also be used to successfully predict the subsequent consolidation of long-term procedural memory. Moreover, this marker dissociated both from the conscious recollection of the stimuli (declarative memory) as well as the behavioral expression of fear during learning. As mentioned before, between-stimulus pattern similarity putatively reflects a state and this state seems to be more related to general arousal and attention than to fear. However, it expresses *a type of arousal or attention that thus far has not been quantified behaviorally* and that carries unique information about future consolidation processes. While it seems probable that in declarative memory research other behavioral measures during learning, such as heightened attention or elaborate conscious processing (Schurger et al., 2010; Xue et al., 2010) could predict future retrieval, thus far, only neural pattern similarity has provided a marker for predicting long-term fear memory based on data from the encoding phase. Importantly, this teaches us that the selection of information for storage in long-term procedural memory occurs during learning. Furthermore, since this marker is based on the learning-dependent changes in a trial-by-trial manner, it allows us to see what is happening during encoding. While single-trial analysis is standard for behavioral data in the majority of animal and human fear-conditioning studies, it is still very uncommon in neuroimaging research due to the low signal-to-noise ratio of the BOLD-signal.

Having a neural measure that is sensitive enough for single-trial analysis brings us one step closer to bridging the gap between behavioral and neuroimaging studies on associative fear learning and memory.

The direct utility of the marker in the identification of premorbid risk factors for psychopathology has not been examined in this thesis, but seems limited. First of all, in the studies presented in chapter 3 and 4 no relation was found between the marker and personality characteristics, such as trait anxiety and anxiety sensitivity. Furthermore, our indices for fear learning and retention are calculated by comparing stimuli that predict threat to control stimuli. Higher indices signify more differentiation between threatening and safe stimuli, and thus the formation of a selective fear association. However, the formation of fear memory is in itself very efficient for survival. A hallmark of anxiety disorders is not a strong fear response to a stimulus that predicts threat, but rather the inability to recognize safety cues and the tendency to generalize fear to perceptually or conceptually similar stimuli (Bishop et al., 2015; Dymond et al., 2014; Haaker et al., 2015; Kong et al., 2014; Lissek et al., 2005, 2014; Mineka & Zinbarg, 2006). In conditioning paradigms, (sub)clinical fear has been characterized by *less* differentiation between CS+ and CS- during learning, a slower rate of extinction and more generalization of fear (Blechert, Michael, Vriends, Margraf, & Wilhelm, 2007; Gazendam, Kamphuis, & Kindt, 2013). Our index on the other hand reflects *more* differentiation between CS+ and CS- stimuli during learning and only predicts retention of fear (i.e., the first few trials from the memory phase), not the persistence or generalization of fear. Based on these observations, we would expect that more differential pattern similarity is a sign of adaptive fear learning, rather than a vulnerability factor, reflecting that an individual is in a 'memory-encoding state'. In this state, relevant associations are tagged for subsequent consolidation, using pattern completion to extract commonalities across the different stimulus-outcome associations and using pattern separation to prevent generalized responses to similar stimuli (i.e., the CS-).

The clinical implications of the present work seem to lie in the possibilities for experimentally studying mechanisms involved in fear memory. A marker that indicates whether information is going to be stored opens up new avenues for the identification of (state-dependent) factors that contribute to the aberrant encoding of threatening events. Similarly, the use of MVPA in the study of fear memory has the potential to distinguish between different accounts of overgeneralization of fear in anxiety disorders (although chapter 6 illustrates that this requires special design considerations). Knowing at what stage in the information processing stream stimuli are (mis)classified as threatening could hypothetically guide the selection of interventions, either favoring those that directly target perceptual processes in phobic fear (Grey & Mathews, 2000;

Hakamata et al., 2010), those that target cognitive and behavioral aspects of fear, including in vivo exposure therapy (Choy, Fyer, & Lipsitz, 2007; Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014), or those that are aimed at permanently eliminating fear responses by pharmacologically disrupting the reconsolidation of the affective component of a fear memory (Kindt et al., 2009; Soeter & Kindt, 2015a). In general, using multiple behavioral and neural measures to unravel dissociating systems involved in fear responding could help clarify why rational cognitions so often seem incapable of controlling our subjective feelings of fear.

Integrating memory and response systems: future directions and concluding remarks

In the introduction of this thesis I referred to the once hotly debated question of how memory should be studied, using the term ‘modern behaviorism’ for the idea that cognitive neuroscience offers ways to directly observe processes occurring inside the ‘black box’. Observing changes in patterns of BOLD-MRI however is in many ways just as indirect as any other physiological measure, as it is still unclear what happens between input (a stimulus) and output (the signal that we measure). The question of whether we can ever directly observe a mental concept such as memory is essentially a philosophical issue: at the level at which we can see the neuronal and synaptic changes that code for memory engrams, we lose sight on the broader scale at which memories reside and form associative networks. Although patterns of BOLD-MRI may be closer related to neural networks than patterns of responses on questionnaires and neuropsychological tests, the latter may be more sensitive for detecting (processes related to) post-traumatic stress disorder or preclinical dementia. While cognitive neuroscience has added an indispensable explanatory level between neuronal processes and psychological experiences, the relation between these levels remains to be elucidated. In fact, it is not entirely obvious how, or even whether, we will be able to develop translations from one explanatory level to another (or whether we even should, Lamme, 2006). This challenge is known as the ‘reduction problem’ (Kievit, 2014; Kievit et al., 2011).

Despite these challenges, the last decade has witnessed an upsurge in exciting new developments that at least seem promising for detecting individual memory engrams. A recently developed technique called ‘optogenetics’ uses light to control neurons that have been genetically sensitized to light. This has been used to causally test the real-time contribution of populations of neurons in rodents (Deisseroth, 2011) and has led to a number of groundbreaking discoveries on dissociating mechanisms underlying the formation and expression of fear memory (Huff, Miller, Deisseroth, Moorman, & LaLumiere, 2013; Johansen et al., 2014; Redondo et al., 2014). In humans, advances have been made by combining pattern analysis with additional methods that explicitly test

assumptions about the dimensionality or the content of those patterns (Davis et al., 2014), such as representational geometry (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008) and encoding models (Dumoulin & Wandell, 2008; Kay, Naselaris, Prenger, & Gallant, 2008). While the latter approach, so-called ‘population receptive field’ imaging, has been successful in early visual cortex where there are well-defined models of neuronal population dynamics, it seems challenging to apply the same approach to higher-level regions such as those involved in the formation of memories. Knowing what properties could be usefully modeled in, for example, a hippocampal voxel is harder than modeling activity in early visual cortex, where voxels respond to a relatively constrained set of dimensions (e.g., orientation, spatial location, spatial frequency). The number and type of possible dimensions that a hippocampal voxel could respond to is infinitely more complex (Chadwick et al., 2012). However, the use of high-resolution and high field-strength imaging in humans (Carr, Rissman, & Wagner, 2010), combined with single-unit electrophysiology in rats (Doeller, Barry, & Burgess, 2010), provide important steps toward bridging the gaps between computational models of memory and observations at the cellular level and the voxel level.

Fortunately, even without knowing the precise relations between different levels, each level at which memory can be studied has the potential to yield unique insights in different aspects of fear memory. It should be stressed though that a crucial element of memory, which is still often overlooked in neuroimaging research, is the fact that it evolves over time. ‘Learning’ can - by definition - not be assessed at a single moment and testing for ‘memory’ effects immediately after learning does not aid in identifying factors that contribute to the long-term storage of information. In this thesis we therefore focused on moment-to-moment changes in memory and separated learning and testing phases by periods ranging from several days to weeks. To capture fear memory in its full complexity, including its malleability and its context-dependent reinstatement, the study of emotional memory also cannot rely on a single method. And even if multiple methods are used, the focus should not only be on how these converge, but also on how these dissociate. Neuroimaging research often aims to characterize brain activation that correlates with behavior, which is a valuable method for cross-validation. Yet, it is arguably also the lack of such correlates – the specificity of findings – that fuels our understanding of the processes that build up to behavior.

In sum, there are many ways to assess different levels of associative fear memory in humans, using subjective, behavioral and neural measures. Our findings provide a new perspective on the complex and dynamic nature of fear memory, including 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.