The role of spatial representations in MTL regions in generalizing fear responses

By: Anouk van der Heide (a.vanderheide@donders.ru.nl)

Supervisor: Dr. Erno J. Hermans

Second reader: Dr. Christian F. Döller

Master's Thesis in Cognitive Neuroscience

At the Donders Graduate School for Cognitive Neuroscience

Radboud University Nijmegen, the Netherlands

August 3, 2016

Table of contents

Abstract

Not only exaggerated reactions to threats, but more importantly deficits in suppressing fear in safe contexts can lead to fear-related disorders. It is therefore important to study the link between threat and context encoding. Animal studies demonstrated the importance of hippocampal place cells in representing contextual information in fear learning. Human fMRI virtual reality (VR) studies conducted so far showed that spatial locations within VR environments could be decoded from the (para)hippocampus and that this region also showed different activity patterns for threat and safe contexts, suggesting a role of medial temporal lobe (MTL) regions in contextual encoding. The relation between contextual representations and fear generalization has not been investigated. We hypothesised that inadequate context representations in MTL regions play a major role in overgeneralization of fear. To assess how the strength of contextual representations influences fear generalization, we designed a differential contextual fear conditioning VR paradigm for fMRI and tested it in 24 healthy individuals. We investigated whether we could decode the contexts within the VR from patterns in four MTL regions. Fear generalization was calculated from skin conductance responses (SCR) and compared the differential conditioning response between the safe and threat context. Multi voxel pattern analysis (MVPA) was used to decode activity patterns from MTL regions in the two contexts, where accurate decoding meant that the area contained information to dissociate the contexts. Our results demonstrate that spatial location could be decoded from patterns in the right entorhinal cortex, but not in other MTL regions. No correlation was found between individual SCR generalization scores and decoding accuracies in any MTL region. In conclusion, contextual information seems to be encoded in MTL regions but it remains unclear how this relates to fear generalization.

Keywords: fMRI, virtual reality, MVPA, hippocampus, entorhinal cortex, contextual fear conditioning

Introduction

People suffering from anxiety disorders usually show an exaggerated fear response in a threatening situation (Shin & Liberzon, 2010). The core problem in these disorders however, seems to be that they are unable to suppress this fear response in safe situations. It was therefore suggested that their brain might not encode contextual representations in a sufficient way (Duits et al., 2015; Maren et al., 2013). This suggests that inaccurate encoding of the threatening context results in more fear generalization to safe contexts (Kadosh et al., 2015; Lissek et al., 2014), thereby playing a key role in development of pathological anxiety. However, there have been no studies yet investigating the role of these contextual representations in human fear learning.

Dysregulated fear responses are seen in several anxiety disorders such as social phobia and panic disorder, as well as in post traumatic stress disorder (PTSD) (Parsons & Ressler, 2013). Fear is a normal adaptive reaction of the human stress system to a real or imagined threat. We form memory of this specific situation that required a fear response to be better prepared for future threatening events (Gullone, 2000). This fear learning process consists of two aspects: on the one hand the cue that predicted the threat, on the other hand the context in which the threat appeared. We use contexts to understand the meaning of an event for us in a particular situation (Barrett & Kensinger, 2010; Maren et al., 2013). When these contextual representations are weak, we misinterpret threat in contexts which should actually be recognized as safe. Consequently, overgeneralization of fear is to a smaller extent also present in healthy individuals, caused by weak contextual representations. Therefore, we hypothesise that the relationship between context representation and fear generalization could explain differences in fear responses in healthy individuals.

Fear learning in an experimental setting is typically studied using a fear conditioning paradigm. Fear conditioning involves learning an association between a neutral conditioned stimulus (CS) and an aversive unconditioned stimulus (US), in human research typically a mild shock. This results in a conditioned response (CR) that can eventually be evoked by the CS alone, in absence of a shock. Specific memory of this fear conditioning is assessed by presenting the CS but not the US and measuring the fear response (Parsons & Ressler, 2013). In differential conditioning, two or more CSs are presented, and only one of those presented stimuli (CS^+) is reinforced with a shock (US), whereas the other stimuli (CS-s) are never paired with the US. This results in a learned response to the CS^+ and not to the CS-s. These responses to the different stimuli are typically monitored from psychophysiological measures related to stress such as skin conductance and pupil dilation. This makes statistical comparison possible between responses to CS^+ and CS^- trials. From both animal and human differential conditioning studies it is well

known that the amygdala plays a critical role in encoding, storing and retrieving associations between a cue and an aversive US (Maren et al., 2013; Maren & Quirk, 2004).

The fear response can also be dependent on the context in which the US was initially administered (Alvarez et al., 2008). When the context determines the consequence of the CS, the context is referred to as the occasion setter (Maren et al., 2013). This phenomenon is used in contextual fear conditioning paradigms. Apart from the amygdala, the hippocampus plays an important role in contextual fear conditioning (Alvarez et al., 2008; Baeuchl et al., 2015). Animal contextual fear conditioning studies showed that the amygdala stores information about the CS-US and the context-US association and the hippocampus encodes the context in which the conditioning took place (Kim & Fanselow, 1992; Maren et al., 2013).

There have been several fMRI studies investigating contextual fear conditioning in humans where different contexts were represented using virtual reality (VR) environments (Åhs et al., 2015; Alvarez et al., 2008; Andreatta et al., 2015; Garfinkel et al., 2014). These studies showed that patients with anxiety disorders differ from healthy people not so much in the degree to which they show fear expression to threatening cues, but rather in the degree to which they generalize fear to normal situations. They could not use the safety context to suppress extinguished fear memory, but they were also less effective in using the danger context to enhance fear (Dunsmoor et al., 2011; Garfinkel et al., 2014). This supports the hypothesis that highly anxious people are not able to contextualize the representation of fear cues accurately, so that they cannot distinguish between a safe and unsafe context. Furthermore, these studies confirmed the role of the amygdala in context-US and CS-US encoding in humans, and in addition to this found differences in hippocampal activation between the threat and safe context (Alvarez et al., 2008). Hence, while the role of the amygdala in cued fear acquisition is quite clear and studied intensively, the role of these different hippocampal representations, as well as how cue information is integrated with contextual information in human fear learning, remains unclear from these studies. Studying the role of contextual representations in fear learning could have important implications for development of treatments for anxiety disorders.

Electrophysiological studies, mostly in animals, showed that the hippocampus has specialized place cells that fire only to specific spatial locations within a spatial environment. Each place within an environment is defined by a unique combination of active place cells (Alme et al., 2014; O'Keefe & Burgess, 1996). This spatial mapping allows animals to integrate events with their spatial contexts. When an important property of the context has been changed, this causes a process called remapping, meaning that a new firing pattern is formed. Not only every environment provides distinct and sparsely distributed patterns of stable place fields, even within a single environment the hippocampus represents the spatial, temporal and behavioural context. This role of hippocampal place cells in representing contextual information in

contextual fear conditioning in animals has been confirmed, as well as remapping of relevant information and subsequent memory formation (Moita, 2004). This could provide a causal mechanism by which the hippocampus plays an important role in dissociating contexts in fear learning. However, there is no evidence for this relation from human brain imaging studies yet.

Hassabis et al. (2009) were the first to show that four corners within a virtual room could be discriminated from a participants' hippocampal activity pattern at the moment participants were localized in those corners. Information about the two different virtual environments they used in their experiment (regardless of specific spatial location within that environment) however, was present within activity patterns in the posterior parahippocampal gyrus, but not in the hippocampus. This suggests that both the hippocampus and the parahippocampal gyrus play an important role in representing contextual information in humans. In addition, two other studies by Brown et al. (2016) and Rodriguez (2010) were able to decode specific spatial representations of goal locations from hippocampal activity patterns in an fMRI study where participants navigated in a virtual environment.

In sum, on the one hand there have been various studies investigating contextual fear conditioning using virtual reality paradigms, and on the other hand three studies showed that contextual information about virtual environments could be decoded from the human hippocampus and parahippocampal gyrus using an fMRI virtual reality paradigm. However, to investigate the role of these contextual representations in fear learning, we must combine these two lines of research so that contextual representations can be compared during a fear conditioning paradigm.

Therefore, in earlier phases of this project a virtual reality paradigm was designed that combines differential conditioning with contextual conditioning and can be applied during fMRI scanning. In this virtual environment, there are two buildings on different locations that represent a 'safe' and 'threat' context. In the threat context, mild shocks were administered after CS^+ presentation to stimulate fear learning. In the safe context, there were no shocks and this building should stimulate safety learning. This way, participants learned to dissociate the CS^+ from the CS^- but they also needed to learn that the CS^+ is only signalling the US in the threat, and not in the safe context. Fear generalization was calculated from the skin conductance response, which was recorded during the whole MRI experiment. We first wanted to replicate the findings by Hassabis et al. (2009) by looking whether it was possible to dissociate between the two buildings in the hippocampus and parahippocampal gyrus in the part of the task where participants were trained in the two contexts without fear conditioning. Based on the previous findings on place cells and grid cells we expected that context is represented in sparsely distributed patterns which would probably not be detected by univariate fMRI analysis techniques. Therefore, we used Multi-Voxel Pattern Analysis (MVPA) to decode activity patterns

from four medial temporal lobe (MTL) regions. MVPA differs from standard fMRI analyses by focusing on whether information relating to specific stimuli is encoded in patterns of activity across multiple voxels. MVPA therefore has the advantage to be more sensitive to information coded on a coarsely distributed spatial scale. If the context where the participant is during a certain trial can be accurately predicted by the classifier, this means that there is information about that context represented in the brain region where that pattern across voxels was identified (Chadwick, Bonnici & Maguire, 2012).

The results by Brown et al. (2016), Hassabis et al. (2009) and Rodriguez (2010) suggest that abstract spatial representations are present in the hippocampus as well as the parahippocampal gyrus and that they can be detected from fMRI activity patterns in these regions. In addition, these studies showed that both the hippocampus and the parahippocampal gyrus are viable targets for MVPA approaches. In the current study we therefore investigated both these regions as a priori regions of interest (ROI) for our MVPA approach. Given that not only two-third of all cortical projections to the hippocampus are received through the entorhinal cortex (Burwell & Agster, 2008; Shah, Jhawar & Goel, 2012), but also numerous studies showed that grid cells in the medial entorhinal cortex have a unique contribution to spatial memory (Döller, Barry & Burgess, 2010; Hafting et al., 2005; Sasaki, Leutgeb & Leutgeb, 2015;) and play an important role in learning and retrieval of context-dependent memories in humans (Brown, Hasselmo & Stern, 2014), we also included the entorhinal cortex as ROI to investigate whether it also contains information about spatial contexts. Finally, we chose to include the amygdala as ROI to be able to reconfirm previous findings showing that the amygdala is not coding contextual information in fear conditioning paradigms (Maren et al., 2013).

Taken together, we predicted to find dissociable spatial fMRI brain activity patterns for the two buildings within the VR during the context training part of the task in the hippocampus, the parahippocampal gyrus and the entorhinal cortex, but we did not expect to show this dissociation in the amygdala. We also expected to find dissociable patterns within these areas in the fear conditioning task. Next, we predicted that this dissociation between the buildings as found in the context training task remains detectable within these brain areas during the fear conditioning part, since the buildings remain the same, although some remapping might be present due to fear learning. Furthermore, we expected the generalization of the fear response as measured through skin conductance to be negatively correlated with the accuracy with which the classifier is able to dissociate the two contexts.

Materials and methods

2.1 Participants

Of the twenty-eight participants tested, four had to be excluded due to technical problems with the MRI scanner. Consequently, twenty-four healthy right-handed participants were included in the final sample of this study of which twelve were male (mean age: 24.06 ± 4.38 yrs) and twelve were female (mean age: 22.10 ± 1.98 yrs). We only included participants with no history of psychiatric, neurological or endocrine disorders and with normal or corrected-to-normal vision. All participants gave written informed consent prior to the study and received a compensation of 31 Euros (ten Euros per hour and five Euros extra for electrical stimulation) after completing the experiment. In addition, fMRI screening and medical screening forms were completed before the start of the experiment to determine whether the participant satisfied all criteria for MR safety clearance and whether all inclusion criteria for the experiment were met. The ethical approval for this study was obtained from the local institutional review board (CMO Region Arnhem-Nijmegen, The Netherlands).

2.2 Experimental design and procedure

In this experiment, we developed a contextual differential fear conditioning paradigm in which the context was the occasion setter. We included two contexts and two conditioned stimuli (CS^+ and CS^-).

2.2.1 Contexts

To simulate the two contexts (i.e. threat and safe context for fear conditioning), we created a virtual environment with the Unreal Development Kit (based on Unreal Engine 3; Epic Games, Inc., Cary, NC, United States). In this environment two buildings (i.e. threat and safe context) and a booth were designed, which were all needed to carry out the context training and fear conditioning task (figure 1A). The buildings were different from the outside but identical on the inside. The location of the booth changed throughout the tasks so that the participants constantly needed to be aware of their spatial location within the environment. In addition, there were various objects that were explored during the familiarization task (e.g. a bucket and a fruit stall), and a few trees as well as mountains surrounded the environment. A top view of entire the environment can be viewed in figure 1B. Participants were able to navigate through the virtual environment using three buttons on the button box. These buttons were used to turn left, turn right or walk straight ahead.

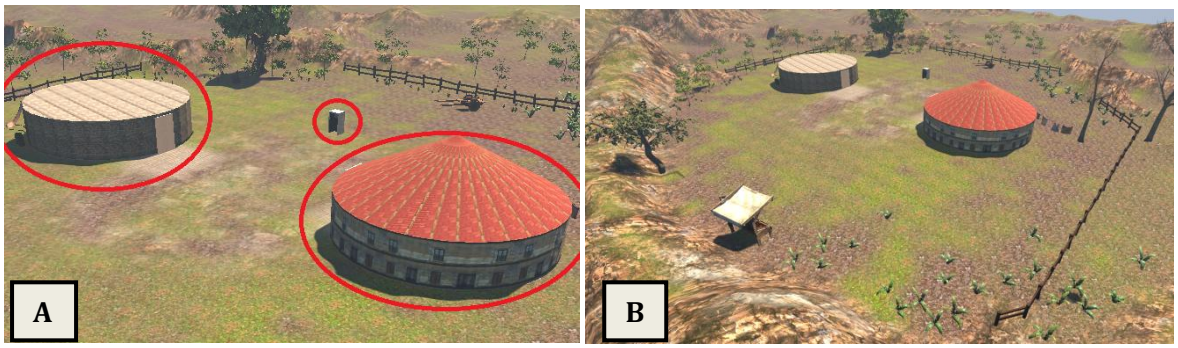


Figure 1: Overview virtual environment. In (A) the two buildings representing the safe and threat context and the booth in which the number was shown that indicated which building would be unlocked for the next trials, are marked with red circles. (B) Shows the virtual environment including the surrounding fence and mountains.

2.2.2 Procedure

Each participant visited the Donders Centre for Cognitive Neuroimaging (Radboud University Nijmegen, the Netherlands) once for a session of approximately three hours. The experiment was scheduled as follows: The participant started with filling in questionnaires. After that, task 1, the familiarization task took place outside the MRI scanner. Task 2, the object localizer task, was the first task in the MRI scanner. After this task, field maps were recorded, followed by task 3, the context training task. Subsequently, we recorded an anatomical T1 scan and finally task 4 took place, the fear conditioning task. Detailed task descriptions will be given below.

2.2.3 Questionnaires, familiarization task and scanner preparation

The first part of this session consisted of six questionnaires assessing symptoms of depression, state and trait anxiety (STAI) and personality traits, followed by an instruction text and some images about the task, all performed on a computer.

Familiarization took place on a computer screen in the scanner room, showing the same virtual environment that was later used within the MRI scanner. The goal of this task was to get familiar with the environment. First, participants got specific instructions about certain objects that they should find in the environment to make sure that they would have been in all corners of the environment. After they discovered all seven objects, they walked to the booth and there they were presented with either number one or number two on the screen (figure 2A). Their first task was to find out which building corresponded to which of the two numbers. Only the door of the building with the number that was just presented opened automatically when participants navigated to that door. The numbers for the buildings remained the same during the whole experiment but between participants this was counterbalanced. Both buildings were empty except for a booth that looked similar to the one outside (figure 2B). When the participant walked straight into the booth the trial started as soon as the viewpoint was transitioned in a way that only the screen was visible, to make sure that the visual input was equal across trials

and most importantly, equal in both buildings. A trial in this part of the task consisted of viewing the fixation cross on the screen for 7 seconds (figure 3A). After the trial ended, the participant walked to the door again. If the door did not open, another trial should be performed in the building (figure 2C). However, if the door opened (figure 2D), the participant navigated outside the building and to the booth outside again, which was now relocated, to find out which building he or she was supposed to enter next.

After the familiarization task, we put the participant in the scanner and attached the psychophysiological measures and shocker electrodes. Shocker electrodes were attached to the distal phalanges of the thumb and little finger of the right hand. Then, the electrical shock intensity was adjusted to the subjective pain threshold of the participant following a standardized shock intensity adjustment. During this calibration procedure, each participant received five shocks which were then subjectively rated on a scale from one to five. This resulted in a final shock intensity which was experienced as annoying but not painful.

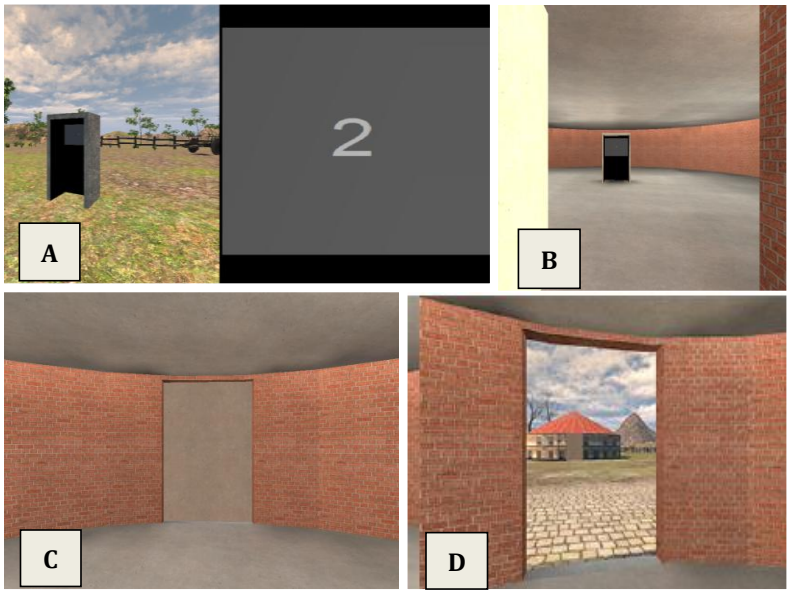


Figure 2: The procedure of a trial. In (A) the booth is shown in which the number is presented of the building the participant should enter next. (B) shows what you see when you enter one of the buildings. When returning from the booth, the door either is still closed (C), or opens (D).

2.2.4 Object localizer task

During the object localizer task, which was the first task that took place in the MRI scanner, twelve different exemplars of two different object types were presented in various blocks. All participants were shown one animate (seagull or tortoise) and one inanimate (pumpkin or zucchini) object type. These were the same object types that we later used as CSs in the fear conditioning task. We chose these specific object categories because previous studies have shown that they result in category-selective clusters of activity in the inferior temporal cortex (Kriegeskorte et al., 2008; Bell et al., 2009; Proklova, Kaiser, & Peelen, 2016). This

allowed us to distinguish between the animate and inanimate object type in the brain, of which one would become associated with an electrical shock in the fear conditioning task. It is important that inferior temporal cortex activation following the CS^+ presentation can be dissociated from the activation following the CS^- presentation, to be able to study activations in other brain areas related to those stimuli. Each block lasted twelve seconds and consisted of twelve exemplars (1 second per image) of one of these two object categories. These were all different exemplars than those that were used in the fear conditioning task, to avoid habituation to the specific images. Between blocks only a fixation cross was presented. Furthermore, two blocks with scrambled images were presented, created through Fourier transformations yielding amplitude and phase components of each image, which were used to rebuild the images (inverse Fourier transformation) with original spatial frequency distribution, luminance and contrast, but with randomly scrambled phase information (Reinders, den Boer, & Büchel, 2005). This resulted in scrambled images that could not be recognized as the object types but could be used as a baseline visual input for the same level of spatial frequency input as the original images.

2.2.5 Context training task

The context training task was included to quantify contextual representations separately from CS representations. The task consisted of 48 trials: 24 per building. Also, the task was divided into two sessions, because the MRI sequence that we used had a maximum duration of 18 minutes. The task was very similar to the familiarization task outside the scanner. Participants used the middle three fingers of their right hand to navigate using the button box. The number of trials before the door of the building opened and participants switched to another building was varied between two and four. Each participant switched between the buildings 16 times in total. The duration of the task differed between participants and depended on the speed with which they navigated through the environment, on average this took 18 minutes.

2.2.6 Fear conditioning task

The fear conditioning task consisted of 60 trials, which were structured differently than in the context training task, as shown in figure 3. Participants received mild shocks during this task and were specifically informed about this beforehand. First the fixation cross was shown again for 7 seconds, and after this the CS was shown for 5 seconds, paired with a mild shock only during the last 500 milliseconds of CS presentation in cases when the CS^+ was presented in the threat context (i.e. the reinforcement rate was 100%)(figure 3B). For the CSs (CS^+ and CS^-) we used an image of the animate object type and an image of the inanimate object type as viewed in the object localizer task, although different exemplars. This was counterbalanced across

participants. We included 15 trials per experimental condition (i.e. CS^+ threat, CS^+ safe, CS^- threat and CS^- safe). Importantly, participants were neither instructed about the fact that one building was safe and one was not, nor about the fact that only one image was paired with a shock. They were only instructed to try and find out in which situations they received a shock. In this task participants switched between the buildings 20 times. On average the duration of this final task was 28 minutes.

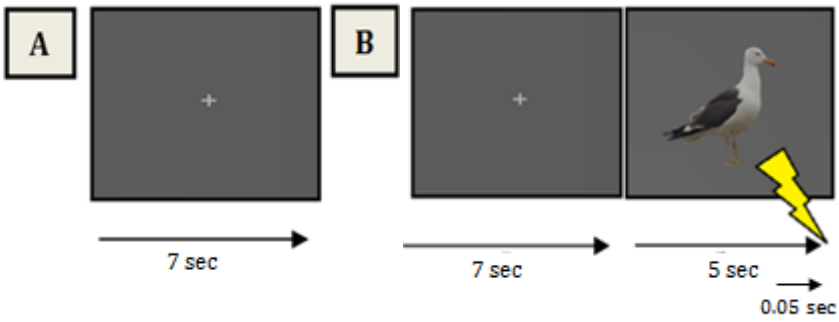


Figure 3: (A) shows the trial structure used in the familiarization and the context training task and (B) shows the trial structure in the fear conditioning task, in which a fixation cross is shown first and thereafter either the CS+ or CS-. The CS+ was then during the last 0.05 second of stimulus presentation paired with a shock.

2.3 Psychophysiological measures

2.3.1 Acquisition of psychophysiological measures

To be able to correct for physiological noise and movement artefacts, we continuously recorded participants' heart rate and respiration during all the tasks performed in the fMRI scanner. In addition, we continuously recorded skin conductance (SCR) and pupil dilation throughout MRI scanning, as a physiological measure of acquired fear, since these are sensitive parameters for emotional and cognitive states, arousal and pain (Bach, Friston & Dolan, 2010).

Heart rate, respiration and SCR were recorded by using an MR-compatible BrainAmp EXG MR 16 channel recording system (Brain Products, Gilching, Germany) at a sampling rate of 5000 Hz. To monitor heart rate we attached a pulse oximeter (Brain Products, Gilching, Germany) to the distal phalanges of the left ring finger. The respiration rate was monitored with a Respiration Belt MR, a sensor for the acquisition of respiratory signals within MR environments, which we attached around the abdomen positioned over the diaphragm. Finally, to measure SCR, electrodermal activity was monitored using two Ag/AgCl electrodes attached to the distal phalanges of the participant's left index and middle finger, to be able to measure conductance directly using 0.5 V constant voltages. We assessed pupil dilation responses using the MR-compatible eye tracking device (MEye Track-LR long range camera unit; SensoMotoric Instruments) and the iView System in the MRI scanner.

2.3.2 Analysis of psychophysiological measures

First, we performed visual artefact correction and peak detection on the raw pulse and respiration data, which we then used as input for retrospective image-based correction of physiological noise artefacts using RETROICOR. This method performs well for reducing both cardiac-induced and respiration-induced noise (Glover, Li, & Ress, 2000; van Buuren et al., 2009). This correction method assumes that the physiological processes are quite periodic so that cardiac and respiratory phases can be uniquely assigned for each recorded image. We used 5th order Fourier basis sets to determine the relation between cardiac and respiratory phase-related modulation of the BOLD signal. This way, we generated ten regressors for cardiac noise and ten for respirational noise, and six additional regressors for heart rate frequency, heart rate variability, abdominal circumference, respiratory frequency, respiratory amplitude, and respiration volume per unit time. In total, this resulted in 26 RETROICOR regressors to be added in the General Linear Model.

To score SCR responses to the different trial types we used Autonomate, which provides automated scoring of stimulus-locked SCR amplitudes (Green et al, 2014). We checked and if needed adapted these automated scores manually. Responses were only taken into account if they started between 0.5 and 3.8 seconds after stimulus onset and if the rise-time was between 0.5 and 4.5 seconds. We always used the first response that occurred. Responses for all trial types were cut off at the time the shock could be given (4.5 sec), to make sure the response to the shocks was never included. Average scores per participant and development of SCR responses over trials for the different trial types could then be evaluated. Additionally, we calculated individual scores for fear generalization to the safe context, based on the average SCR responses per trial type. For this, we used the following formula:

Generalization score = ($SCR_{safe\ CS+}$ - $SCR_{safe\ CS-}$) - ($SCR_{threat\ CS+}$ - $SCR_{threat\ CS-}$)

2.4 fMRI acquisition, preprocessing and analysis

2.4.1 fMRI acquisition

fMRI data was collected at the Donders Centre for Cognitive Neuroimaging. Scanning was performed on a 3T Siemens MAGNETOM Skyra MRI scanner (Siemens Medical Solutions, Erlangen, Germany) using a 32-channel head coil. Participants underwent structural and functional scanning. Functional scans were acquired with a T2*-weighted gradient-echo echo planar imaging (EPI) sequence, using generalized auto-calibrating partially parallel acquisition (GRAPPA) reconstruction (factor two accelerated) (Griswold et al., 2002) and multiband parallel excitation (factor four accelerated) (Blaimer, Choli, Jakob, Griswold, & Breuer, 2013) with full-brain coverage (TR = 909 ms, TE = 24.6 s, FA = 59°, FOV = 212*212 mm, slice-matrix size =

106*106, slice thickness = 2 mm with no gap between slices, voxel size = 2 mm isotropic, 60 axial slices acquired in an interleaved order).

A T1-weighted structural image was acquired for every participant using a Magnetization-Prepared Rapid Gradient Echo (MP-RAGE) sequence (TR = 2300 ms, TE = 3.03 ms, FA = 8°, FOV = 256*256*192 mm, slice-matrix size = 256*256, voxel size = 1 mm isotropic) as an anatomical reference for visualization of activation patterns and to extract our ROIs from the functional scans. In addition, magnetic field inhomogeneity was measured with a fieldmap for each participant to reduce distortion in the functional images (TR = 1020 ms, TE_1 = 10 ms, TE_2 = 12.46 ms, FA = 90°, FOV = 224*224 mm, slice matrix size = 64*64).

2.4.2 fMRI preprocessing

fMRI imaging preprocessing was conducted using Statistical Parametric Mapping SPM8 package (Welcome Department of Cognitive Neurology, University College London, www.fil.ion.ucl.ac.uk) and implemented in MATLAB (The Mathworks Inc., Natick, MA). The five first volumes of each run were discarded to allow for T1equilibration effects. Preprocessing was conducted separately for context training task and the fear conditioning task, each divided into two sessions.

First, fieldmaps were used to calculate a voxel displacement map. Then we realigned and unwarped the functional images to correct for head movement and coregistered them with the structural scan. Additionally we performed a segmentation step to separate grey matter, white matter and CSF. Furthermore, for the univariate analysis we normalized each participant's EPI scans to the Montreal Neurological Institute International Consortium for Brain Mapping (MNI-152) template (Tzourio-Mazoyer et al., 2002). We resampled into 2 mm isotropic voxels and smoothed these normalized images using a 6 mm full-width-at-half-maximum (FWHM) Gaussian kernel. The multivariate analyses (MVPA) were performed in native space of the participants so for these analyses we did not perform normalisation. Finally, we estimated accuracies with the classifier on images that were smoothed with a 3 mm FWHM Gaussian kernel since this was most frequently done in previous MVPA studies. This moderate level of spatial smoothing (2-3 mm kernel) seems to have an optimal effect on MVPA quality with fMRI data, since on the one hand this smoothing increases the signal to noise ratio (SNR) and can control for residual noise artefacts such as head motion in the data, but on the other hand information encoded in sparsely distributed patterns will not be lost due to a larger kernel size because high spatial frequency information is preserved (Gardumi et al., 2016; Misaki, Luh, & Bandettini, 2013; Op de Beeck, 2010).

2.4.3 fMRI analysis: Univariate

In the first level model we included the CS presentations of the four trial types, the shocks and navigation periods. For the fear conditioning task, responses to CS presentation of the four different trial types were modeled as separate regressors using box car functions with a duration of 5 sec. Shocks were modeled as an additional regressor using a 500 ms delta function. The model additionally included six motion parameter regressors (three translations and three rotations) derived from the motion correction and 26 RETROICOR respiratory and cardiac noise regressors, a high pass filter of 1/128 Hz and AR(1) serial correlation correction, and each regressor was convolved with a canonical haemodynamic response function (HRF).

In the second level factorial random effects analysis, betas from the first level analysis were used as input for a 2x2 (context by CS) ANOVA model. We constructed t-contrasts to view main effects of context and CS as well as the interaction effect between context and CS, and activity positively as well as negatively related to this interaction was investigated. Furthermore, we looked at the differential conditioning effect and the effect of the generalization score as covariate on the interaction between context and CS. We defined peak voxels and then reported peak voxels which survived whole-brain family-wise error (FWE) correction (α=.05). In addition, we applied a small volume correction (SVC) for the following a priori ROIs: bilateral amygdala, bilateral hippocampus and the bilateral parahippocampal gyrus, because we wanted to see whether we could replicate findings by previous studies showing that these regions are implicated in storing contextual information (Li, Lu & Zhong, 2016; Maren et al., 2013). We located the brain regions activated from the resulting statistical map using the Automated Anatomical Labeling (AAL) atlas (Tzurio-Mazoyer et al., 2002) in MRIcron (www.mccauslandcenter.sc.edu/mricro/mricron/).

2.4.4 fMRI analysis: Multi Voxel Pattern Analysis

We used Multivariate Pattern Analysis (MVPA) to investigate whether there are context-specific hippocampal, parahippocampal, enthorhinal and amygdala response patterns across multiple voxels. We believed this was possible because MVPA focuses on whether information relating to specific stimuli is encoded in patterns of activity across multiple voxels. With MVPA searchlight, Hassabis et al. (2009) showed that from the hippocampus and the parahippocampal gyrus distributed fMRI activation patterns could be identified which were capable of discriminating between four target positions in a virtual room and between two separate virtual rooms. This suggests that in an fMRI virtual reality paradigm like in our study, it should be possible train a classifier on the different activation pattern in the hippocampus for building A and B. In the case of fMRI, classification consists of determining the decision function that takes various fMRI signals in a cluster of voxels in a certain trial and predicts the class (building A or B) of that trial (Mahmoudi, Takerkart, Regragui, Boussaoud, & Brovelli, 2012).

To create first level models, we calculated beta estimates with separate general lineral models (GLMs) for each individual trial, taking all other trials into account as a second regressor. This way, we could test every trial independent of all other trials. The period taken into account as 'trial' was in this case the period of 7 seconds during which the fixation cross was presented, importantly not including the stimulus presentation in the fear conditioning task (figure 3A). The GLMs only included all trials in the session where this trial took place.

ROI-based analyses are often appropriate for hypothesis-driven MVPA analyses (Chadwick et al., 2012). Regions of interest (ROIs) for the left and right hippocampus, parahippocampal gyrus, amygdala and entorhinal cortex were defined in the native space T1-weighted images of every participant. For this, we used the automated anatomical segmentation with the 'recon-all' function in FreeSurfer 5.3 (http://surfer.nmr.mgh.harvard.edu/). From this whole brain segmentation we could then construct separate ROIs for all the areas of interest. An example of ROIs derived from segmentation performed by FreeSurfer is shown in figure 4.

In addition, we manually divided the hippocampus ROIs over the longitudinal axis resulting in additional posterior, medial and anterior hippocampus ROIs. We did this to be able to investigate the relative contributions of these subregions to classification performance, since previous studies indicate that there is an anterior–posterior neurofunctional gradient (Robinson, Salibi, & Deshpande, 2016), in which the posterior hippocampus is thought to be more specialized in spatial processing (Strange et al., 2014). The precise nature of this functional dissociation is not clear yet, so we tested whether a functional dissociation could be shown in our design.

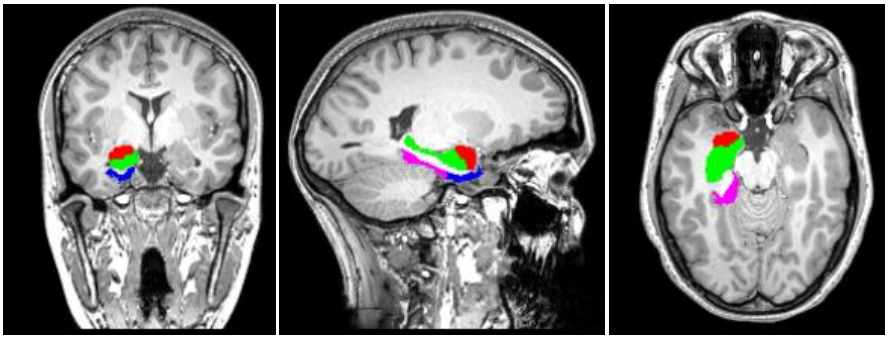


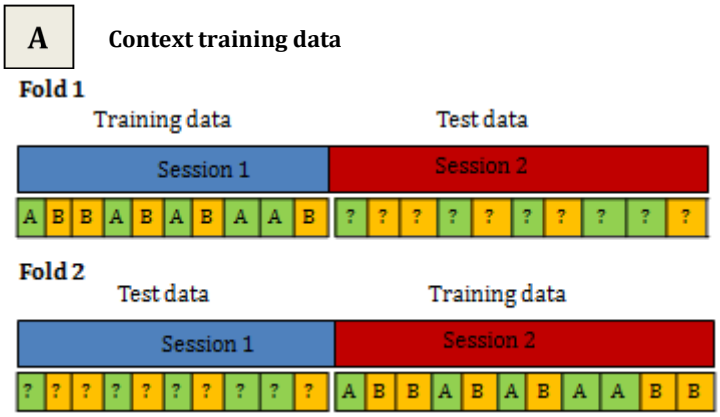
Figure 4: MRIcron view of segmented brain areas as obtained by FreeSurfer, which were used to train the SVM. (Green = hippocampus, red = amygdala, purple = parahippocampal gyrus, blue = entorhinal cortex.)

We then constructed matrices of the estimated single-trial responses across all voxels within the ROIs of interest, both for betas and t-values, and included labels specifying in which building that trial took place. These values and labels were used to train the classifier. In total this yielded separate trial matrices for all ROIs as input for the support vector machine (SVM).

In order for a classifier to successfully decode brain activity, the difference between two conditions must be systematic and consistent across the majority of the training examples. For example this is not the case for the navigation periods between our two contexts which were self-paced and thus led to trial-by-trial variability in terms of the paths taken and the visual input during this route, which in turn could falsely contribute to the classification. To ensure this consistent input we extracted brain activity only when the fixation cross was shown, which was always consistent across contexts. For the same reason, an equal number of trials in building A and B were required as input for the training set to make sure the classifier received the same amount of knowledge about both contexts. For this, some trials at the end of a session were removed if the number of trials in building A and B were unequal.

To ensure independent training and testing data and to avoid temporal correlation between trials, we used a 2-fold cross-validation approach, in which trials that were correlated by being close in time were put in the same fold. Otherwise, the classifier would predict the test trials with a correlated counterpart in the training set (Pereira, Mitchell, & Botvinick, 2009). We first performed separate classifications for the context training and fear conditioning task with 2 folds, equal to the number of sessions in both the tasks, as shown in figure 5A and 5B. In addition, we tested classification accuracy when training on the context training task and applying this to the fear conditioning task, to see whether the representations could be generalized from one task to the other (figure 5C).

The classification was performed with a linear SVM provided by the LIBSVM implementation (Chang & Lin, 2011). Linear mojdels are often preferred to nonlinear approaches in fMRI decoding studies, because they offer more stable results, are less sensitive to overfitting and results are easier to interpret (Kriegeskorte, 2011). Furthermore, we used default settings and z normalized time course data within each trial window.



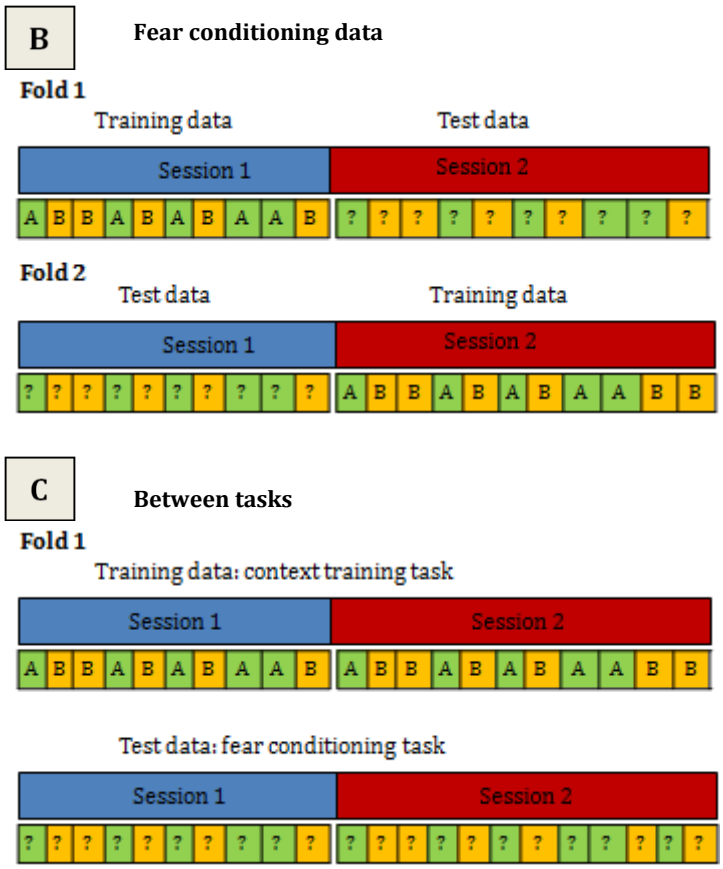


Figure 5: (A) Shows the 2-fold cross validation procedure we used for the classifications of the context training. (B) shows how we used the same 2-fold cross validation we used for the fear conditioning task. Finally, (C) shows how we performed a classification to see whether the training the classifier on the context training task could be generalized to the activation patterns in the fear conditioning task.

The significance levels of decoding accuracies were computed using non-parametric permutation tests (Nichols & Holmes, 2002; Pereira et al., 2009). For each subject, 1000 independent runs of the classification were performed by randomizing the training labels. Voxel accuracies from these permutation runs were concatenated into separate null distributions for each participant, which were converted into z-distributions. Additionally, decoding accuracies for every subject were converted into z-scores using these null distributions to be able to test significance on the individual z-distributions. We performed a one sample t-test to determine whether the null hypothesis could be rejected, i.e. whether the classification significantly above chance on the group level.

2.5 Analysis questionnaires

We first investigated whether participants were able to explicitly mention contingencies of the paradigm using the debriefing questionnaire. Furthermore, we explored interactions between awareness of the paradigm and age, gender and gaming experience of participants. Finally, we expected to find correlations between the STAI score and the generalization score as well as the decoding accuracy.

Results

3.1 Psychophysiological measures

To confirm that our fear conditioning paradigm was successful, we analyzed the skin conductance responses to the CS^+ and CS^- in both the safe and the threat environment.

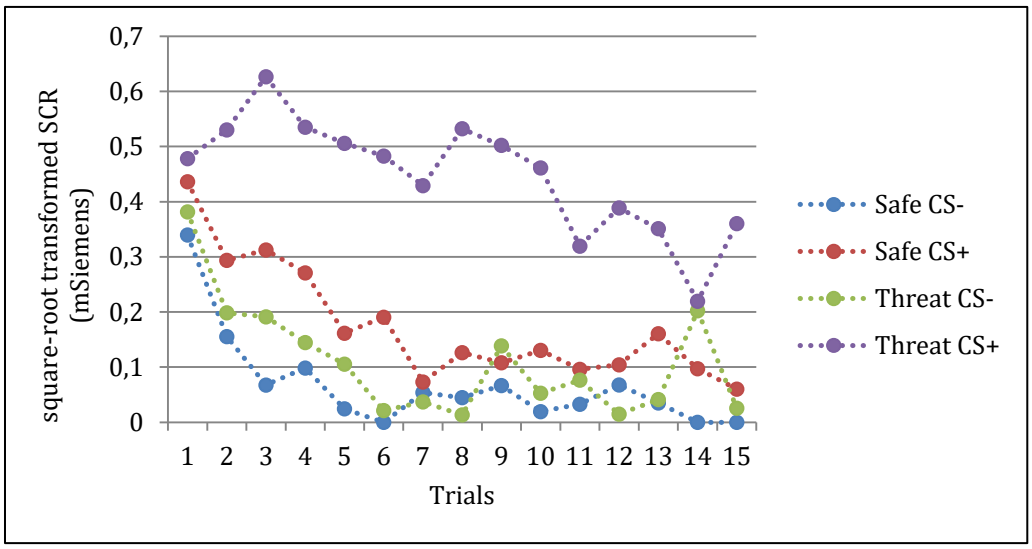


Figure 6: Mean square-root transformed skin conductance response across trials averaged for all participants and split per trial type.

Figure 6 shows the skin conductance response (SCR) across trials averaged over all participants. The responses are split into the four different trial types that were used in the fear conditioning task. We performed a 2x2x15 repeated measures ANOVA on the SCR data with 3 factors: context (safe/threat), CS (CS^+/CS^-) and trials (1-15). The main thing we were interested in was the interaction effect of context * CS [$F(1, 23) = 17.534$, $p < .001$]. When testing this interaction effect separately for both contexts and conditions, we found that the SCR response to the CS^+ was significantly higher than the CS^- for both the threat and safe context ($p < .001$), although the effect was significantly larger in the threat context with a mean difference of .233 ($p < .001$). The fact that we see significantly higher responses in the fear compared to safe and for CS^+ compared to CS^-, shows that contextual differential fear conditioning was successful. Furthermore, we can conclude a generalization effect to the safe context is occurring at the group level, because even in the safe context, participants show a significantly higher SCR response to the CS^+ than to the CS^-.

Additionally, we found a significant main effect of context [$F(1, 23) = 25.862$, $p < .001$], as well as a main effect of CS [$F(1, 23) = 38.922$, $p < .001$]. The main effect of trials was corrected with a Greenhouse-Geisser correction, because the Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated, $\chi^2(104) = 249.1$, $p < .001$, and was also

significant: [$F(14, 10) = 6.937$, $p < .001$]. In addition, the linear contrast of trials showed a significant result ($p < .001$), from which we concluded that this overall decrease over time represented the habituation of the SCR in response to the fear conditioning. No other significant interaction effects were found in the SCR data.

3.2 fMRI analysis: univariate

We were mainly interested in the interaction effect of context * CS, to analyze whether we replicate activations in the same brain regions which are commonly found in differential fear conditioning studies and furthermore to analyze whether the neural substrates giving rise to the differential conditioning effect are different for the threat versus the safe context. We found that this interaction was positively related to the left rolandic operculum spreading to the anterior insula, the left pallidum, the left dorsal anterior cingulated cortex (ACC) and the left cerebellum. After small volume correction we also observed the bilateral hippocampus, bilateral amygdala and the right parahippocampal gyrus to be positively related to this interaction. The interaction showed a negative effect in the left angular gyrus, the left inferior orbitofrontal cortex (OFC) bilateral ventro-medial prefrontal cortex (vmPFC) and the left superior and medial frontal gyrus. These results are shown in table 1.

Contrast	**Region**	**Side**	**x(mm)**	**y(mm)**	**z(mm)**	**p value**	**t value**
Context * CS: Positive effect	Rolandic operculum/insula	L	-56	2	6	.000	7.80
	Pallidum	L	-10	6	-2	.000	7.45
	Midcingulate area/dACC	L	-4	14	34	.000	7.13
	Cerebellum	L	-32	-54	-32	.005	5.73
	Amygdala	L	-28 -22	-8 -6	-12 -12	.003 * .003 *	4.01 4.05
		R	22	-4	-12	.009 *	3.72
	Hippocampus	L	-30	-12	-12	.006 *	4.28
		R	20 30	-4 -10	-12 -12	.030 * .046 *	3.75 3.59
	Parahippocampal gyrus	R	30	-14	-30	.042 *	3.69
Context * CS: Negative effect	Angular gyrus	L	-54 -42	-62 -60	32 30	.001 .026	6.10 5.28
	Inferior orbitofrontal cortex	L	-38 -48	44 40	-12 -8	.005 .030	5.75 5.25
	Medial orbitofrontal cortex/vmPFC	L	-36	52	-8	.021	5.35
		R	4 32	60 48	-6 -10	.004 .040	5.78 5.17
	Superior frontal gyrus	L	-12	40	52	.005	5.72
	Medial frontal gyrus	L	-40	16	48	.017	5.40

*Table 1: Regions of activation, peak voxel coordinates and peak voxel statistics for context*CS interaction effect. All coordinates are defined in MNI152 space. All reported statistics are FWE corrected at the voxel level, and are significant at* $p\text{-}_{FWE}$*<.05. Regions corrected with SVC are marked with *.*

To disentangle the context by CS interaction, we tested the two CSs separately for both contexts and we tested the effects of both contexts separately for both CSs. This way we could show what underlies the contextual modulation of the fear response. Results are shown in table

2. In the CS^{+}_{threat} > CS^{-}_{threat} contrast we observed differential activations for in the left rolandic operculum spreading to the insula, right dorsal ACC spreading to the left supplemental motor area, left cerebellum and right inferior temporal gyrus. After small volume correction we also observed activations in the bilateral hippocampus and bilateral amygdala. In addition, we found the CS^{-}_{threat} > CS^{+}_{threat} contrast associated to the left inferior OFC, the right vmPFC, the bilateral angular gyrus, the pons and left superior frontal gyrus, which all match the negative effects found for the original interaction (table 1). In the safe context, the CS^{+} showed a higher activation only in one peak voxel in the left inferior OFC compared to the CS^{-}. However, when looking at this contrast with a lower threshold it seems to be really close to the peak voxels related to the anterior insula, which would resemble what we showed for this differential contrast in the threat context. This would make sense, because we expected to see some activation in response to the CS+ but these activations are much weaker in the safe context. From this we concluded that differential conditioning was indeed mostly present in the threat context.

When looking at the CSs separately, we found the same regions as in the CS^{+}_{threat} > CS^{-}_{threat} contrast to be activated by the CS^{+} in the threat environment compared to the safe environment, except for the inferior temporal gyrus (CS^{+}_{threat} > CS^{+}_{safe}). Furthermore, we observed CS^{+}_{safe} > CS^{+}_{threat} contrast effects in the inferior OFC, the vmPFC, the superior and medial frontal cortex, the pons and the left angular gyrus, again matching the positive and negative effects found for the original interaction (table 1). The results of this contrast can also be viewed in figure 7. The contrasts for CS^{-} did not show any activations or deactivations. From this we could conclude that the interaction effect between context and CS was driven by the CS^{+} in the threat context.

Contrast	**Region**	**Side**	**x(mm)**	**y(mm)**	**z(mm)**	**p value**	**t value**
CS^{+}_{threat} > CS^{-}_{threat}	Rolandic operculum / insula	L	-56	2	6	.000	9.96
	Midcingulate area / dACC	R	8	12	40	.000	9.91
	Supplemental motor area / dACC	L	6	8	54	.000	9.74
	Cerebellum	L	-32 -44 -34	-54 -58 -58	-32 -32 -54	.000 .009 .008	7.60 5.58 5.60
	Inferior temporal gyrus	R	44	-6	-40	.029	5.26
	Hippocampus	L	-30 -18	-12 -6	-12 -12	.000 * .001 *	5.25 4.83
		R	20 30	-4 -10	-12 -12	.000 * .003 *	5.09 4.49
	Amygdala	L	-28 -20	-6 -6	-12 -12	.000 * .000 *	5.31 5.22
		R	22 30 18	-4 -8 4	-12 -12 -16	.000 * .004 * .030 *	5.34 4.02 3.30
CS^{+}_{safe} > CS^{-}_{safe}	Inferior orbitofrontal cortex	L	-44	20	-2	.016	5.42
CS^{+}_{threat} > CS^{+}_{safe}	Rolandic operculum / insula	L	-56	2	6	.000	9.85
	Midcingulate area / dACC	L	-2	6	38	.000	8.91

	Supplemental motor area / dACC	R	8	12	40	.000	8.72
	Cerebellum	L	-32	-54	-32	.000	6.61
	Amygdala	L	-22	0	-12	.000 *	5.56
		R	30 24	-8 2	-12 -12	.000 * .000 *	4.88 5.35
	Hippocampus	L	-30 -18	-12 -6	-12 -12	.002 * .001 *	4.66 4.68
		R	30 20	-10 -4	-12 -12	.000 * .002	5.06 4.66
$CS^-_{threat} > CS^+_{threat}$	Inferior orbitofrontal cortex	L	-40	46	-12	.000	6.27
	Medial orbitofrontal cortex / vmPFC	R	4	60	-6	.001	6.20
			34	48	-10	.000	5.80
			6	50	-16	.001	6.05
	Pons	R	-4	-18	-48	.000	6.35
	Angular gyrus	R	50	-62	44	.041	5.16
		L	-52	-64	34	.034	5.21
	Superior frontal gyrus	L	-16	38	52	.006	5.69
$CS^+_{safe} > CS^+_{threat}$	Inferior orbitofrontal cortex	L	-48 -38	40 44	-8 -12	.005 .007	5.75 5.65
	Medial orbitofrontal cortex / vmPFC	L	-36	52	-8	.004	5.82
		R	4 4 32	60 50 48	-6 -16 -10	.040 .027 .039	5.17 5.27 5.17
	Superior and medial frontal gyrus	L	-12 -42	40 18	52 48	.001 .022	6.10 5.33
		R	12	40	54	.012	5.49
	Pons	L	-4	-18	-46	.047	5.12
	Angular gyrus	L	-54	-62	32	.003	5.84

*Table 2: Differential conditioning effects. All coordinates are defined in MNI152 space. All reported statistics are FWE corrected at the voxel level, and are significant at $p\text{-}_{FWE}<.05$. Regions corrected with SVC are marked with *.*

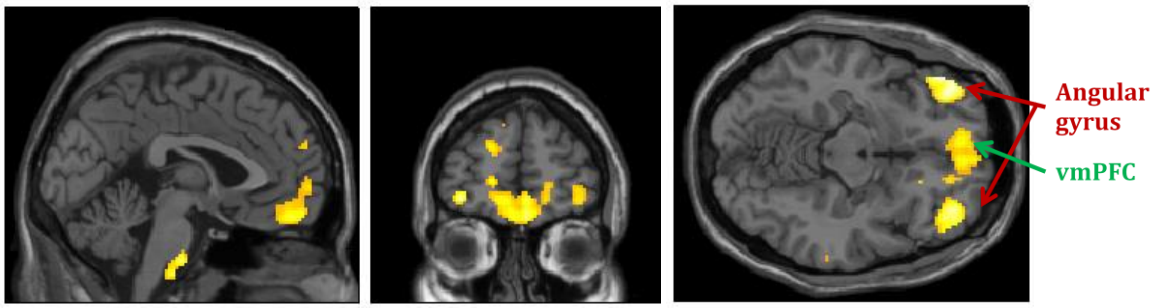


Figure 7: Visualization of the brain activations after CS+ presentation in the safe context compared to the threat context. We clearly see this is related to activations in the vmPFC and angular gyrus. For visualization purposes a threshold of p<.001(uncorrected) was used. The color indicates the t-value at that particular location, lighter meaning a higher t-value.

Finally, we were interested in whether fear generalization on the psychophysiological level could be related to fear generalization at the neural level. For this, we investigated the effect of the generalization score as calculated from the SCR data as covariate on the context by CS interaction. We expected to find that a lower generalization score was related to a higher effect of this interaction, because people who show low generalization are to a higher extent able to suppress the fear response in the safe context. Against our expectations, we did not find an effect of the generalization score on the negative or positive context by CS interaction.

3.3 fMRI analysis: MVPA

First, we tested whether the SVM classifier was able to decode the context participants were in from their BOLD activity pattern in the parahippocampal gyrus, hippocampus, amygdala and entorhinal cortex, all bilaterally. We started by performing classifications for both the context training and fear conditioning task separately. In table 3 we report results when using betas as data input for the classifier, but we did also run classifications with t-values which did not yield different results. When giving the context training task beta values as input, we found a significant classification in the right entorhinal cortex with an average accuracy of 52.62 (p = .004), whereas classifications in other regions did not reach significance (table 3, classification 1). For the fear conditioning task however, we did not find significant accuracies in any ROI (table 3, classification 2). When giving context training trials as training data and fear conditioning trials as test data, we also did not find a significant classification, meaning that representations could not be generalized from the first task in the virtual environment to the second task in any of the studied regions (table 3, classification 3).

After the previous analysis did not seem to yield significant results for the hippocampus, which was against our expectations, we divided the hippocampus into three parts on the longitudinal axis (anterior, medial and posterior) and looked whether one of the subregions showed higher classification accuracies and the effect of the context could be located within the hippocampus even on a smaller scale. However, we did not find significant accuracies in any hippocampal subregion.

	Context training				**Fear conditioning**			**Between tasks**	
	side	**accuracy**	**p-value**		**accuracy**	**p-value**		**accuracy**	**p-value**
hippocampus	L	52.05	.174		51.24	.346		50.12	.825
	R	50.52	.804		52.42	.157		50.74	.514
parahippocampal gyrus	L	48.62	.232 *		49.00	.450 *		50.31	.750
	R	51.03	.474		49.35	.585 *		49.99	.975
entorhinal cortex	L	51.14	.516		50.39	.639		50.18	.976
	R	52.62	.004 **		51.29	.495		50.37	.765
Amygdala	L	51.26	.486		51.82	.149		49.59	.748
	R	48.28	.302 *		50.01	.942 *		48.51	.349

*Table 3: Shows the results of the MVPA classifier. Mean accuracies below the mean of the permutation tests (meaning that the p-value goes in the opposite direction) are marked with *. Significant classifications are marked with **.*

With a 4x2 repeated measures ANOVA with factors area and hemisphere we found no interaction between brain area and the hemisphere when looking at the decoding results. We performed an additional one sampled t-test on the whole entorhinal cortex (i.e. the mean of left and right accuracies) in the context training task, but this did not give a significant result [t (1,23) = 1.364, p = .186]. Therefore we cannot say that the classification appeared in the entorhinal cortex as a whole and the effect seems specific for the right entorhinal cortex.

Finally, one of our main predictions was that a negative correlation would exist between the individual degree of fear generalization and the decoding accuracies of the classifier, representing the strength of contextual representations. However, we did not find any significant correlations between the individual SCR generalization scores and decoding accuracies of the classifier in any of the studied regions. We also did not find correlations between individual prediction accuracies and the scores on the STAI questionnaire for state and trait anxiety, subjective awareness of the contextual conditioning, age or gender.

3.4 Questionnaires

From the feedback questionnaire we found that 23 of the 24 participants recognized that they only received shocks for one of the images, whereas only 19 participants could correctly indicate that they only received shocks in one of the two contexts. This understanding of the paradigm did not correlate significantly with the frequency with which the participants played computer games or with the gender or age of the participants.

Discussion

With the current study, we aimed to investigate the relationship between context representations in the brain and fear generalization. We expected that inaccurate encoding of a threatening context results in overgeneralization, leading to inadequate fear responses in safe contexts. In addition to individual differences in fear generalization amongst healthy people, this could be relevant for understanding what is affected in pathological anxiety. To investigate this in healthy individuals, we studied the role of activity patterns in four MTL regions in dissociating between two buildings (contexts) within a VR, later distinguished as a threat and safe building through contextual fear conditioning. Our results demonstrated that spatial location could be decoded from fMRI activity patterns in the right entorhinal cortex, but not in other MTL regions. Furthermore, safety learning related areas such as the vmPFC and angular gyrus were activated more in response to the CS^+ in the safe context than in the threat context, and more to the CS^- than to the CS^+ in the threat context, although we were not able to relate activation levels in these areas to individual SCR generalization scores. We also did not show a correlation between individual SCR generalization scores and the decoding accuracies of the classifier in the studied MTL regions.

We first tested whether our differential contextual fear conditioning paradigm successfully modelled context as an occasion setter for the association between the CS and the US. This was indeed the case, since analysis of the SCR data showed that the response was higher for CS^+ in the treat context than in other conditions. Secondly, with univariate analysis of fMRI

data, we also confirmed this by showing that on the neural level, the interaction between context and CS in the fear conditioning task was driven by the CS^+ in the threat context. Brain regions related to conditioned fear such as the anterior insula extending to the frontal operculum, dorsal ACC spreading to the supplemental motor area, cerebellum, amygdala and hippocampus (Fullana et al., 2016), were significantly more active during our fear conditioning paradigm not only for the $CS^+_{\text{threat}} > CS^-_{\text{threat}}$ contrast but also for the $CS^+_{\text{threat}} > CS^+_{\text{safe}}$ contrast. In addition, we demonstrated that generalization of the fear response was present in our healthy study population using both SCR data and univariate contrasts. First, the SCR response to the CS^+ was significantly higher than to the CS^- in the safe context, showing that a weak fear response was present even in the absence of conditioning. Second, weak neural effects were found for the CS^+ compared to CS^- in the safe context around the anterior insula, suggesting that brain regions related to conditioned fear might activate in safe environments as well, but to a much weaker extent.

More specifically, our univariate results are also consistent with previous contextual fear conditioning studies using virtual reality. Importantly, these studies showed that contextual stimuli activate the hippocampus, amygdala, but also the ACC, mPFC, inferior and orbital frontal cortices, insula and ventral putamen. This indicates the involvement of a broad network in encoding contextual representations in humans (Alvarez et al., 2008; Lang et al., 2009). The involvement of not only the hippocampus but also various cortical areas in contextual encoding is consistent with evidence from animal studies (e.g. Quinn, Ma, Tinsley, Koch & Fanselow, 2008), suggesting the same mechanism for fear conditioning across species (Maren et al., 2013). Andreatta et al. (2015) expand on this by disentangling activations to initial and sustained fear response. Initial fear responses, when entering the threat context, activated the motor cortex which they implicated in eliciting a direct fear response, the dlPFC and dmPFC and the OFC, which were thought to create awareness of the contingency and appraisal of the threat. When subjects remained in the threat context they showed sustained anxiety responses in the amygdala, which according to Andreatta et al. (2015) resulted in the conditioned response, and hippocampus, which they implicate in generating a spatial map of the threat event. Although we were not able to separate these initial from sustained responses in the current study, Andreatta et al. 2015 provide a clear explanation as to how areas in this broad contextual fear learning network interact and in which order they are activated.

We observed a negative context by CS interaction in amongst others the vmPFC (see table 1). More specifically, the vmPFC was activated more in response to the CS^+ in the safe context than in the threat context, and more to the CS^- than to the CS^+ in the threat context (see table 2 and figure 7). This is consistent with previous studies describing the role of the vmPFC in

mediating safety learning and context dependent fear extinction (Kalish et al., 2006; Milad et al., 2007). Recently, Åhs et al. (2015) used structural equation modelling to compare functional circuitries of prefrontal pathways between fear renewal and extinction recall contexts in their fMRI VR differential contextual fear conditioning paradigm. They showed that the dmPFC fully mediated the effect of the hippocampus on right amygdala activity during fear renewal, whereas the vmPFC partially mediated the effect of the hippocampus on right amygdala activity during extinction recall. This indicates that the context can determine the functional relation of the hippocampus with prefrontal pathways, which can then determine whether fear expression is activated (through dmPFC) or inhibited (through vmPFC). This suggests that actions among the hippocampus, the amygdala and the vmPFC protect us from fear renewal, which might be of relevance to understand the nature of pathological anxiety. Since a change of context in our fear conditioning paradigm shows similar vmPFC effects as during extinction in Åhs et al. (2015), the loss of the fear response during both extinction and change of context seems to be enabled by a more general inhibitory mechanism which signals safety and thereby prevents the expression of conditioned fear. The vmPFC, possibly in interaction with the hippocampus, seems to support this mechanism that enables the modulation of fear expression. It thus seems that fear extinction and context dependent fear expression are two expressions of the same process, particularly supported by the vmPFC. Our results support these findings about the role of the vmPFC in context dependent fear expression, as the vmPFC was significantly more active in the safe context, suggesting that the vmPFC inhibits fear expression by sending a safety signal. A design with context as occasion setter might thus engage a mechanism similar to an extinction design.

We now return to our findings related to contextual representations as investigated with a multivariate approach. Involvement of the entorhinal cortex in representing contextual information was expected because of the close connection of this area to the hippocampal formation (Schultz & Engelhardt, 2014), and since the entorhinal cortex accounts for two third of the hippocampal input (Burwell & Agster, 2008; Shah, Jhawar & Goel, 2012). Also, the presence of grid cells suggests a unique role of the medial entorhinal cortex in contextual representations in humans (Brown, Hasselmo & Stern, 2014; Döller, Barry & Burgess, 2010). Nonetheless, our study in combination with the ones of Hassabis et al. (2009) and Rodriguez et al. (2011) is not in line with the idea that neural decoding of voxel patterns reflects place-cell activity. In our study, spatial location could only be decoded from entorhinal activity patterns and not from the hippocampus or parahippocampal gyrus as shown by Hassabis et al. (2009) and Rodriguez et al. (2011). Since the decoding of spatial representations does not appear to be unique for the hippocampus, and place cells have only been found in this region (O'Keefe & Burgess, 1996), it rather suggests another undetermined type of spatial representation. A possibility is that the whole MTL might play a role in encoding of spatial contextual information,

and it appears to depend a lot on the study design which areas within the MTL you detect to discriminate between your experimental contexts. Furthermore, the BOLD signal is only an indirect measure of neuronal activity and many factors can influence their relationship. To find meaningful BOLD differences with MVPA in fMRI, the large-scale spatial distribution of place cells must show sufficient anisotropy across voxels (Logothetis, 2008). It is questionable whether such a large-scale hippocampal population code exists, since the low spatial resolution of the BOLD signal measures populations of ten thousands of neurons and does not allow for drawing inferences about cell populations. Local field potentials also represent the summed representation of ten thousands of neurons, and are therefore robustly correlated with BOLD responses (Logothetis, 2008). In favour of the anisotropic nature of the hippocampal population code, Agarwal et al. (2014) showed that a rat's position on a maze could be decoded from spatially distributed local field potentials in the hippocampus. How this relates to place cell populations remains not entirely clear. However, more studies are needed to find the general mechanism of context encoding in the human brain and specifically in MTL regions.

We only found significant classification in the entorhinal cortex during context training, and not during the fear conditioning task or when we trained the classifier on the context training and tested it on the fear conditioning data. On the one hand we expected that differences in arousal between the two contexts during the fear conditioning task would elicit different activity patterns for the two contexts, even possibly resulting in more dissociable representations in MTL regions. For example, the awareness of aversive properties of the CS is hippocampus dependent (Phelps & LeDoux, 2005), and is therefore likely to elicit different activity patterns in a threat versus a safe environment. On the other hand, human fear conditioning studies as well as animal studies showed that during a typical contextual fear conditioning procedure, we first encode context representations during early context exploration when no foot shocks (US) are applied yet, as reflected by hippocampal activation. In later phases of acquisition, we learn to associate that representation with the US and fear expression can be shown (Lang et al., 2013; Maren et al., 2013). It would in that case make sense that we were only able to find clear dissociable context representations during the context training task and that during fear conditioning other processes relating to learning correct context-US associations may be stronger than the patterns that were created earlier to dissociate the two contexts. Therefore, during the fear conditioning task and when comparing activity patterns between tasks, we might not have found differences between the two buildings because the representations had changed due to learned associations. However, with the current study we cannot explain why we could not decode these different learned associations for the two buildings from the amygdala and hippocampus.

We only found a significant classification in the right entorhinal cortex and not in the left, suggesting that there is a certain influence of the hemisphere in context encoding. Functional differentiation of hippocampus has been studied a lot: anterior hippocampus was suggested to play a role in hot processing, and posterior in cold processing such as spatial navigation (Strange et al., 2014). In addition, Robinson et al. 2016 found hemisphere-specific specialization in the hippocampal formation, showing differences mostly in functional connectivity with other MTL regions, inferior frontal gyrus and the basal ganglia. The differences between left and right in posterior connectivity were more remarkable than in the anterior hippocampus. This hemisphere-specific functional specialization might be also present in brain areas connected to the hippocampus such as the entorhinal cortex. This might explain why we were able to decode spatial context only from the right entorhinal cortex. However, for this region this functional specialization has never been investigated yet.

With the current analyses, we were not able to show a correlation between decoding accuracy en fear generalization. We predicted that a negative correlation would exist between the individual degree of fear generalization and the decoding accuracies of the classifier, representing the strength of contextual representations. This would imply that people who showed more generalized fear responses show less strong context representations which make them worse in dissociating between safe and threat environments. The fact that we were not able to show a correlation could be partly due to a relatively small study population for showing correlations, and to the fact that there was a low variability in STAI scores in our study population. Future studies using an extended paradigm and a study population which shows more variation in STAI scores, might be able to show this association.

One potential caveat of the current design is that we only compare brain activity patterns between two contexts. It would have been useful to include more than two contexts, because in our design we are only able to detect differences in decoding accuracies given two buildings and therefore given the difference without being able to assess them independent of each other. However, the reason why we limited trials to two buildings was to be able to perform neural decoding of place information on a sufficient number of trials per building. Since navigating through the environment took a large part of the duration of the scan session, it was not possible to include more trials per building without making the task too long. Another limitation is that the number of trials per building is not necessarily equal in every session, due to individual differences in navigational speed. For this reason, some trials had to be excluded in our MVPA analysis to make sure the classifier was trained on an equal number of A and B trials. In a next version of this paradigm, the number of trials must be fixed. In addition, decoding would become more accurate if more runs were included, because then the classifier would have a larger training data set. However, this would also have resulted in a task that was too long. It could be

considered to divide the experiment over two days, but this would result in many more confounds relating to for example the time of the day.

Studying the role of contextual representations in fear learning could have important implications for the optimization of psychotherapy in anxiety disorders. It would therefore be interesting to use our paradigm in patients suffering from anxiety disorders or PTSD to see whether contextual representations are less strong compared to healthy subjects. If we can confirm that pathological anxiety is related to overgeneralization of fear responses because of inadequate use of contextual information by MTL regions (Kalisch et al., 2006; Haaker et al., 2015), various forms of therapy could be improved by integrating this information. Firstly, the aim of exposure therapy is to reduce the person's fearful reaction to the stimulus (Foa & Kozak, 1986). Patients are exposed to the situations or objects they fear, and after repeatedly experiencing safety they generate fear inhibitory safety memories. However, fear often returns because safety memory retrieval fails. Because exposure during therapy sessions is context-specific, too little generalization of extinction learning occurs. To improve the adequate use of contextual information, contextual representations in extinction learning should also be taken into account. Cognitive behavioural therapy as well as exposure therapy should focus more on incorporating methods to recognize safety signals and suppress fear responses in safe contexts (Haaker et al., 2015), but they should also find ways to make extinction learning more generalized to allow adequate safety learning. In addition, pharmacotherapy could be more effective if it focused more on fear suppression in safe environments. This was also suggested by Haaker et al. (2013), who showed that administration of the dopamine precursor L-dopa improved dopamine availability and increased neural activity in the vmPFC during consolidation, which improved extinction memory expression and thereby prevented the return of fear.

In conclusion, our results suggest that contextual information can be decoded from right entorhinal cortex, although extended paradigms might reveal general contextual encoding in all MTL regions. Furthermore, we showed that safety learning was related to stronger activations in the vmPFC and angular gyrus, although we were not able to show that these activations were higher for individuals showing a lower degree of fear generalization. We also could not support our hypothesis that low decoding accuracies were related to a high degree of fear generalization.

References

Agarwal, G., Stevenson, I. H., Berényi, A., Mizuseki, K., Buzsáki, G., & Sommer, F. T. (2014). Spatially distributed local fields in the hippocampus encode rat position. *Science*, *344*(6184), 626-630.

Åhs, F., Kragel, P. A., Zielinski, D. J., Brady, R., & LaBar, K. S. (2015). Medial prefrontal pathways for the contextual regulation of extinguished fear in humans. *NeuroImage*, *122*, 262-271.

Alme, C. B., Miao, C., Jezek, K., Treves, A., Moser, E. I., & Moser, M. B. (2014). Place cells in the hippocampus: Eleven maps for eleven rooms. *Proceedings of the National Academy of Sciences*, 111(52), 18428-18435.

Alvarez, R. P., Biggs, A., Chen, G., Pine, D. S., & Grillon, C. (2008). Contextual fear conditioning in humans: cortical hippocampal and amygdala contributions. *The Journal of Neuroscience*, 28(24), 6211-6219.

Andreatta, M., Glotzbach-Schoon, E., Mühlberger, A., Schulz, S. M., Wiemer, J., & Pauli, P. (2015). Initial and sustained brain responses to contextual conditioned anxiety in humans. *Cortex*, *63*, 352-363.

Bach, D. R., Friston, K. J., & Dolan, R. J. (2010). Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations.*International Journal of Psychophysiology*, *76*(1), 52-55.

Baeuchl, C., Meyer, P., Hoppstädter, M., Diener, C., & Flor, H. (2015). Contextual fear conditioning in humans using feature identical contexts. *Neurobiology of learning and memory*, 121, 1 11.

Barrett, L. F., & Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. *Psychological Science*, 21(4), 595-599.

op de Beeck, H. P. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943-1948.

Bell, A. H., Hadj-Bouziane, F., Frihauf, J. B., Tootell, R. B., & Ungerleider, L. G. (2009). Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging. *Journal of neurophysiology*, *101*(2), 688-700.

Blaimer, M., Choli, M., Jakob, P. M., Griswold, M. A., & Breuer, F. A. (2013). Multiband phase constrained parallel MRI. *Magnetic Resonance in Medicine, 69*(4), 974–980.

Brown, T. I., Carr, V. A., LaRocque, K. F., Favila, S.E., Gordon, A. M., Bowles, B., Bailenson, J. N., Wagner, A. D. (2016). Prospective representation of navigational goals in the human hippocampus. *Science* 352 (6291), 1323-1326.

Brown, T. I., Hasselmo, M. E., & Stern, C. E. (2014). A High-resolution study of hippocampal and medial temporal lobe correlates of spatial context and prospective overlapping route memory. *Hippocampus*, *24*(7), 819-839.

Burwell, R. D., & Agster, K. L. (2008). Anatomy of the hippocampus and the declarative memory system. *Memory systems*, *3*, 47-66.

van Buuren, M., Gladwin, T. E., Zandbelt, B. B., van den Heuvel, M., Ramsey, N. F., Kahn, R. S., & Vink, M. (2009). Cardiorespiratory effects on default-mode network activity as measured with fMRI. *Human Brain Mapping*, *30*(9), 3031–3042.

Chadwick, M. J., Bonnici, H. M., & Maguire, E. A. (2012). Decoding information in the human hippocampus: a user's guide. *Neuropsychologia*, *50*(13), 3107-3121.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27. (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm)

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature, 463*(7281), 657–661.

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., ... & Baas, J. M. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and anxiety*,32(4), 239-253.

Dunsmoor, J. E., Prince, S. E., Murty, V. P., Kragel, P. A., & LaBar, K. S. (2011). Neurobehavioral mechanisms of human fear generalization. *NeuroImage*, *55*(4), 1878-1888.

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, *99*(1), 20.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... & Montillo, A. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341-355.

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: an updated and extended meta analysis of fMRI studies.*Molecular Psychiatry*, *21*(4), 500-508.

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., & Uludağ, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *NeuroImage*, *132*, 32-42.

Garfinkel, S. N., Abelson, J. L., King, A. P., Sripada, R. K., Wang, X., Gaines, L. M., & Liberzon, I. (2014). Impaired contextual modulation of memories in PTSD: an fMRI and psychophysiological study of extinction retention and fear renewal. *The Journal of Neuroscience*, *34*(40), 13435-13443.

Glover, G. H., Li, T. Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic resonance in medicine: official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *44*(1), 162–167.

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology*, *91*(3), 186-193.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., ... Haase, A. (2002), Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine, 47*(6), 1202–1210.

Gullone, E. (2000). The development of normal fear: A century of research. *Clinical psychology review*, 20(4), 429-451.

Haaker, J., Gaburro, S., Sah, A., Gartmann, N., Lonsdorf, T. B., Meier, K., ... & Kalisch, R. (2013). Single dose of L-dopa makes extinction memories context-independent and prevents the return of fear. *Proceedings of the National Academy of Sciences*, *110*(26), 2428-2436.

Haaker, J., Lonsdorf, T. B., Schümann, D., Menz, M., Brassen, S., Bunzeck, N., ... & Kalisch, R. (2015). Deficient inhibitory processing in trait anxiety: Evidence from context-dependent fear learning, extinction recall and renewal.*Biological psychology*, *111*, 65-72.

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801-806.

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P. D., & Maguire, E. A. (2009). Decoding neuronal ensembles in the human hippocampus. *Current Biology*, 19(7), 546-554.

Kadosh, K. C., Haddad, A. D., Heathcote, L. C., Murphy, R. A., Pine, D. S., & Lau, J. Y. (2015). High trait anxiety during adolescence interferes with discriminatory context learning. *Neurobiology of learning and memory*, 123, 50-57.

Kalisch, R., Korenfeld, E., Stephan, K. E., Weiskopf, N., Seymour, B., & Dolan, R. J. (2006). Context dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *The Journal of neuroscience*, *26*(37), 9503-9511.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381(6581), 425-428.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science*, 256(5057), 675-677.

Kriegeskorte, N. (2011). Pattern-information analysis: from stimulus decoding to computational model testing. *Neuroimage*, *56*(2), 411-421.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126-1141.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7(1), 54-64.

Lang, S., Kroll, A., Lipinski, S. J., Wessa, M., Ridder, S., Christmann, C., ... & Flor, H. (2009). Context conditioning and extinction in humans: differential contribution of the hippocampus, amygdala and prefrontal cortex. *European Journal of Neuroscience*, *29*(4), 823-832.

Li, M., Lu, S., & Zhong, N. (2016). The Parahippocampal Cortex Mediates Contextual Associative Memory: Evidence from an fMRI Study. *BioMed research international*, *2016*.

Lissek, S., Bradford, D. E., Alvarez, R. P., Burton, P., Espensen-Sturges, T., Reynolds, R. C., & Grillon, C. (2014). Neural substrates of classically conditioned fear-generalization in humans: a parametric fMRI study. *Social cognitive and affective neuroscience*, *9*(8), 1134 1142.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI.*Nature*, *453*(7197), 869-878.

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel pattern analysis for fMRI data: A review. *Computational and mathematical methods in medicine*, *2012*.

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience*, 14(6), 417 428.

Maren, S., & Quirk, G. J. (2004). Neuronal signalling of fear memory. *Nature Reviews Neuroscience*, 5(11), 844-852.

Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., & Rauch, S. L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry, 62*(5), 446-454. doi:10.1016/j.biopsych.2006.10.011

Misaki, M., Luh, W. M., & Bandettini, P. A. (2013). The effect of spatial smoothing on fMRI decoding of columnar-level organization with linear support vector machine. *Journal of neuroscience methods*, *212*(2), 355-361.

Moita, M. A., Rosis, S., Zhou, Y., LeDoux, J. E., & Blair, H. T. (2004). Putting fear in its place: remapping of hippocampal place cells during fear conditioning. *The Journal of neuroscience,* 24(31), 7015-7023.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1-25.

Parsons, R. G., & Ressler, K. J. (2013). Implications of memory modulation for post-traumatic stress and fear disorders. *Nature neuroscience*, 16(2), 146-153.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*(1), S199-S209.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron*, *48*(2), 175-187.

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of cognitive neuroscience*.

Quinn, J. J., Ma, Q. D., Tinsley, M. R., Koch, C., & Fanselow, M. S. (2008). Inverse temporal contributions of the dorsal hippocampus and medial prefrontal cortex to the expression of long-term fear memories. *Learning & Memory*,*15*(5), 368-372.

Reinders, A. A. T. S., Den Boer, J. A., & Büchel, C. (2005). The robustness of perception. *European Journal of Neuroscience*, 22(2), 524-530.

Robinson, J. L., Salibi, N., & Deshpande, G. (2016). Functional connectivity of the left and right hippocampi: Evidence for functional lateralization along the long-axis using meta analytic approaches and ultra-high field functional neuroimaging. *NeuroImage*, *135*, 64 78.

Rodriguez, P. F. (2010). Neural decoding of goal locations in spatial navigation in humans with fMRI. *Human brain mapping*, *31*(3), 391-397.

Sasaki, T., Leutgeb, S., & Leutgeb, J. K. (2015). Spatial and memory circuits in the medial entorhinal cortex. *Current opinion in neurobiology*, *32*, 16-23.

Schultz, C., & Engelhardt, M. (2014). Anatomy of the hippocampal formation. In *The Hippocampus in Clinical Neuroscience* (Vol. 34, pp. 6-17). Karger Publishers.

Shah, A., Jhawar, S. S., & Goel, A. (2012). Analysis of the anatomy of the Papez circuit and adjoining limbic system by fiber dissection techniques. *Journal of Clinical Neuroscience*, *19*(2), 289-298.

Shin, L. M., & Liberzon, I. (2010). The neurocircuitry of fear, stress, and anxiety disorders. *Neuropsychopharmacology*, *35*(1), 169-191.

Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. *Nature Reviews Neuroscience*, *15*(10), 655-669.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., ... & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273-289.