



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, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Chapter 3

Neural pattern similarity predicts long-term fear memory

Renée M. Visser
H. Steven Scholte
Tinka Beemsterboer
Merel Kindt

This chapter is based on the article that is published as:

Visser, R.M., Scholte, H.S., Beemsterboer, T. & Kindt, M. (2013). Neural pattern similarity predicts long-term fear memory. *Nature Neuroscience*, 16(40), 388-390.

Abstract

Although certain changes in the brain may reflect fear learning, no marker is yet available that indicates whether an aversive experience will develop into fear memory. Here we unveil the moment-to-moment dynamics of human fear learning, by applying MVPA to single-trial BOLD-MRI data. We demonstrate that the long-term behavioral expression of fear memory can be predicted from neural patterns at the time of learning.

Introduction

Despite everything we know about the processes that can weaken or strengthen memory for fearful events, our understanding of how these events are represented, processed and eventually engraved into the neural architecture of the brain remains remarkably poor. An inherent restriction of memory research, including fear conditioning, is that we can only infer fear memory from the degree to which behavior during learning (e.g., freezing in rats, physiological responding in humans) overlaps with behavior at a later retention test. Although the expression of fear during learning is certainly related to long-term memory, much of what we learn does eventually not transform into long-term memory. Currently, no signature exists that at the time of learning permits a read-out of subsequent consolidation (Dudai, 2012).

The application of Multi-voxel pattern analysis (MVPA), a technique for analyzing distributed patterns of BOLD-MRI data (Haxby et al., 2001), may be especially promising for fear memory research, given that there is naturally more information in patterns than in mean activation (Haxby et al., 2001; Norman et al., 2006). Using this technique, it has been shown that the neural representations of stimuli change as a function of fear conditioning (Bach, Weiskopf, & Dolan, 2011; Li et al., 2008; Visser, Scholte, & Kindt, 2011). While it is tempting to interpret these changes as a marker for fear memory, they were not examined in relation to a later, independent memory test (Bach et al., 2011; Li et al., 2008; Visser et al., 2011). The question is whether patterns of activation also inform about upcoming consolidation processes, indicating that, at least in fear memory research, MVPA is more than a cross-validation technique for assessing transient learning-dependent changes. Although neural pattern similarity during encoding can successfully predict relatively short-term (1-6 h) declarative memory performance (Xue et al., 2010), this effect could be explained by enhanced conscious processing of the subsequently remembered items (Schurger et al., 2010; Xue et al., 2010). Central to fear memory, however, is not the conscious processing of threat and the factual recollection of events, but the automatic emotional expression of fear.

Here, we tested whether the formation of long-term procedural fear memory can be assessed with neural pattern similarity during learning. To this end, we combined fMRI with a behavioral measure of associative fear learning (i.e., pupil dilation responses) during differential fear conditioning and a memory-test (Supplementary Figure 1a). We repeatedly presented six stimuli, derived from two distinct categories (faces and houses), as the to-be-conditioned stimuli (CS). One face and one house (CSs^{+neg}) co-terminated with a mild electric stimulus (unconditioned stimulus, UCS) in half of the trials (Supplementary Figure 1b). Another face and another house (CSs^{+neut}) co-terminated with a neutral sound (neutral associative stimulus, AS) in half of the trials, to control for non-aversive associative learning. The third face and house (CSs^{-}) were never reinforced. For the

fMRI-analyses, we only analyzed CS^{+}_{neg} and CS^{+}_{neut} trials in which no shock or sound was presented, and corresponding CS^{-} trials (together referred to as ‘target’ trials; Supplementary Figure 1c).

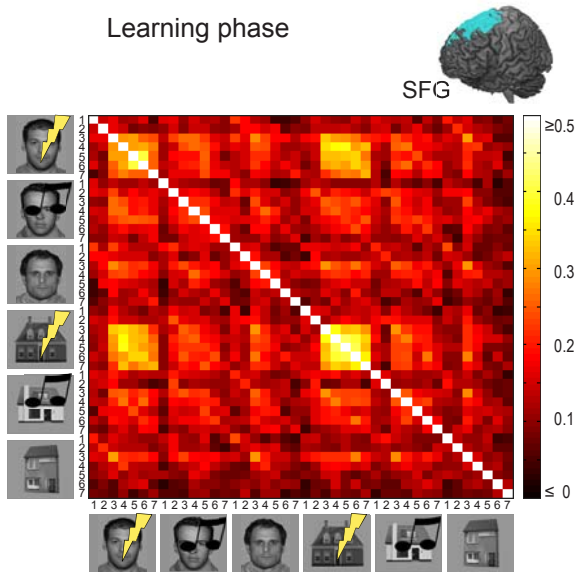


Figure 1 A 42 x 42 correlation matrix, containing within-stimulus and between-stimulus correlations of BOLD-MRI patterns in the superior frontal gyrus (SFG) during the learning phase ($n = 38$). The off-diagonal represents correlations between two consecutive target trials (within-stimulus correlations). Between-stimulus correlations include correlations between stimuli from the same original category (faces and houses) and between stimuli that share a particular outcome (shock vs. sound vs. no consequence).

We employed trial-by-trial similarity analysis (Kriegeskorte et al., 2008; Visser et al., 2011) in order to 1) examine the formation, activation and extinction of fear associations and, importantly, 2) assess whether pattern similarity predicts the long-term expression of fear memory. Hereto, we modelled trials as separate events. For each of the 38 participants, we created a vector containing the spatial pattern of BOLD-MRI signal related to one stimulus presentation, for six anatomically defined regions of interest (ROIs): the anterior cingulate cortex (ACC), the insula, the amygdala, the hippocampus, the ventromedial prefrontal cortex (vmPFC) and the superior frontal gyrus (SFG). Next, we correlated each vector with all other vectors related to the other stimulus presentations, resulting in a similarity matrix of 42-by-42 target trials (Figure 1). Correlations of interest included the correlations between two consecutive presentations of the same stimulus (within-stimulus) and between pairs of different stimuli (between-stimulus), including stimuli that belonged to the same category (original categories: faces and houses) and stimuli that shared a specific outcome (categories based on learned association: shock vs. sound vs. no consequence) (Supplementary Figure 2). Between-stimulus pattern similarity is characterized by a focus on a common outcome rather than a focus on single predictors and putatively reflects a type of higher-order fear learning (Visser et al., 2011).

Methods

Participants

Fifty-four students of the University of Amsterdam participated in the initial fear learning session. Forty-seven participants returned for the subsequent memory phase. Data were discarded from analyses if participants did not return for the second session, if fMRI-data were confounded by substantial head motion ($> 2\text{mm}$ in any direction, $n = 4$) or by excessive sleepiness, judged on the basis of eye-tracker data combined with self-report ($n = 3$). Additionally, two participants were excluded from the analysis, because they had not learned the contingencies (see *Experimental design*). For the fMRI analyses, the final sample included 38 participants (15 male, 36 right-handed) between 18 and 33 years of age (mean 23.7 ± 3.8 s.d.). The time between the two sessions ranged from 13 to 43 days (mean 22.18 ± 6.4 s.d.). From three participants, part of the eye-tracker data was missing (*not* the retention trials) and from six participants part of the AS-expectancies was missing. Therefore, when data from the three experimental phases are compared directly, eye-tracker data are reported for 35 participants and US- and AS-expectancies are reported for 32 participants. Participants earned course credit or €22,- per session. All participants gave their written informed consent before participation and were naive to the purpose of the experiment. Procedures were executed in compliance with relevant laws and institutional guidelines, and were approved by the local ethics committee (2011-CP-1565).

Experimental design

The experiment consisted of two phases: a learning phase and a memory phase. The latter consisted of a baseline and an extinction phase (Supplementary Figure 1a). For the learning phase, a differential fear conditioning paradigm was used, with delay conditioning and partial reinforcement (Supplementary Figure 1b). Three faces (NimStim set, Tottenham et al., 2009) and three houses (collected from the internet and separated from their background) were repeatedly presented for 4.5 seconds and served as the to-be-conditioned stimuli (CS; see main text). Stimuli were converted to grayscale and presented on a grey background.

The electrical stimulation served as aversive unconditioned stimulus (UCS) and was delivered twice for 2 ms, with a delay of 300 ms, applied by a Digitimer D57A through MRI-compatible carbon electrodes attached to the right shin-bone. Prior to the experiment, the intensity of the electric stimulus was individually adapted to be aversive but not painful (intensity range 8-71 mA, mean 34.6 ± 15.6 s.d.). For the neutral associative stimulus (AS) a sound (500 ms) was selected that is standard implemented in Windows ('Chimes'). Subjective assessments (Supplementary Table 3) confirmed the intended valence of both UCS and AS.

Inter-stimulus intervals were fixed and long enough (19.5 seconds) to limit intrinsic noise correlations. The onset of each trial was triggered by the start of the acquisition of a BOLD-MRI volume. The order of stimulus presentation was fixed (counterbalanced across participants) and consisted of a repeating sequence of six target trials, with filler trials of the same stimuli in between (Supplementary Figure 1c). In total, the learning phase consisted of 78 trials: 42 target trials (7 per stimulus type) and 36 filler trials (6 per stimulus type), including all CS^{+neg} trials that co-terminated with a shock and all CS^{+neut} trials that co-terminated with a sound. The baseline phase (without electrodes) consisted of 24 trials: 18 target trials (3 per stimulus type) and 6 filler trials (1 per stimulus type). Finally, the extinction phase (with electrodes) consisted of 78 trials: 42 target trials (7 per stimulus type) and 36 filler trials (6 per stimulus type). For fMRI-analyses, we constrained our analyses to target trials (Visser et al., 2011).

Pupil dilation

Pupil dilation responses and eye-movements were recorded continuously throughout MRI-scanning, using a remote non-ferromagnetic infrared Eyelink-1000 Long Range Mount eye-tracker (SR Research). Before task onset, a nine-point calibration procedure was performed.

Data were sampled at 250 Hz. The baseline pupil diameter was the average value -during the 500 ms prior to each CS onset. The pupil response to the CS was calculated as the peak change from baseline in a window from 0 to 4 s after picture onset, discarding any data samples that were obscured by eye blinks. Trials that suffered substantial signal loss (*not* related to sleepiness), affecting either the baseline or more than 2 seconds within 4 seconds after stimulus onset were eliminated (0 -19 % per participant, median = 0.6 %) and replaced using the linear trend at point. Next, data were Fisher-transformed for each of the three experimental phases separately.

Image acquisition

Scanning was performed on a 3T Philips Achieva TX MRI scanner using a 32-channel head-coil. Functional data were acquired using gradient echo, echo planar imaging (TR = 2000 ms; TE = 27.63 ms; FA = 76.1°; 39 sagittal slices with interleaved acquisition; 3.0 × 3.0 × 3.3 mm voxel size; 64 × 64 matrix; 192 × 192 × 141.24 FoV) covering the whole brain. The learning phase consisted of 946 dynamics, the baseline of 295 and the extinction phase again of 946 dynamics. Foam pads minimized head motion. A high-resolution 3D T1-weighted image (TR = 8.124 ms, TE = 3.72 ms, FA = 8°; 1 × 1 × 1 mm voxel size; 256 × 256 × 160 FoV) was additionally collected for anatomical visualization.

Preprocessing

FEAT (fMRI Expert Analysis Tool) version 4.1.6, part of FSL (Oxford Centre for Functional MRI of the Brain [FMRIB] Software Library [www.fmrib.ox.ac.uk/fsl]) was used to analyze the (f)MRI data. Pre-processing steps included slice-time correction, motion correction, spatial smoothing (Beeck, 2010) using a 5 mm full-width-at-half-maximum Gaussian kernel, high-pass filtering in the temporal domain ($\sigma = 50$) and prewhitening (Woolrich et al., 2001). Structural images were co-registered to the functional images and transformed to MNI standard space (Montreal Neurological Institute) using FLIRT (FMRIB's Linear Image Registration Tool, FSL). The resulting normalization parameters were applied to the functional images.

Region of interest selection

Regions of interest (ROI) were selected based on their role in fear learning and (extinction) memory (LaBar & Cabeza, 2006; Sehlmeier et al., 2009) and included the anterior cingulate cortex (ACC), the insula, amygdala, hippocampus and ventromedial prefrontal cortex (vmPFC). We additionally included the superior frontal gyrus (SFG), to illustrate that robust learning-dependent changes - as revealed by similarity analysis - can also be observed outside the salience network (Visser et al., 2011). ROIs were obtained from the Harvard-Oxford cortical and subcortical structural atlases (Harvard Center for Morphometric Analysis). Additional analyses were performed with functional and equal-sized ROIs (Supplementary Table 4, 5, 8 and 12).

Univariate analysis

To visualize trial-by-trial learning dependent changes in average activation, we modelled all trials as separate events and extracted for each event the average response amplitude per ROI using Matlab (version 7.11; MathWorks). For additional whole brain analyses, see Supplementary Table 9.

Trial-by-trial similarity

For the trial-by-trial representational similarity analysis, trials were also modelled as separate events. The resulting single-trial data were further analyzed in Matlab, by calculating pair-wise Pearson correlations between event-related spatial patterns of activation (Z-values per voxel). This resulted in a similarity matrix containing correlations among trials, for each participant, for each ROI. From this matrix (Figure 1), three different types of correlations were selected (see main text; Supplementary Figure 2). The strength of these correlations was used as a metric of similarity. Correlations were then Fisher-transformed for each experimental phase separately. Figures, however, display raw data, as this facilitates interpretation of the results.

Statistical analyses

Fisher-transformed pupil dilation responses, within-stimulus correlations and correlations for original-related stimuli were averaged over face- and house stimuli. This was done to reduce the number of comparisons and because we were not interested in the difference between faces and houses with regard to the experimental manipulation (but see Figure 1).

Statistical comparisons of the learned associations were performed by within-subjects Analysis of Variance (ANOVA), using Statistical Package for the Social Sciences (SPSS, version 17; SPSS Inc.). Differential fear learning was assessed by the interaction of all trials \times stimulus, but was only tested when there was also a significant main effect of stimulus type. For the baseline phase, no main effects of stimulus type were expected. Reactivated fear memory was assessed by a main effect of stimulus type during the extinction phase, and if significant, extinction was assessed by a significant interaction of all extinction trials \times stimulus type.

The predictive value of the different measures was examined by assigning participants to a 'Retention'-group and a 'No retention'-group, according to their pupil dilation responses on the first three extinction trials (see main text). To compare groups, we first calculated difference scores for each trial from the learning phase: $2(CS^{+}_{neg}) - CS^{+}_{neut} - CS^{-}$ for single-trial activation data and pupil data; $3(CS^{+}_{neg}) - CS^{+}_{neut} - CS^{-}$ - *original category* for between-stimulus correlations. Then, we subtracted the average difference over the first target trials (when participants did not know the contingencies; Supplementary Figure 1c) from the average difference over the later target trials (when contingencies could have been learned), yielding a 'conditioning'-index that expressed the relative increase of CS^{+}_{neg} responses over the course of learning. These 'conditioning'-indices were then compared between the 'Retention' and 'No retention'-group and also used for the continuous assessment of fear memory.

Predictions were tested while correcting for multiple comparisons (6 ROIs) by limiting the false discovery rate (FDR; Benjamini & Hochberg, 1995). In case that the assumption of sphericity was violated a Greenhouse-Geisser correction was applied. All p -values are reported two-sided.

Results & Discussion

Consistent with previous work (Visser et al., 2011), the application of trial-by-trial similarity analysis revealed clear learning curves that indexed the formation of associative fear (within-stimulus: trial (6) \times stimulus (3), between-stimulus: trial (7) \times stimulus (4); all $ps \leq 0.004$, η_p^2 ranging from 0.06-0.22; Figure 2a; Supplementary Figure 3a; Supplementary Table 1). Two to six weeks later (mean 22.18 days \pm 6.4 s.d.) participants returned for the memory phase. Without electrodes attached, exposure to previously learned CS^{+}_{neg} stimuli did not elicit differential pattern similarity. With electrodes attached, differential pattern similarity reappeared and - as the aversive outcome was no

longer delivered - eventually extinguished (Supplementary Table 2). To compare our approach with standard univariate analysis, we examined single-trial activation, averaged over all voxels in a ROI. This yielded similar learning and extinction curves as obtained with similarity analysis in ACC and insula (Figure 2b), although interaction effects did not reach FDR-corrected significance (all $p \geq 0.033$; Supplementary Table 1).

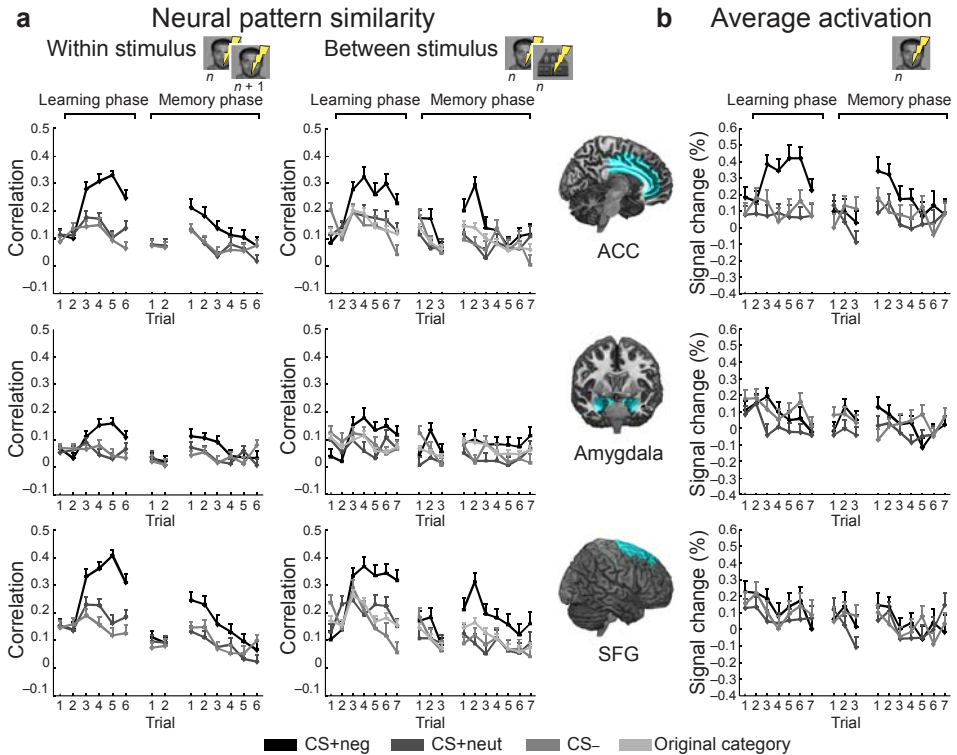


Figure 2 Neural pattern similarity versus average activation during different experimental phases. Within-stimulus pattern similarity (averaged over faces and houses) and between-stimulus pattern similarity (a) and average single-trial activation (averaged over faces and houses) (b) in the anterior cingulate cortex (ACC), amygdala and superior frontal gyrus (SFG) ($n = 38$) show the formation, re-expression and extinction of fear-associations. The high sensitivity of similarity analysis compared to univariate analysis reveals the involvement of brain areas outside the traditional fear-circuit (e.g., SFG). Error bars represent SEM.

In parallel with the results from the similarity analysis, successful fear conditioning was also evident from increased pupil dilation in response to CS^+_{neg} stimuli compared to CS^+_{neut} and CS^- stimuli over the course of conditioning (trial (13) \times stimulus (3); $F_{24, 816} = 9.15$, $p < 0.0005$, $\eta p^2 = 0.21$). Without electrodes attached, the presentation of previously learned CS^+_{neg} stimuli did not elicit enhanced pupil responses. With electrodes attached, differential pupil responses were present again (main

effect of stimulus, $p < 0.0005$, $\eta_P^2 = 0.43$), and eventually extinguished (trial (13) \times stimulus (3), $F_{24, 816} = 2.49$, $p < 0.0005$, $\eta_P^2 = 0.07$; Supplementary Figure 4). These findings confirm the validity of the conditioned pupil dilation response as a behavioral measure of fear (Reinhard, Lachnit, & König, 2006).

After the first three trials of the extinction phase, differential pupil responses rapidly diminished (2nd versus 3rd extinction trial, $F_{1, 37} = 0.14$, $p = 0.713$; 3rd versus 4th extinction trial, $F_{1, 37} = 8.10$, $p = 0.007$), indicating that extinction learning started after the third trial. To predict the long-term expression of fear memory, we classified the 38 participants according to their behavioral expression of fear memory (Figure 3a), averaged over these first three ‘retention’-trials before extinction learning became apparent. We assigned participants to the ‘Retention’-group ($n = 22$) if they showed a stronger pupil response to the CS^{+neg} -face stimulus as well as to the CS^{+neg} -house stimulus, compared to any of the four control stimuli (CSs^{+neut} and CSs^{-}). The remaining participants were assigned to the ‘No retention’-group ($n = 16$). Groups did not differ on subjective and procedural variables (Supplementary Table 3 and Supplementary Figure 5).

We assessed the predictive value of fear learning curves derived from neural pattern similarity, by comparing groups on mean ‘conditioning’-indices (*Methods*), which expressed the relative increase of the CS^{+neg} responses over the course of learning. For within-stimulus pattern similarity no substantial differences were found (Supplementary Table 4; Supplementary Figure 6). For between-stimulus pattern similarity, we found more stimulus differentiation during initial fear learning in the ‘Retention’-group compared to the ‘No retention’-group, in the ACC ($F_{1, 36} = 6.53$, $p = 0.015$, $\eta_P^2 = 0.15$), insula ($F_{1, 36} = 5.10$, $p = 0.030$, $\eta_P^2 = 0.12$), amygdala ($F_{1, 36} = 4.96$, $p = 0.032$, $\eta_P^2 = 0.12$), hippocampus ($F_{1, 36} = 5.16$, $p = 0.029$, $\eta_P^2 = 0.13$) and vmPFC ($F_{1, 36} = 8.47$, $p = 0.006$, $\eta_P^2 = 0.19$; Figure 3b), with a trend observed in the SFG ($p = 0.079$) (Supplementary Figure 7; Supplementary Table 5). Notably, the ‘Retention’- and ‘No retention’ group did not differ in pupil responses at the time of fear learning ($F_{1, 33} = 2.19$; $p = 0.148$; Figure 3a).

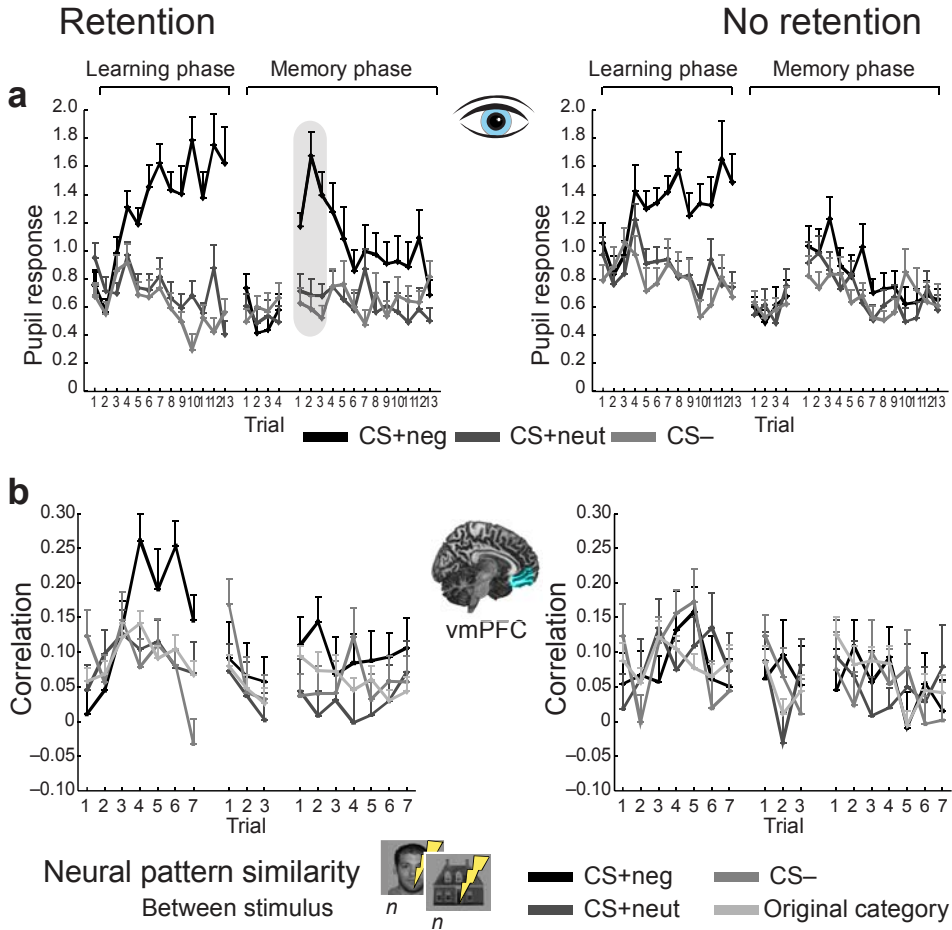


Figure 3 Pupil dilation and neural pattern similarity split by long-term procedural fear memory. Pupil dilation responses ($n = 35$) show differential learning and extinction of fear (a). Based on differential pupil responses during the first three trials from the extinction phase, participants were assigned to the 'Retention'-group ($n = 19$ for pupil data, $n = 22$ for MRI-data) or the 'No retention'-group ($n = 16$) (a, b). While differential pupil responses during learning did not predict subsequent pupil responses, differential pattern similarity in several brain regions, including the ventromedial prefrontal cortex (vmPFC; b), did. Error bars represent SEM.

Follow-up tests for both groups separately (Supplementary Table 6) revealed strong fear learning – as indexed by differential pattern similarity - in the 'Retention'-group in all areas ($ps \leq 0.018$; Figure 3b; Supplementary Figure 7). The 'No retention'-group showed a different pattern. For this group, no fear learning was observed in the amygdala, hippocampus (Supplementary Figure 7) and vmPFC (Figure 3b), but some fear learning was observed in the ACC, the insula and the SFG (Supplementary Figure 7). A continuous assessment of the predictive value of pattern similarity (Supplementary

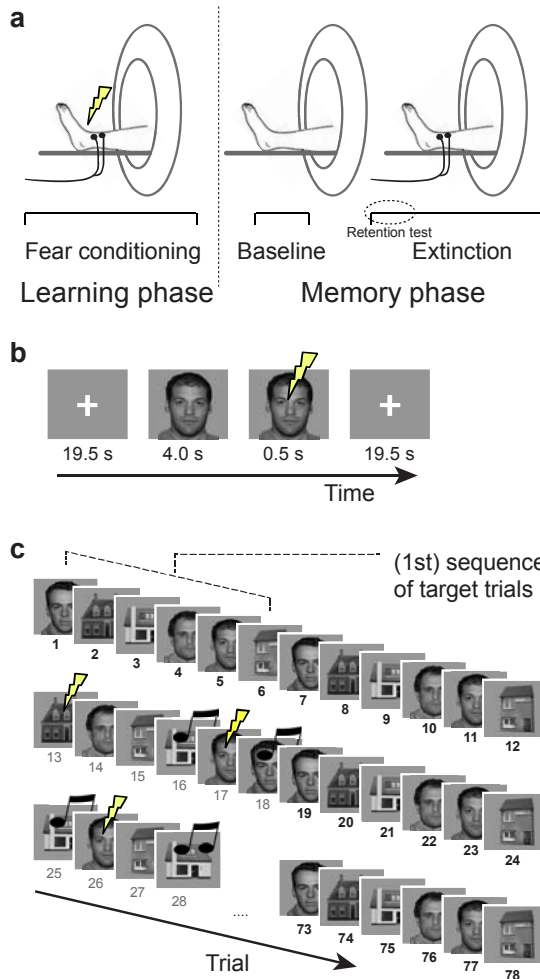
Material) revealed a linear relationship between differential between-stimulus pattern similarity in the ACC, insula, vmPFC and SFG during fear learning and subsequent differential pupil dilation during the retention trials (Supplementary Table 7). Again, differential pupil responses during fear learning did not predict differential pupil responses at test ($p = 0.464$).

Crucially, we found that average activation did not predict the later expression of fear memory in any of the ROIs (Supplementary Figure 8; Supplementary Table 7 and 8), or anywhere else in the brain (*Supplementary Material*; Supplementary Table 9). We repeated the ROI analyses while only considering voxels that were responsive to the task (Haxby et al., 2001) (*Supplementary Material*). This enhanced learning effects for both pattern similarity and mean activation (Supplementary Table 10), but again revealed that the prediction of procedural fear memory was restricted to between-stimulus pattern similarity (Supplementary Table 4, 5 and 8). Although within-stimulus pattern similarity did not clearly predict procedural fear memory, we replicated other findings on declarative memory (Xue et al., 2010) by showing that in the hippocampus within-stimulus pattern similarity was higher for subsequently remembered stimuli compared to subsequently forgotten stimuli ($F_{1,15} = 5.28$, $p = 0.036$, $\eta_p^2 = 0.26$, uncorrected; Supplementary Table 11).

In sum, the refinement of an individual stimulus representation (within-stimulus) seems different from the formation of a higher-order fear association (between-stimulus). The cortical distribution of between-stimulus and within-stimulus pattern similarity further substantiates this idea (Visser et al., 2011). Whereas these two types of neural pattern similarity both reflect fear- and extinction learning, they apparently contain different information about future consolidation processes, given that only between-stimulus neural pattern similarity predicts the later behavioral expression of fear memory. Higher-order fear learning expressed as changes in neural patterns may offer a promising signature to examine the determinants of fear memory.

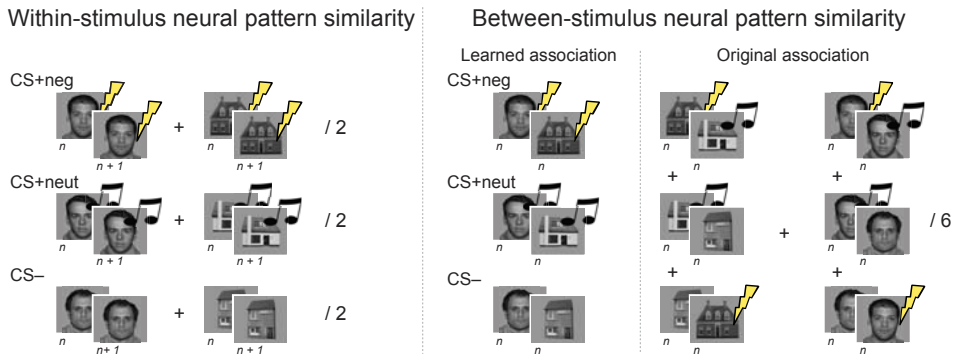
Supplementary Material – chapter 3

Supplementary Figures

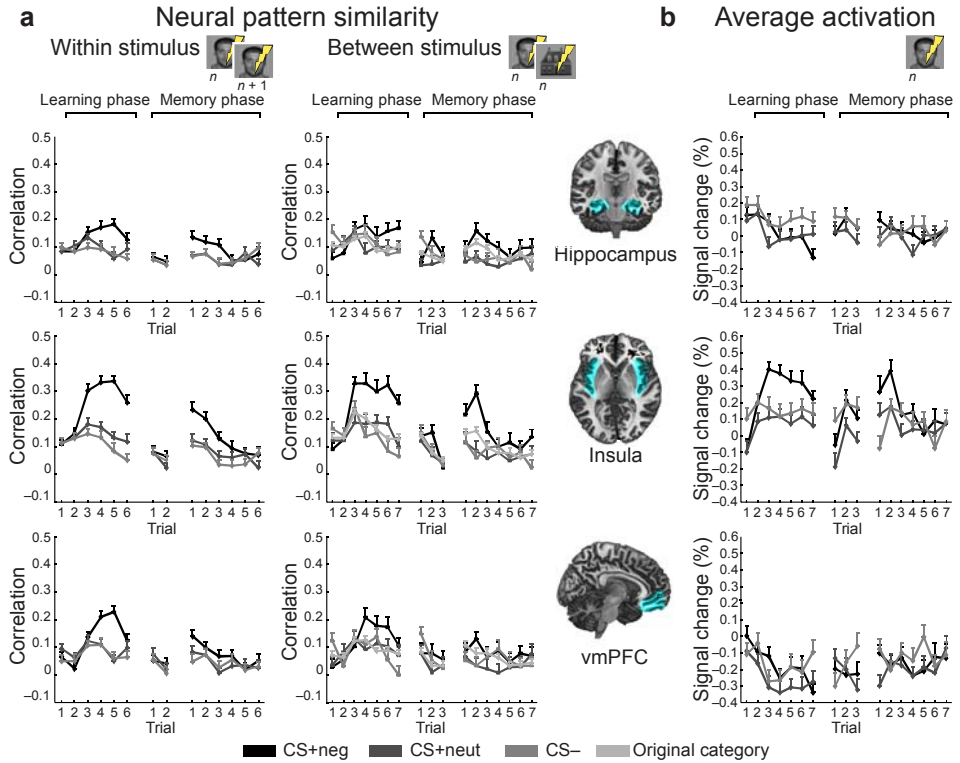


Supplementary Figure 1 Experimental design. Schematic drawing of different experimental phases, consisting of a learning phase and a memory phase, all of which took place during functional MRI-scanning (a). During the learning phase, fear associations were acquired through differential fear conditioning. Participants were told that two out of six stimuli could be followed by the shock, that two other stimuli could be followed by a sound, while the remaining two would never be followed by a shock or a sound, and that they had to learn these contingencies. During the memory phase, participants were told that no shocks or sounds were delivered during the first run, and that this run served as a baseline. In between the two runs, the

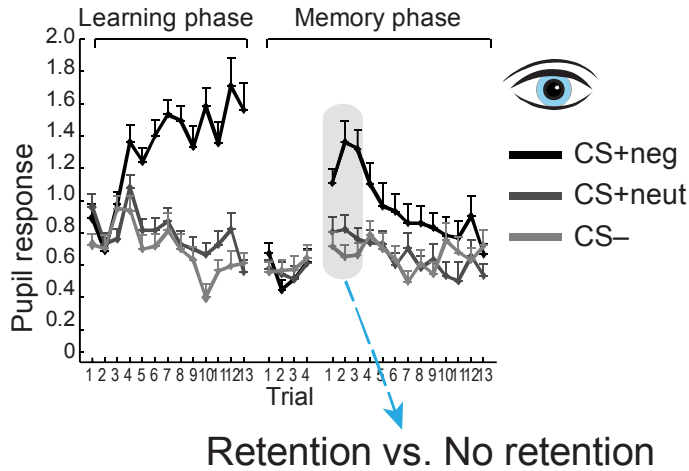
experimenter entered the scanner room and attached the electrodes. During this run, participants were instructed to remember what they had learned during the previous session, to prevent participants from erroneously expecting a different contingency scheme. The shock intensity was explicitly set at the individual level as determined in the previous session, but no actual shocks were delivered. Six stimuli, which were backward-projected onto a screen that was viewed through a mirror attached to the head-coil, were repeatedly presented for 4.5 seconds. During fear conditioning (b), two of these co-terminated with a shock on 46% of the trials (CS^+_{neg}), two co-terminated with a sound on 46% of the trials (CS^+_{neut}), and two were never reinforced (CS^-). Inter-trial intervals were fixed. The paradigm consisted of repeating sequences of target trials (bold), presenting the six different stimuli in a fixed order such that the time between two consecutive target trials was equal over the six conditions (c). In between, the semi-random presentation of filler trials ensured the unpredictability of stimuli. Administration of a UCS or AS only occurred on filler trials, which were discarded from the fMRI analyses to be certain that shock- or sound-related activity did not confound CS-related activity. The first paired CS^+ trials were presented between the second and third sequence of target trials. Images are not to scale.



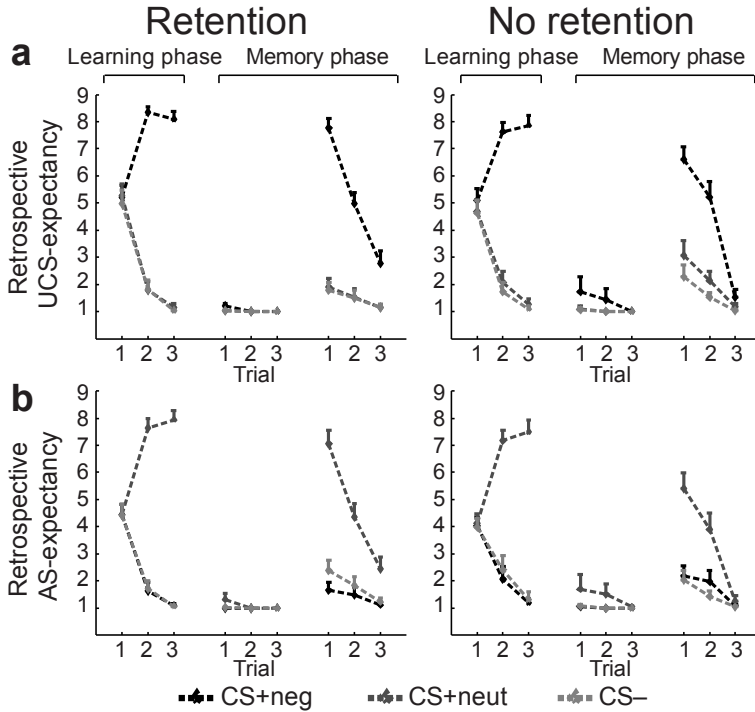
Supplementary Figure 2 Assessment of neural pattern similarity. In order to assess neural pattern similarity, correlations were calculated between patterns evoked by consecutive trials of the same stimulus (within-stimulus), trials of stimuli that share (non)reinforcement (categories based on learned association [i.e., a similar outcome]: CS^+_{neg} -face with CS^+_{neg} -house, CS^+_{neut} -face with CS^+_{neut} -house and CS^- -face with CS^- -house) and trials of stimuli that belong to the same category (original associations: faces and houses). Note that the number of between-stimulus correlations is equal to the number of target trials, whereas the number of within-stimulus correlations is equal to the number of target trials minus one. Results reported for original category correlations and within-stimulus correlations are averaged over faces and houses (see *Methods*). Images are not to scale.



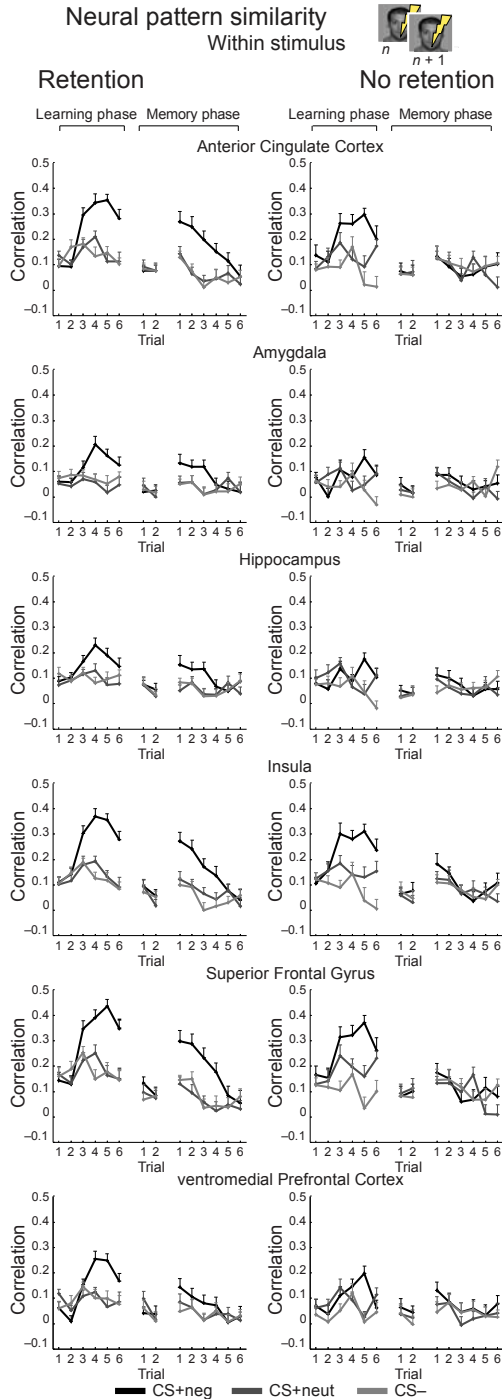
Supplementary Figure 3 Neural pattern similarity versus average activation during different experimental phases. Graphs refer to neural pattern similarity, within-stimulus and between-stimulus, (a) and average single-trial activation (b) in the hippocampus, insula and ventromedial prefrontal cortex (vmPFC) ($n = 38$). This illustrates how the processing of stimuli based on affective significance is reflected throughout the cortex, during different experimental phases. Differential correlations emerge during fear conditioning. After a few weeks, no differential correlations are visible at baseline, when electrodes are not attached. However, the presence of a threat causes reactivation of the fear associations, which extinguish when participants learn that the aversive consequence is no longer delivered. Error bars represent SEM.



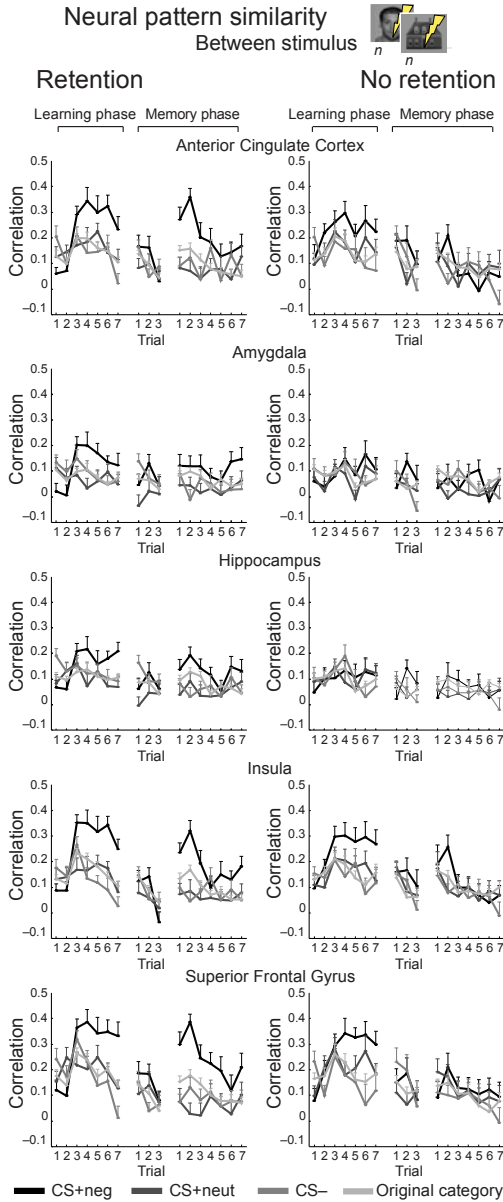
Supplementary Figure 4 Pupil dilation responses. Pupil dilation responses to the CS ($n = 35$), calculated as the peak change from baseline in a window from 0 to 4 s after picture onset, show differential learning and extinction of fear associations. The relatively high temporal resolution of the pupillary response allowed for the inclusion of filler trials (i.e., CSs⁺ that co-terminated with a UCS or AS), as responses triggered by UCS or AS could not confound the CS-related signal. Each condition thus contained 13 trials from the learning and extinction phase and 4 trials from the baseline phase. Based on the behavioral expression of fear memory during the first three trials from the extinction phase, participants were assigned to the 'Retention'-group ($n = 19$ for pupil data, $n = 22$ for MRI-data) or the 'No retention'-group ($n = 16$). Next, we compared data from the learning phase to predict subsequent procedural fear memory (group membership). Our main dependent measure for fear learning (between-stimulus pattern similarity) is based on the correlation between separate stimuli, not the average of individual stimulus presentations (as with within-stimulus pattern similarity). To match this as closely as possible, we used a classification criterion that ensured that for anyone in the retention group both CS⁺_{neg} stimuli, not just one, were encoded as predictors of threat (the UCS). The 'Retention'-group therefore consisted of individuals that had a stronger pupil dilation response (average over first three retention-trials) to both CS⁺_{neg} stimuli separately, compared to any of the other four stimuli. Note that in the 'No retention'-group individuals could potentially have retained their fear for only one CS⁺_{neg} stimulus, which would be expressed as an enhanced pupil response to either the CS⁺_{neg}-face or the CS⁺_{neg}-house. Hence, after averaging pupil data over faces and houses (see *Methods*), pupil dilation responses in the CS⁺_{neg} condition still appear to be slightly increased in the 'No retention'-group as well (Figure 3a). Error bars represent SEM.




Supplementary Figure 5 UCS- and AS-expectancy scores. Mean retrospective UCS- and AS-expectancy scores ($n = 32$, see *Participants*), collected after the learning phase and the memory phase, displayed separately for the 'Retention'-group ($n = 17$) and the 'No retention'-group ($n = 15$). Participants were asked to identify which houses and which faces were followed by shocks and by sounds, and to indicate retrospectively their expectation of the UCS or AS on a continuous rating scale consisting of 9 points labeled from "certainly no electric stimulus/ sound" (1) through "uncertain" (5) to "certainly an electric stimulus/ sound" (9), for three time points during the scan. Mean retrospective UCS-expectancies revealed significant acquisition of fear associations (interaction effect of time point (3) \times stimulus type (3); $F_{2,17, 67.37} = 146.01$, $p < 0.0005$, $\eta_p^2 = 0.83$) and reactivation and extinction of fear associations (main effect of stimulus type and interaction effect of time point (3) \times stimulus type (3) during the extinction phase; $F_{1,31, 40.70} = 98.10$, $p < 0.0005$, $\eta_p^2 = 0.76$ and $F_{2,43, 75.32} = 34.63$, $p < 0.0005$, $\eta_p^2 = 0.53$ respectively) (a). Similarly, mean retrospective AS-expectancies revealed significant acquisition ($F_{2,39, 73.94} = 56.01$, $p < 0.0005$, $\eta_p^2 = 0.64$) and reactivation and extinction ($F_{1,40, 43.26} = 54.48$, $p < 0.0005$, $\eta_p^2 = 0.64$ and $F_{2,17, 67.17} = 20.13$, $p < 0.0005$, $\eta_p^2 = 0.39$) of neutral associations (b). Long-term expression of fear memory was not predicted by UCS-expectancies at the time of encoding (difference last time-point minus difference first time point, between-subject ANOVA; $F_{1, 30} = 2.71$, $p = 0.11$). Error bars represent SEM.

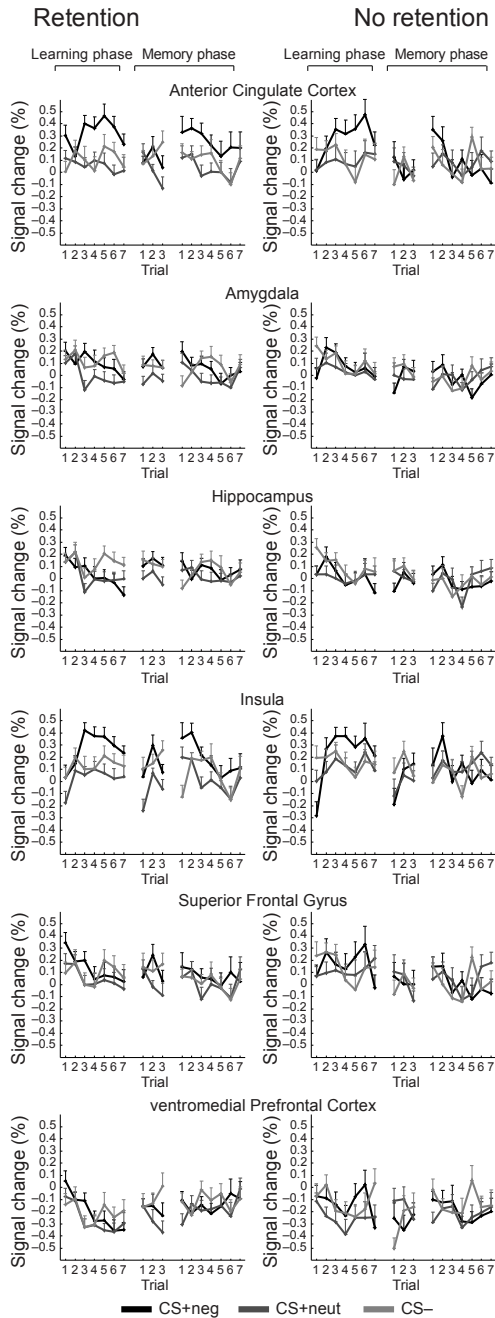


Supplementary Figure 6 Neural pattern similarity split by group. Graphs present within-stimulus neural pattern similarity in 6 ROIs for the 'Retention'-group ($n = 22$) and 'No retention'-group ($n = 16$) separately. Differential correlations at the time of encoding seem slightly enhanced in the 'Retention'-group compared to the 'No retention'-group, although the difference between groups with regard to learning effects is only significant (uncorrected) in the vmPFC. Interestingly, in all areas the behavioral expression of fear memory was paralleled by pattern similarity during the memory phase, as indicated by the absence of differential correlations during the memory phase in the 'No retention'-group. Error bars represent SEM.



Supplementary Figure 7 Neural pattern similarity split by group. Graphs present between-stimulus neural pattern similarity in 5 ROIs for the 'Retention'-group ($n = 22$) and 'No retention'-group ($n = 16$) separately. Differential correlations at the time of encoding seem enhanced in the 'Retention'-group, and predictive of subsequent memory expression, although learning effects in the 'No retention'-group reached significance, or trend significance in the anterior cingulate cortex (ACC), insula and superior frontal gyrus (SFG) as well. Interestingly, in all areas the behavioral expression of fear memory was paralleled by pattern similarity during the memory phase, as indicated by the absence of differential correlations during the memory phase in the 'No retention'-group. Error bars represent SEM.

Average activation 
n



Supplementary Figure 8 Average activation split by group. Graphs present single-trial activation in 6 ROIs for the 'Retention'-group ($n = 22$) and 'No retention'-group ($n = 16$) separately. In none of these ROIs does differential activation predict the long-term behavioral expression of fear memory. Error bars represent

Supplementary Tables

Supplementary Table 1 fMRI data for the learning phase

Region	Within-stimulus				Between-stimulus				Mean activation			
	Main effect (3)		Interaction (3 x 6)		Main effect (4)		Interaction (4 x 7)		Main effect (3)		Interaction (3 x 7)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
ACC	37.69	<0.0005	9.31	<0.0005	16.66	<0.0005	3.61	<0.0005	20.54	<0.0005	1.49	0.161
Amygdala	10.12	<0.0005	3.32	<0.0005	3.50	0.032	2.62	0.004	5.37	0.007^a	0.93	0.512
Hippocampus	10.18	<0.0005	3.12	0.004	3.24	0.034	2.35	0.009	6.24	0.003^a	1.01	0.435
Insula	56.66	<0.0005	10.36	<0.0005	25.37	<0.0005	5.61	<0.0005	13.77	<0.0005	2.16	0.033
SFG	33.23	<0.0005	9.26	<0.0005	16.62	<0.0005	6.36	<0.0005	1.54	0.221	NT	NT
vmPFC	11.41	<0.0005	5.40	<0.0005	3.47	0.034	2.33	0.009	4.63	0.019^a	1.31	0.234

All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for acquisition effects. ^a effect not caused by significantly higher values for CS^{neg} stimuli

Supplementary Table 2 fMRI data for the extinction phase

Region	Within-stimulus				Between-stimulus				Mean activation			
	Main effect (3)		Interaction (3 x 6)		Main effect (4)		Interaction (4 x 7)		Main effect (3)		Interaction (3 x 7)	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
ACC	13.23	<0.0005	1.04	0.408	7.49	0.001	2.16	<i>0.018</i>	4.35	<i>0.016</i>	0.81	0.593
Amygdala	4.73	0.012	2.36	0.010	6.28	0.002	0.66	0.775	0.79	0.458	NT	NT
Hippocampus	5.20	0.008	2.12	0.022	7.44	<0.0005	0.69	0.737	0.28	0.754	NT	NT
Insula	17.28	<0.0005	2.15	0.020	13.66	<0.0005	1.86	<i>0.049</i>	3.23	<i>0.045</i>	1.59	0.132
SFG	11.55	<0.0005	1.63	0.095	14.52	<0.0005	0.87	0.569	0.18	0.834	NT	NT
vmPFC	4.00	0.022	0.92	0.514	3.05	0.032	0.77	0.675	1.73	0.185	NT	NT

All significant values ($p < 0.05$) are in italics; those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for extinction effects.

Supplementary Table 3 Participant- and UCS- and AS-characteristics

	Retention (<i>n</i> = 22)		No retention (<i>n</i> = 16)		Main effect (2)	
	Mean (\pm s.d.)		Mean (\pm s.d.)		<i>F</i>	<i>p</i>
ASI	8.9 (\pm 5.3)		10.3 (\pm 5.9)		0.60	0.445
STAI-S	31.3 (\pm 6.3)		33.9 (\pm 9.2)		1.04	0.316
STAI-T	34.2 (\pm 8.6)		36.2 (\pm 8.8)		0.49	0.487
UCS intensity (mA)	36.4 (\pm 16.5)		32.1 (\pm 14.5)		0.68	0.414
UCS evaluation	3.4 (\pm 1.4)		3.5 (\pm 1.3)		0.01	0.934
AS evaluation	8.9 (\pm 0.4)		8.8 (\pm 1.0)		0.24	0.629
# of days between scans	22.3 (\pm 7.5)		22.0 (\pm 4.8)		0.02	0.883
	# of participants (%)		# of participants (%)		χ^2	<i>p</i>
Correct declarative memory for CS ⁺ _{neg} stimuli	17 (77.3)		10 (62.5)		0.983	0.321

Summary of participant- and UCS- and AS-characteristics (*n* = 38, between-subjects ANOVA and chi-square), split by group. Prior to the experiment, the Anxiety Sensitivity Index (ASI; Peterson & Reiss, 1993) was used to assess one's tendency to respond fearfully to anxiety-related symptoms. In addition, state and trait anxiety were assessed with the State and Trait Anxiety inventory (STAI-S and STAI-T; Spielberger, 1983). Furthermore, participants evaluated the unconditioned stimulus (UCS) and neutral associative stimulus (AS) on a 9-point scale (1 = "very unpleasant", 9 = "not unpleasant at all"). No significant differences were found on any of these measures. Although the number of participants that did not correctly identify both CS⁺_{neg} stimuli was slightly higher in the 'No retention'-group than in the 'Retention'- group, this difference was not significant, nor was the average retention-index larger for individuals with a correct declarative memory (*p* = 0.234), suggesting that declarative fear memory and the emotional expression of fear memory are at least to some extent independent.

Supplementary Table 4 Within-stimulus pattern similarity during learning compared between groups

Region	Within-stimulus							
	Anatomical ROI				Functional ROI			
	Mean conditioning-index		Main effect (2)		Mean conditioning-index		Main effect (2)	
	(\pm s.d.)				(\pm s.d.)			
Retention	No retention	<i>F</i>	<i>p</i>	Retention	No retention	<i>F</i>	<i>p</i>	
(<i>n</i> = 22)	(<i>n</i> = 16)			(<i>n</i> = 22)	(<i>n</i> = 16)			
ACC	2.33 (\pm 2.05)	1.51 (\pm 2.06)	1.49	0.230	2.21 (\pm 1.73)	1.43 (\pm 1.95)	1.71	0.199
Amygdala	1.34 (\pm 1.73)	0.96 (\pm 1.25)	0.56	0.459	1.36 (\pm 1.78)	0.80 (\pm 1.23)	1.20	0.280
Hippocampus	1.12 (\pm 1.61)	1.19 (\pm 1.51)	0.02	0.879	1.16 (\pm 1.93)	1.00 (\pm 1.45)	0.08	0.778
Insula	2.24 (\pm 1.50)	2.16 (\pm 1.95)	0.02	0.883	2.08 (\pm 1.40)	1.92 (\pm 1.79)	0.09	0.761
SFG	2.21 (\pm 1.71)	1.43 (\pm 1.81)	1.85	0.183	2.32 (\pm 1.67)	1.35 (\pm 1.71)	3.07	0.088
vmPFC	1.78 (\pm 1.54)	0.60 (\pm 1.93)	4.37	<i>0.044</i>	1.81 (\pm 1.90)	0.55 (\pm 1.77)	4.36	<i>0.044</i>

Summary of statistics of within-stimulus neural pattern similarity for the learning phase (*n* = 38, between-subjects ANOVA), compared between groups, for atlas-based ROIs (left). Within each ROI 'conditioning'-indices, which expressed the relative increase of the CS+neg responses over the course of learning, were compared between groups. It is possible that MVPA suffers less from functionally unresponsive voxels than an analysis in which activity is averaged across all voxels. We therefore repeated the analysis for each of the six anatomical ROIs, but only considering those voxels whose responses explained variance in our model (F-test on target trials, voxels thresholded at *Z* > 1.7; hereafter referred to as functional ROIs; right). All significant values (*p* < 0.05) are in italics; none of the values survives FDR-correction.

Supplementary Table 5 Between-stimulus pattern similarity during learning compared between groups

Region	Between-stimulus							
	Anatomical ROI				Functional ROI			
	Mean conditioning-index (± s.d.)		Main effect (2)		Mean conditioning-index (± s.d.)		Main effect (2)	
	Retention (n = 22)	No retention (n = 16)	F	p	Retention (n = 22)	No retention (n = 16)	F	p
ACC	3.63 (±2.92)	1.24 (±2.75)	6.53	0.015	3.43 (±2.70)	1.48 (±2.56)	5.06	<i>0.031</i>
Amygdala	3.29 (±2.39)	1.53 (±2.41)	4.96	0.032	2.58 (±2.46)	1.10 (±1.74)	4.26	<i>0.046</i>
Hippocampus	2.99 (±2.38)	1.11 (±2.69)	5.16	0.029	2.72 (±2.65)	0.83 (±2.55)	4.86	<i>0.034</i>
Insula	4.10 (±2.16)	2.48 (±2.21)	5.10	0.030	3.87 (±2.55)	2.83 (±2.25)	1.72	0.198
SFG	3.74 (±2.18)	2.43 (±2.28)	3.26	0.079	3.75 (±2.39)	2.62 (±2.03)	2.34	0.135
vmPFC	2.62 (±2.94)	-0.18 (±2.89)	8.47	0.006	2.55 (±2.79)	0.87 (±2.94)	3.21	0.082

Summary of statistics of between-stimulus neural pattern similarity for the learning phase ($n = 38$, between-subjects ANOVA), compared between groups. The assessment of the predictive value of between-stimulus neural pattern similarity was done for atlas-based ROIs (left), and for ROIs based on individual functional ROIs (right). Within each ROI 'conditioning'-indices, which expressed the relative increase of the CS_{neg} responses over the course of learning, were compared between groups. All significant values ($p < 0.05$) are in italics; those that reach FDR-corrected significance are in bold.

Supplementary Table 6 Between-stimulus pattern similarity during learning split by groups

Region	Retention (n = 22)				No retention (n = 16)			
	Main effect		Interaction		Main effect		Interaction	
	(4)		(4 x 7)		(4)		(4 x 7)	
	F	p	F	p	F	p	F	p
ACC	9.19	<0.0005	3.13	0.003	7.03	0.003	1.48	0.098
Amygdala	3.62	0.031	2.57	0.009	0.43	0.661	NT	NT
Hippocampus	4.12	0.010	2.37	0.018	0.50	0.682	NT	NT
Insula	14.86	<0.0005	4.34	<0.0005	9.97	<0.0005	1.88	0.018
SFG	9.85	<0.0005	4.42	<0.0005	6.91	0.003	3.08	<0.0005
vmPFC	5.23	0.008	2.59	0.012	0.02	0.996	NT	NT

Summary of statistics of between-stimulus neural pattern similarity for the learning phase ($n = 38$, between-subjects ANOVA), split by subsequent behavioral expression of fear memory, in six anatomical ROIs. All significant values ($p < 0.05$) are in italics; those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for acquisition effects.

Supplementary Table 7 Continuous assessment of procedural fear memory

Region	Within-stimulus		Between-stimulus		Mean activation	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
ACC	0.259	0.117	0.447	0.005	-0.013	0.939
Amygdala	0.111	0.507	0.194	0.243	0.056	0.736
Hippocampus	0.063	0.706	0.210	0.206	0.090	0.592
Insula	0.038	0.821	0.469	0.003	0.036	0.832
SFG	0.301	0.066	0.542	<0.0005	-0.044	0.792
vmPFC	0.329	<i>0.043</i>	0.363	0.025	-0.163	0.332

Correlations between conditioning-indices and differential pupil dilation responses at the retention test ($n = 38$), in six anatomical ROIs. Aside from the split-group analysis, the prediction of the behavioral expression of fear was also continuously assessed, by calculating the average difference between the CS^{+neg} stimuli and the other stimuli (thus the relative response to the CS^{+neg} stimuli) in pupil dilation responses. This is substantially different from our split-group analysis, as in the continuous assessment a strong pupil response to either one CS^{+neg} stimulus (compared to the other stimuli) could drive the difference score up. Some participants showed a large difference score on the retention trials, while not meeting the requirements for the 'Retention'-group (i.e., in the case that not both CS^{+neg} stimuli elicited stronger responses than the other stimuli). Nevertheless, results from the split-group analyses and continuous assessment of memory largely overlap (compare with Supplementary Table 5). All significant values ($p < 0.05$) are in italics; those that reach FDR-corrected significance are in bold.

Supplementary Table 8 Average activation during learning compared between groups

Region	Mean activation							
	Anatomical ROI				Functional ROI			
	Mean conditioning-index (± s.d.)		Main effect (2)		Mean conditioning-index (± s.d.)		Main effect (2)	
	Retention ($n = 22$)	No retention ($n = 16$)	<i>F</i>	<i>p</i>	Retention ($n = 22$)	No retention ($n = 16$)	<i>F</i>	<i>p</i>
ACC	0.34 (±0.94)	0.52 (±0.70)	0.44	0.513	0.73 (±1.21)	0.76 (±0.81)	0.01	0.936
Amygdala	0.13 (±0.66)	0.11 (±0.80)	0.00	0.959	0.32 (±1.01)	0.15 (±1.01)	0.27	0.609
Hippocampus	-0.03 (±0.59)	-0.08 (±0.71)	0.04	0.844	-0.01 (±0.86)	-0.12 (±0.89)	0.15	0.698
Insula	0.38 (±0.94)	0.60 (±0.81)	0.60	0.442	0.75 (±1.07)	0.87 (±0.82)	0.13	0.720
SFG	-0.14 (±0.95)	0.10 (±0.82)	0.66	0.421	0.00 (±1.17)	0.20 (±0.93)	0.30	0.586
vmPFC	-0.13 (±0.95)	0.10 (±0.75)	0.65	0.427	-0.30 (±1.20)	0.15 (±1.04)	1.46	0.236

Summary of statistics of mean activation for the learning phase ($n = 38$, between-subjects ANOVA), compared between groups. The assessment of the predictive value of mean activation was done for atlas-based ROIs (left), and on ROIs based on individual functional ROIs (right). Within each ROI 'conditioning'-indices (see Methods), which expressed the relative increase of the CS^{+neg} responses over the course of learning, were compared between groups. None of the values reached statistical significance.

Supplementary Table 9 Brain areas showing differential activation during fear learning

Brain region (COG)	MNI coordinates			Volume	
	x	y	z	Size (mm ³)	Max. Z
Learning phase, CS ⁺ _{neg} > CS ⁺ _{neut} & CS-					
Group mean (n = 38)					
Anterior cingulate cortex	0	8	38	11396	7.42
Brain stem, thalamus	0	-14	-6	5593	6.28
Insula right	46	16	1	4298	6.75
Insula left	-44	13	-3	3644	7.80
Posterior supramarginal gyrus right	62	-40	25	2375	6.26
Posterior supramarginal gyrus left	-62	-35	28	1878	7.07
Cerebellum left	-35	-61	-38	1085	5.65
Frontal pole left	-34	48	24	567	4.79
Retention (n = 22) > No retention (n = 16)					
No significant clusters					
No retention (n = 16) > retention (n = 22)					
No significant clusters					

Whole brain activation ($Z > 2.3$, cluster-corrected at $p < 0.05$) that discriminates the threat-associated stimuli from the control stimuli, and within this contrast, activation that discriminates between groups. Univariate results are reported in order to facilitate integration and interpretation of the data in the light of the previous findings (LaBar & Cabeza, 2006; Sehlmeier et al., 2009) and assess whether there are any clusters of voxels outside the six anatomically defined ROIs that predict subsequent procedural fear memory. The three conditions (CS⁺_{neg}, CS⁺_{neut} and CS-, split by house and face stimuli) were separately modeled in a GLM by convolution with a double-gamma response function, taking into account the onset times and duration of the trials. We modeled target and filler trials separately and the UCS- and AS-trials in a similar fashion as the CS-trials. All trials and head motion parameters were included in the model. However, only target trials were used for higher-level analysis. Subsequently, a mixed-effect group analysis was conducted using the FMRIB FLAME stages 1 and 2. Main effects of fear learning were assessed by contrasting CS⁺_{neg} trials with CS⁺_{neut} and CS- trials, and comparing this contrast between the two groups. For all contrasts, a threshold value of $Z > 2.3$ was set with a cluster probability of $p < 0.05$, thereby correcting for whole-brain multiple comparisons (using Gaussian random field theory). No significant clusters were observed that distinguished between the groups. Uncorrected, at a threshold of $p < 0.001$, the largest cluster that distinguished between the groups consisted of 93 voxels. Coordinates are in MNI-space and depict for each significant cluster the Center of Gravity (COG),

Supplementary Table 10 fMRI data for the learning phase: functional ROIs

Region	Within-stimulus				Between-stimulus				Mean activation			
	Main effect		Interaction		Main effect		Interaction		Main effect		Interaction	
	(3)	(3 x 6)	(4)	(4 x 7)	(3)	(3 x 7)	(3)	(3 x 7)	(3)	(3 x 7)		
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
ACC	39.25	<0.0005	9.41	<0.0005	17.20	<0.0005	3.61	<0.0005	27.05	<0.0005	2.37	0.018
Amygdala	17.61	<0.0005	3.40	<0.0005	8.87	<0.0005	2.50	0.005	5.05	0.009	1.15	0.320
Hippocampus	12.38	<0.0005	3.20	0.001	7.74	<0.0005	2.04	0.027	3.46	<i>0.037^a</i>	0.84	0.571
Insula	68.46	<0.0005	10.38	<0.0005	30.14	<0.0005	7.20	<0.0005	18.89	<0.0005	3.28	0.001
SFG	45.53	<0.0005	10.65	<0.0005	22.44	<0.0005	7.80	<0.0005	2.69	0.075	NT	NT
vmPFC	13.61	<0.0005	5.33	<0.0005	11.24	<0.0005	3.17	<0.0005	3.32	<i>0.042^a</i>	1.38	0.173

Summary of statistics of the fMRI data for the learning phase ($n = 38$, within-subjects ANOVA) in six functional ROIs. Although interaction effects now become significant for mean activation in ACC and insula (compare with Supplementary Table 1), effect sizes are increased for similarity analysis as well and thus remain substantially larger for similarity analysis than for activation analysis. All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. NT = not tested; Areas without significant main effect of stimulus type are not tested for acquisition effects. ^a effect not caused by significantly higher values for CS^{neg} stimuli.

Supplementary Table 11 Within-stimulus pattern similarity predicting declarative memory

Region	Within-stimulus			
	Mean Fisher-transformed correlation (\pm s.d.)		Main effect (2)	
	<i>forgotten</i>	<i>remembered</i>	<i>F</i>	<i>p</i>
ACC	0.19 (± 0.38)	0.30 (± 0.25)	0.67	0.424
Amygdala	0.11 (± 0.25)	0.30 (± 0.23)	4.27	0.057
Hippocampus	0.09 (± 0.27)	0.31 (± 0.26)	5.28	<i>0.036</i>
Insula	0.16 (± 0.31)	0.27 (± 0.22)	0.84	0.375
SFG	0.23 (± 0.42)	0.36 (± 0.30)	0.70	0.417
vmPFC	0.19 (± 0.26)	0.22 (± 0.25)	0.10	0.760

Summary of statistics of within-stimulus pattern similarity for the learning phase ($n = 16$, within-subjects ANOVA) in six anatomical ROIs. Comparisons are made between stimuli based on long-term declarative memory for the CS-UCS and CS-AS associations. Starting the memory phase, participants received a sheet presenting the 6 stimuli they had seen before and were asked to identify the stimuli ('forced choice') and to rate their degree of certainty. Twenty-two participants correctly identified all six stimuli and 16 participants mixed up at least two stimuli. To establish the predictive value of pattern similarity for declarative memory we compared subsequently recognized items with subsequently forgotten items (averaged over experimental conditions and trials) within the 16 participants that had mixed up at least one stimulus. Stimuli that were subsequently remembered elicited more similar patterns of activation in the hippocampus on consecutive presentations than stimuli that were subsequently forgotten. Although not completely comparable (declarative memory in this study refers to memory for the association, whereas in the study by Xue and colleagues (2010) declarative memory referred to the recognition of the item itself) this is in line with Xue et al. (2010). All significant values ($p < 0.05$) are in italics; none of the values survives FDR-correction.

Supplementary Table 12 Between-stimulus pattern similarity during learning compared between groups: equal-sized ROIs

Region	Between-stimulus					
	Mean conditioning-index (\pm s.d.)		Main effect (2)			
	Retention (n = 22)	No retention (n = 16)	<i>F</i>	<i>p</i>	<i>r</i>	<i>p</i>
ACC (2967 voxels)	2.58 (\pm 3.00)	1.35 (\pm 3.10)	1.53	0.224	0.230	0.164
Hippocampus (2967 voxels)	2.76 (\pm 2.70)	0.98 (\pm 2.47)	4.30	<i>0.045</i>	0.223	0.179
Insula (2967 voxels)	3.31 (\pm 2.23)	1.44 (\pm 2.37)	6.20	<i>0.018</i>	0.351	<i>0.031</i>
SFG (2967 voxels)	3.14 (\pm 2.39)	1.81 (\pm 2.22)	3.05	0.089	0.376	<i>0.020</i>
vmPFC (2967 voxels)	2.28 (\pm 2.88)	-0.33 (\pm 2.93)	7.50	<i>0.010</i>	0.336	<i>0.039</i>

Summary of statistics of between-stimulus pattern similarity for the learning phase ($n = 38$). ROIs are downsized to match the number of voxels from the smallest ROI (amygdala). First, we calculated the distance between each voxels within a ROI and the center of gravity (COG; based on probability values) of that particular ROI. Next, voxels were subtracted from this ROI, in order of distance to the COG (highest to lowest), until the ROI contained 2967 voxels. The main predictive analyses were repeated for these equal-sized ROIs, to examine whether the number of voxels within each ROI could explain the dissociation between different brain regions with regard to how between-stimulus pattern similarity predicts fear memory (continuous versus discrete). Hereto 'conditioning'-indices, which expressed the relative increase of the CS_{neg}^+ responses over the course of learning, were compared between groups (discrete, left) and correlated with differential pupil dilation responses during the memory phase (continuous, right). Results are, with the exception of the ACC, comparable to results from the original ROIs (Supplementary Table 5, 7), so the size of the ROI cannot account for the dissociation between the results from the continuous assessment of fear memory and the results from the split-group analysis. All significant values ($p < 0.05$) are in italics; those that reach FDR-corrected significance are in bold.