Neural signals of vicarious extinction learning

Armita Golkar*, Jan Haaker, Ida Selbing & Andreas Olsson

Department of Clinical Neuroscience, Karolinska Institutet

Nobels väg 11, 17177

Stockholm, Sweden

***Corresponding author:**

Armita Golkar

email: Armita.golkar@ki.se

Department of Clinical Neuroscience, Karolinska Institutet,

Nobels väg 11, 17177

Stockholm, Sweden

Abstract

Social transmission of both threat and safety is ubiquitous, but little is known about the neural circuitry underlying vicarious safety learning. This is surprising given that these processes are critical to flexibly adapt to a changeable environment. To address how the expression of previously learned fears can be modified by the transmission of social information, two conditioned stimuli (CS+s) were paired with shock and the third was not. During extinction, we held constant the amount of direct, non-reinforced, exposure to the CSs (i.e. direct extinction), and critically varied whether another individual – acting as a demonstrator - experienced safety ($CS+_{vic\ safety}$) or aversive reinforcement ($CS+_{vic\ reinf}$). During extinction, vmPFC responses to the $CS+_{vic\ reinf}$ increased but decreased to the $CS+_{vic\ safety}$. This pattern of vmPFC activity was reversed during a subsequent fear reinstatement test, suggesting a temporal shift in the involvement of the vmPFC. Moreover, only the $CS+_{vic\ reinf}$ association recovered. Our data suggest that vicarious extinction prevents the return of conditioned fear responses and that this efficacy is reflected by diminished vmPFC involvement during extinction learning. The present findings may have important implications for understanding how social information influence the persistence of fear memories in individuals suffering from emotional disorders.

Keywords: Social learning, vicarious learning, extinction, amygdala, vmPFC

Introduction

The neural processes underlying how socially transmitted information influence prior, direct, learning are unknown. This is surprising given that these processes are likely to be critical to functioning adaptively in a changeable environment for both humans and other animals (Laland, 2004). Here, we focused on understanding the neural processes involved in using social information gleaned through observation to attenuate the expression of previously learned fear responses. Learning safety through observing others (vicarious safety learning) is ubiquitous in human and non-human animals, and serves a key role in the development of both healthy and dysfunctional behavior (Bandura, 1977).

Recently, we established an experimental model to study vicarious safety learning in humans through vicarious extinction of directly conditioned fear (i.e. learning from the safety experience of another individual –the so called *demonstrator*) (Golkar et al., 2013). Critically, and in contrast to direct extinction, vicarious extinction augmented safety learning by blocking the return of learned fear responses, as measured by skin conductance responses (SCR). Return of fear is commonly observed after standard, direct, extinction and has strong clinical relevance as a model for relapse after successful exposure treatment of anxiety disorders (Hartley and Casey, 2013; Maren et al., 2013). Moreover, in spite of a growing understanding of safety learning through direct fear extinction, which is known to involve the ventromedial prefrontal cortex (vmPFC) in both rodents (Milad and Quirk, 2002) and humans (Phelps et al., 2004), the neural circuitry underlying vicarious extinction learning remains unexplored.

In order to address how the expression of previously learned fears can be modified by the transmission of social information, we used a within-subject design in which two conditioned

stimuli (CS+s) were paired with shock and the third was not (CS-). During extinction, we held constant the amount of direct safe, non-reinforced, exposure to the CSs (i.e. direct extinction), and critically varied whether the demonstrator experienced safety (non-reinforced exposure) or danger (reinforced exposure). Finally, we assessed the recovery of fear by reinstating the CR through unsignaled presentations of the shock (see Figure 1). Based on our previous study on vicarious extinction learning (Golkar et al., 2013), we expected that the return of conditioned fear responses would be evident only for the vicariously reinforced cue ($CS+_{vic\ reinf}$), and that this return of fear would be accompanied by an increase in threat-related amygdala activity, as typically observed following standard extinction (Agren et al., 2012; Lonsdorf et al., 2014). In contrast, the shared experience of safety during exposure to the vicariously extinguished cue ($CS+_{vic\ safety)}$ was expected to strengthen the retention of extinction learning and attenuate the psychophysiological and neural expressions of fear recovery. If the efficacy of vicarious extinction reflects an augmentation of safety learning, we expected this safety learning to be reflected by increased activity in the vmPFC to the $CS+_{vic\ safety}$ vs. the $CS+_{vic\ reinf}$, in accordance with its role in direct fear extinction in both human (Phelps et al., 2004; Kalisch et al., 2006; Milad et al., 2007) and non-human animals (Milad and Quirk, 2002), as well as in safety signaling more generally (Schiller et al., 2008).

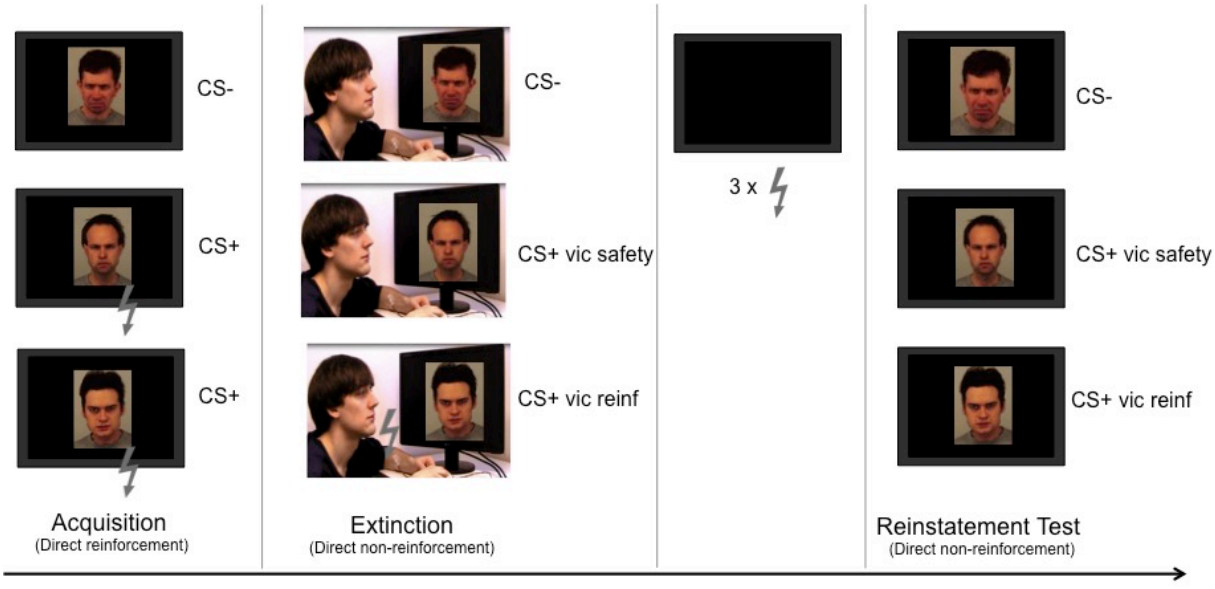

Figure 1. Experimental design. The experiment was divided into different stages. Within each stage, all conditioned stimuli (CSs) were presented 8 times each in a pseudorandomized order. During *Acquisition*, two angry faces (CS+s) were repeatedly paired with a mild electric shock (Unconditioned stimulus, US) given to the participants' wrist (6 reinforced presentations/CS). The third angry face (CS-) was never paired with the shock. During *Extinction*, participants watched a video depicting an individual (the demonstrator) acting calmly when exposed to non-reinforced presentations of the CS- and to one of the previously reinforced CS+s ($CS+_{vic\ safety}$), but received shocks on the presentations of the other CS+ ($CS+_{vic\ reinf;}$ 6 reinforced presentations). The demonstrator reacted to the shocks by twitching the arm and blinking. Critically, the participants did not receive any shocks during this stage. Finally, participants were then re-exposed to all three CSs after receiving three reminder shocks during the *Reinstatement-test*.

Materials and Methods

Participants. Based on sample sizes in previous research on vicarious fear learning (Olsson et al., 2007) and vicarious extinction learning (Golkar et al., 2013) we planned to include 20 participants in the current study. Therefore, we recruited a total of 23 male, right-handed

112 participants who were free from self-reported life-time psychiatric or neurological disease and
113 medication. We used the stopping rule that all participants had to complete the functional
114 magnetic resonance (fMRI) task and the questionnaires. Prior to analysis, we excluded 2
115 participants because they failed to report the contingency between the conditioned stimuli
116 (CSs) and the unconditioned stimulus (US) and 1 participant with abnormal brain anatomy
117 leaving a final sample of 20 participants with a mean age of 25 years (SD = 1.25). All
118 participants gave written informed consent and were paid 350 SEK (approximately 50 USD)
119 for their participation. Due to technical problems, the skin conductance data were missing for
120 1 participant, who was excluded from all statistical analyses of the skin conductance data.
121
122 **Stimuli.** Three pictures depicting angry male faces from the Karolinska Directed Emotional
123 Faces database (Lundqvist, 1998) served as CSs (Items AM02ANS; AM04ANS; AM06ANS).
124 During each stage of the experiment, each CS was presented 8 times, with a duration of 6s.
125 The inter-trial interval between each CS was jittered between 11-15 s. The US consisted of a
126 100-ms DC-pulse electric stimulation applied to the participant's right wrist. The coupling
127 between a specific conditioned face stimulus and the US, and the order of presentation of the
128 two CS+s (CSs that were coupled to the US) was counterbalanced between participants. For
129 the extinction stage, we created two movies (counterbalancing the order of the CS+
130 presentations) using Adobe Premiere Pro CS5.5 that was each 4 min and 18 s in length. The
131 movies showed the demonstrator sitting in front of a computer screen watching the CS
132 presentations. Which face that served as $CS+_{vic\ safety}$ and the $CS+_{vic\ reinf}$ was counterbalanced
133 between participants. A shock electrode was visibly attached to the demonstrator's right wrist.
134 Apart from the order of CS+ presentations, the movies were identical in terms of content and
135 timing.
136

Vicarious extinction learning

Experimental procedure. The experiment consisted of three experimental stages: Acquisition, Extinction and Reinstatement testing. Before starting the experimental task, participants were attached to SCR and shock electrodes and underwent a standard work-up procedure in order to adjust the level of the shock to be experienced as "uncomfortable but not painful". Following this, participants underwent a direct acquisition task during which each CS was presented eight times, out of which six presentations of each of the CS+s co-terminated with a 100-ms shock given to the wrist of the participant. The presentation of the CS- was never paired with a shock. After the direct acquisition stage, participants were given the following instructions: *"During the next stage you will watch a movie of another person, attached to the same equipment as you, who will undergo a similar experiment as the one you are participating in. Remember to attend to the picture display."* During the observational extinction stage that followed, participants watched a movie (see Figure 1) depicting the demonstrator in front of a screen on which the CSs were presented again (each presented eight times). In the movie, the demonstrator acted calmly while watching the presentations of the CS-, and one of the previously reinforced CS+s (the $CS+_{vic\ safety}$), but received shocks on 75% of the presentations of the other CS+ ($CS+_{vic\ reinf}$). The model reacted to the shocks by slightly twitching the arm and blinking. After the end of the observational extinction stage (i.e. after completion of the movie), participants read the following instructions: *"You will now watch the images on your screen again. The setup of the experiment will be the same as before you watched the movie. The presentation will begin with a black screen. Remember to attend to the picture display.*

To assess the return of fear, these instructions were followed by a standard reinstatement procedure during which participants received unsignaled reminder shocks before they were directly re-exposed to the CSs. This procedure has been shown to reinstate the expression of

the original fear memory in both animals (Bouton, 2002) and humans (Haaker et al., 2014) and is a commonly employed to model clinical relapse of anxiety symtoms. During the reinstatement procedure, participants were exposed to a black screen for 30s, after which they received three reminder presentations of the US. This procedure was followed by the reinstatement test stage, in which each CS was again presented without the US eight times in a pseudorandom order with the first trial always a CS− to capture the orienting response.

Subjective ratings. Participants completed a post-experimental interview assessing CS-US contingency awareness, and rated on a scale from 1 (not at all) to 7 (very much) how much discomfort they experienced when observing the person in the movie receiving shocks and how much discomfort they thought that the person in the movie experienced when receiving shocks, how much they identified themselves with the person in the movie and how much empathy they felt for the person in the movie on a scale. Finally, they rated how much they liked the person in the movie on a scale from -3 (disliked) to 3 (liked).

Psychophysiological assessment. Skin conductance responses (SCRs) to each CS were measured throughout the experiment and the raw signal was off-line filtered with a low-pass filter at 1 Hz and a high-pass filter at .05 Hz. For each CS trial, conditioned SCRs were measured as the peak-to-peak amplitude difference in skin conductance to the largest response (in microsiemens, μS) in the .5 to 4.5 second window following stimulus onset. Responses below .02 μS were scored as zero and data was z-transformed prior to analysis, (Boucsein et al., 2012).

Image acquisition and pre-processing. fMRI data were obtained with a 3 Tesla MR scanner (General Electrics 750) using an 8-channel head coil. Each functional image volume

187 comprised 46 continuous axial slices (2.3 mm thick, no gap) that were acquired using a T2*-
188 sensitive gradient echo-planar imaging (EPI) sequence [repetition time (TR): 3000 ms; echo
189 time (TE): 31 ms; flip angle: 85°; field of view (FOV): 96 x 96 mm, 3 x 3 mm in-plane
190 resolution]. To account for T1 equilibrium effects, the first 5 volumes of each time series were
191 discarded. High-resolution T1-weighted structural images (1x1x1mm) were acquired after the
192 experimental session. Pre-processing using Statistical parametric mapping (SPM8,
193 (www.fil.ion.ucl.ac.uk/spm)) running on Matlab2013b (The MathWorks, Natick, MA)]
194 involved realignment, unwarping co-registration and normalization to a sample-specific
195 template, using DARTEL (Ashburner, 2007). Normalized data series were spatially smoothed
196 with a 6 mm FWHM isotropic Gaussian kernel and manually inspected for excessive head
197 movement. Further processing included temporal high-pass filtering (cut-off 128 s) and
198 correction for temporal auto-correlations using first-order autoregressive modelling.
199
200 **Regions of interest (ROI) selection.** The pre-defined ROIs included two key structures of
201 fear and safety memory processing in humans: the amygdala (LaBar et al., 1998; Phelps et al.,
202 2004) and the vmPFC (Phelps et al., 2004; Kalisch et al., 2006; Milad et al., 2007). The
203 amygdala ROI was defined as an anatomical mask derived from the automatic anatomical
204 labelling atlas (Tzourio-Mazoyer et al., 2002). The vmPFC ROI was defined as a box
205 (20x16x16mm) around the average peak coordinate [xyz (MNI) = 0,41,-12] of previous
206 human fMRI studies (Phelps et al., 2004; Kalisch et al., 2006; Milad et al., 2007; Spoormaker
207 et al., 2010; Haaker et al., 2013; Rabinak et al., 2013; Lonsdorf et al., 2014) testing for
208 extinction recall.
209
210 **Statistical analyses.** For the SCR data, each stage of the experiment (Acquisition, Extinction
211 and Reinstatement-test) was analyzed with separate repeated measures analysis of variance

(ANOVA). For the fMRI data, a general linear model (GLM) with a total of 15 regressors was set up for statistical first-level (single-subject) analysis: one regressor per CS type in each phase named after their functional significance during the extinction stage (Acquisition: $CS+_{vic\ safety\ to\text{-}be}$, $CS+_{vic\ shock\ to\text{-}be}$, CS-; Extinction: $CS+_{vic\ safety}$, $CS+_{vic\ reinf}$, CS-; Reinstatement: $CS+_{previously\ vic\ safety}$, $CS+_{previously\ vic\ shock}$, CS-) which modelled the onset of each cue as an event using a stick function. Two regressors were included to model each onset (as a stick function) of the US to the CS+s during acquisition ($US_{vic\ safety\ to\text{-}be}$, $US_{vic\ shock\ to\text{-}be}$). During extinction, we modelled the vicarious US (administered to the model) to $CS+_{vic\ shock}$ and the omission of each shock to the $CS+_{vic\ safety}$. In addition, two nuisance regressors were included to factor out experimental effects of no interest: one regressor modeled the whole duration (as a boxcar function) of each ITI (including the rest period after the reinstatement-USs) and another nuiscance regressor modelled the reinstatement-USs (as a stick function). All regressors were convolved with a canonical hemodynamic response function. Random effect analysis on the group level was performed using SPM's "full factorial" model and focused on comparisons between the $CS+_{vic\ safety}$ and the $CS+_{vic\ reinf}$ for the effect of vicarious extinction. Separate analyses for each session included beta-estimates for each CS (1 factor, 3 levels), derived from individual single subjects general linear modeling. We also included a comparison for the effects of reinstatement between extinction and reinstatement (2 factors with 3 levels each) to test the enhancement of responses through reinstatement. P-values inside our ROIs were corrected for multiple testing (small volume correction, SVC) using family-wise error (FWE) correction. For illustrative purposes, estimated responses were calculated and plotted within the rfx plot toolbox (http://rfxplot.sourceforge.net/), displaying the mean estimated time course within each ROI, scaled to the onset of each CS. Hypothesis generating effects outside our ROIs with a high uncorrected p-value ($p < .001$) and a liberal threshold of $k>5$ voxel are reported for each analysis in Supplementary Table 1. To examine condition-specific

functional connectivity during extinction and reinstatement tetsing, each participant's BOLD signal time-course at the individual peak within the vmPFC ROI (for extinction: from the $CS+_{vic\ reinf}$ > $CS+_{vic\ safety}$ contrast; for reinstatement: from the $CS+_{vic\ safety}$ > $CS+_{vic\ reinf}$, thresholded at $p < .05$ uncorrected) was extracted as an eigenvariate. The time course was deconvolved and multiplied with the condition specific onset (e.g. onset of the $CS+_{vic\ reinf}$ or $CS+_{vic\ safety}$ during extinction or reinstatement). This Psycho-physiological interaction (PPI) was entered as a regressor into a general linear model for each participant including the vmPFC time-course and the regressors for the different conditions, as nuisance regressors (see Supplementary information). Parameter estimates for each CS condition were then contrasted using one-sample t-test (for extinction: $CS+_{vic\ reinf}$ > $CS+_{vic\ safety}$; for reinstatement: $CS+_{vic\ safety}$ > $CS+_{vic\ reinf}$).

Results

Subjective ratings. On a scale from 1 (not at all) to 7 (very much), participants rated how much discomfort they experienced when observing the model receive shock (M = 2.28; SD = 1.13), how much discomfort they though the model experienced when he received a shock (M = 3.5; SD = 1.30), how much they could identify themselves with the person in the movie (M = 4.17; SD = 1.15), how much empathy they felt for the person in the movie (M = 3.17; SD = 1.54), and how much they liked the person in the movie on a scale from -3 (disliked) to 3 (liked) (M = 0.44; SD = .07). None of these ratings was significantly related to the extinction or reinstatement data as assessed with correlation analysis (all p's < .05).

SCR. Mean SCRs to each CS are displayed in Figure 2A and demonstrate a replication of our previous findings on the efficacy of vicarious extinction (Golkar et al., 2013). During acquisition, there was a predicted main effect of Stimulus ($F(2,36) = 13.27$, $p < .001$; η^2 =

.42), showing that mean SCRs to both CS+s were larger than to the CS- ($CS+_{vic\ safety\ to\text{-}be:}$ t(18) = 4.20, p = .001, 95% confidence interval (CI) for the difference between conditions = [0.12, 0.36]; ($CS+_{vic\ reinf\ to\text{-}be:}$ t(18) = 5.05, p < .001; 95% CI = [0.21, 0.51]), and that the CS+s did not differ from each other (t(18) = 1.45, p = .16) CI = [-0.05, 0.30]. The conditioned fear responses diminished during extinction (no main effect of Stimulus: F(2,36) = 1.39, p = .26 and a significant main effect of Trial: F(7,126) = 3.60, p = .002; η^2 = .17) and there were no between-stimulus differences left at the end of extinction, defined as the mean response during the last two trials of each CS (all t's < 1). To establish whether vicarious extinction successfully reduced the return of fear, we analyzed the change in mean SCRs from extinction to reinstatement-testing using a Stimulus (3) x Time (2) ANOVA that resulted in a significant interaction (F(2,36) = 3.70, p = .03; η^2 = .17). Follow-up t-tests revealed a smaller increase in SCR to the $CS+_{vic\ safety}$ (t(18) = 2.10, p = .05), 95% CI = [-2.11, 0.00]) than to the $CS+_{vic\ reinf}$ (t(18) = 3.58, p = .002; 95% CI = [-0.41, -0.11]; CS- t(18) = 2.90, p = .009), 95% CI = [-0.47, -.08]), and planned comparisons revealed that mean SCRs during the reinstatement-test were significantly lower to the $CS+_{vic\ safety}$ than to the $CS+_{vic\ reinf}$ (t(18) = 2.58; p = 0.019), 95% CI = [0.03, 0.31]). The trial-by trial data from extinction to reinstatement-testing are displayed in Figure 2B. Inspection of the data revealed an increase in conditioned fear responding that generalized to the CS-, which is commonly reported in the reinstatement literature (Haaker et al., 2014). Given that the first CS presentation always was a CS- (to capture the orienting response, i.e. the immediate response to a change in the environment), we ran an additional analysis in which we excluded the first CS- trial (see also Schiller et al., 2010 for the same rationale) that resulted in a significant interaction (F(2,36) = 4.22, p = .02; η^2 = .19) explained by a significantly higher SCR to the $CS+_{vic\ reinf}$ vs. the CS- (t(18) = 2.11, p = .049), 95% CI = [0.00, .377]), and no differences between the $CS+_{vic\ safety}$ vs. the CS- (t(18) = .36, p = .72), 95% CI = [-0.134, 0.095]).

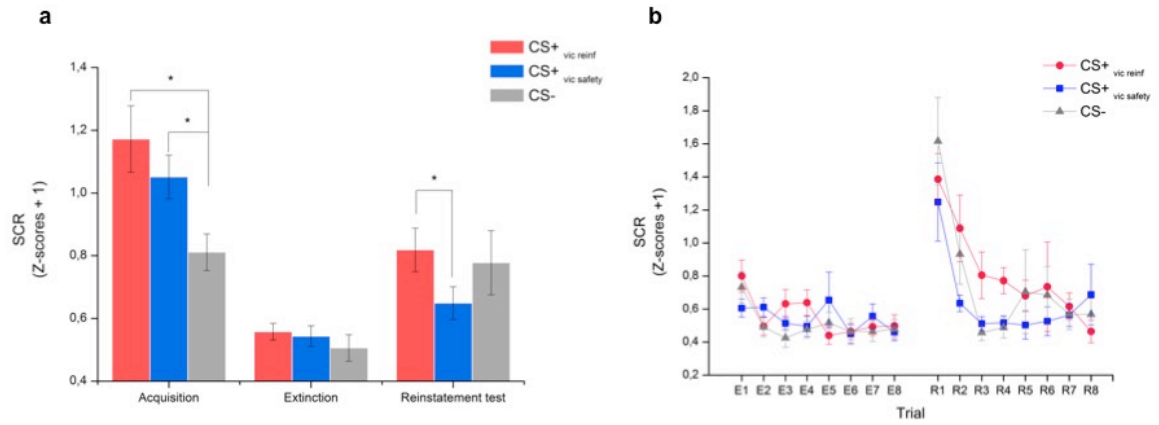

Figure 2 (a)**.** Mean skin conductance responses (SCR) as a function of experimental stage and conditioned stimulus (CS). (b). Trial-by-trial data for Extinction (E1-E) and Reinstatement (R1-R8). Note that in order to capture the immediate response to a new context (i.e. the orienting response), the first CS presentation during the reinstatement test was always a CS-. Error bars indicate standard error of the mean (SEM). Asterisks indicate statistically significant differences ($p < .05$).

fMRI. Using fMRI, we sought to identify the neural processes underlying how vicariously transmitted information modulated learned fear. Specifically, we examined the Blood-oxygen-level dependent (BOLD) signal differences between the $CS+_{vic\ safety}$ and the $CS+_{vic\ reinf}$ during extinction learning and the reinstatement-test. Based on the existing literature on direct extinction learning (Phelps et al., 2004; Kalisch et al., 2006; Milad et al., 2007; Milad et al., 2009), we specified two separate regions-of-interest (ROI); the amygdala and the vmPFC.

Acquisition and extinction of threat memory. First, we confirmed that activity in the amygdala was greater to the CS+s compared to the CS- during acquisition [x,y,z (MNI) = 31,0,-27; T = 3.11; Z = 2.98; p(SVC) = .047, p(uncorrected) <.001] (see Supplementary Table 2). Mirroring the SCR data, extinction learning revealed no significant activation differences in the amygdala between the CS+s or between either of the CS+ and the CS-. The only

difference that emerged was a strong trend towards an increased activity in the vmPFC ROI to the $CS+_{vic\ reinf} > CS+_{vic\ safety}$ [x,y,z(MNI) = 9,45,-9; Z = 3.21; p(SVC) = .064, p(uncorrected) <.001]. To further characterize this difference, we ran a separate model including parametric regressors for each CS modelling linearly increasing (trial-by-trial) responses (see Figure 3a). This model revealed that activity within the vmPFC ROI increased to the $CS+_{vic\ reinf,}$ whereas responses to the $CS+_{vic\ safety}$ decreased [x,y,z (MNI) = -9,48,-6; Z = 3.75; p(SVC) = .01, p(uncorrected) <.001], consistent with previous studies comparing responses between a conditioned threat and a safe cue during extinction (Phelps et al., 2004; Milad et al., 2007).

Condition-specific functional connectivity during extinction. We further examined the condition-specific functional connectivity during extinction using the seed region inside the previously defined vmPFC ROI that displayed a difference in activity between the $CS+_{vic\ reinf}$ > $CS+_{vic\ safety}$ ($p < .05$). We found that the connectivity between the vmPFC and a region located in the lateral amygdala and the anterior hippocampus (see Figure 3b, Table 1) was more positive for the $CS+_{vic\ reinf}$ vs. the $CS+_{vic\ safety}$.

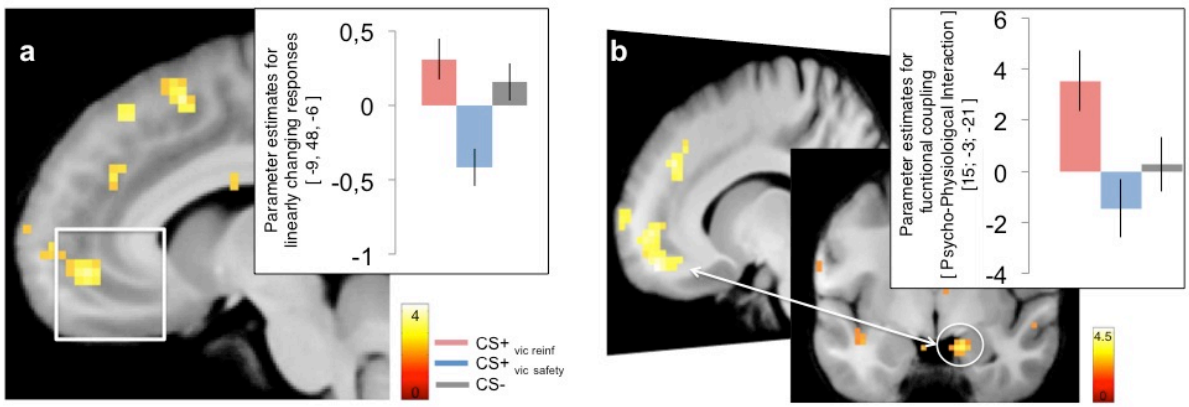

Figure 3. (a) Activity in the vmPFC ROI increased linearly to $CS+_{vic\ reinf}$ during extinction learning whereas responses in this region decreased to the $CS+_{vic\ safety.}$ (b). Functional connectivity during extinction using the vmPFC ROI as seed revealed coupling with a region

located in the lateral amygdala and anterior hippocampus that was stronger during the $CS+_{vic\ rein}$ vs. the $CS+_{vic\ safety}$. T-values are superimposed on a normalized average structural image. fMRI display threshold: $p < .005$, uncorrected for illustrative purposes. Error bars represent standard error of the mean (SEM).

Table 1. Whole brain results of the Psycho-physiological interaction (PPI) [p(uncorr) < .001 and clustersize (k) > = 5]

Contrast / Region	**T**	**Z**	**Coordinates**
$CS+_{vic\ reinf} > CS+_{vic\ safety}$			
right lateral amygdala / anterior para-hippocampus	4.25	3.52	15; -6; -21
left temporal sulcus	4.24	3.51	-54; 12; 9
left dorsal cingulate	4.10	4.34	-12; 24; 27
$CS+_{vic\ safety} > CS+_{vic\ reinf}$			
no cluster above threshold			

Note: Coordinates are given in MNI space

Reinstatement of threat association: role of the amygdala. During the reinstatement-test, we confirmed that our reinstatement manipulation successfully engaged the amygdala by directly contrasting the neural responses to the $CS+_{vic\ reinf}$ with the CS-. This analysis revealed a greater activity to the $CS+_{vic\ reinf}$ vs. the CS- in the left amygdala [xyz(MNI) = -24,0,-21; Z = 2.98; p(SVC) = .04, p(uncorrected) <.001], but no difference between the $CS+_{vic\ safety}$ and the CS- (no voxel above p(uncorrected)<0.01 in the ROI), demonstrating that vicarious extinction blocked the reinstatement of defensive responses. In fact, amygdala responses to the $CS+_{vic\ safety}$ was intermediate between the amygdala responses to the CS- and the to the $CS+_{vic\ reinf}$, and there were no significant difference between the amygdala response between the CS+s.

(no voxel above p(uncorrected)<0.01 in the ROI), (Figure 4a). Additionally, we observed that the difference in reinstated amygdala response in the $CS+_{vic\ reinf} > CS+_{vic\ safety}$ contrast was positively correlated with reinstated SCRs ($CS+_{vic\ reinf} > CS+_{vic\ safety}$) during the reinstatement-test ($r = .78$, $p < .001$; 30, -9, -12; $t = 5.01$; $p(FWE) = .005$) (Figure 4b).

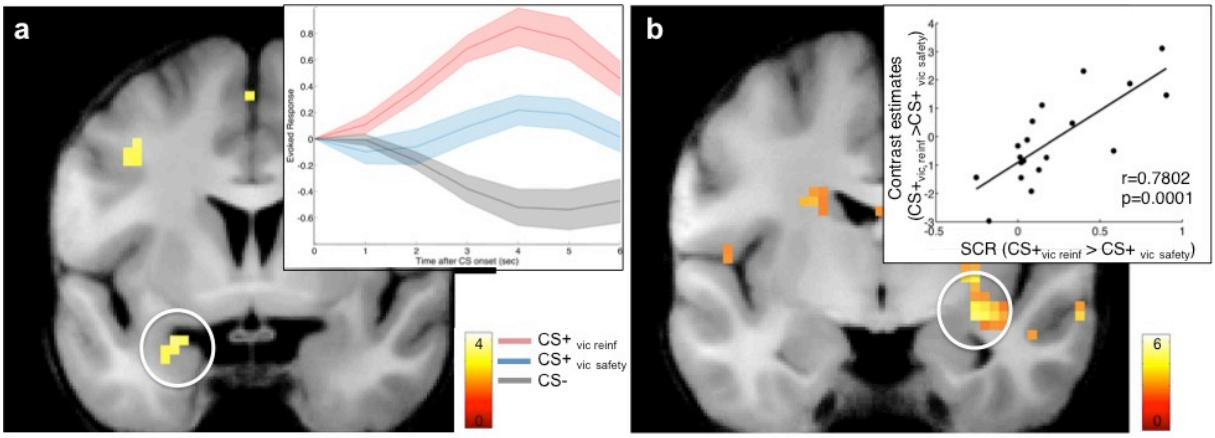

Figure 4. (a). Mean estimated evoked responses (arbitrary units) within the amygdala during reinstatement testing scaled to the response at onset of each CS. Shaded areas represent standard errors of the means (SEM). (b). Correlation between amygdala activity during Reinstatement-test in the contrast $CS+_{vic\ reinf} > CS+_{vic\ safety}$ and the reinstated SCRs for $CS+_{vic\ reinf} > CS+_{vic\ safety}$. T-values are superimposed on a normalized average structural image. fMRI display threshold: $p < .005$, uncorrected for illustrative purposes.

Reinstatement of threat association: role of the vmPFC. To investigate the differences between the vicariously learned cues during reinstatement, we investigated the change in BOLD activity from extinction to reinstatement testing within the vmPFC ROI. This analysis revealed a significant interaction that was explained by a larger increase in vmPFC activity from extinction to reinstatement testing for the $CS+_{vic\ safety}$ as compared to the $CS+_{vic\ reinf}$ [x,y,z (MNI) = 9,45,-9; $Z = 3.73$; $p(SVC) = .011$, p(uncorrected) <.001], see Figure 5a. As predicted, mean activity within the vmPFC during reinstatement was larger to the $CS+_{vic\ safety}$ as compared to the $CS+_{vic\ reinf}$ [x,y,z (MNI) = -9,48,-15; $Z = 3.99$: $p(SVC) = .004$,

p(uncorrected) <.001], but did not differ from the CS- (p(SVC) = .236; Z = 2.51). The individual average peak responses within the vmPFC ROI were enhanced to the $CS+_{vic\ safety}$, intermediate to the CS- and decreased to the $CS+_{vic\ reinf}$ (Figure. 5b), echoing the ordinal pattern in the SCR data.

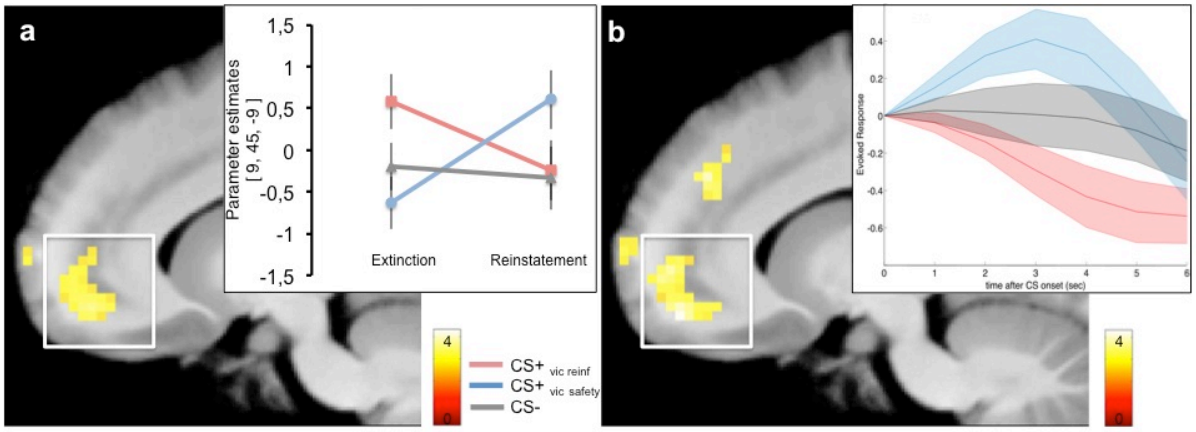

Figure 5. (a). Change in vmPFC activity from extinction to reinstatement testing displayed for all CSs separately (b). Mean estimated evoked responses (arbitrary units) within the vmPFC during Reinstatement-test scaled to the response at onset of each CS. Shaded areas represent standard errors of the means (SEM). T-values are superimposed on a normalized average structural image. fMRI display threshold: $p < .005$, uncorrected for illustrative purposes.

Conditions-specific functional connectivity during reinstatement. Finally, we examined the condition-specific functional connectivity during reinstatement using the vmPFC ROI (from the $CS+_{vic\ safety} > CS+_{vic\ reinf}$ contrast; $p < .05$) as seed. We found that connectivity between the vmPFC and the anterior hippocampus, as well as the inferior temporal gyrus (see Table 2) was stronger between the vicariously extinguished CS+ versus the vicariously reinforced CS+. Whole brain results for all stages of the experiment are reported in Supplementary Table 2.

Table 2. Whole brain results of the Psycho-physiological interaction (PPI) [p(uncorr) < .001 and clustersize (k) > = 5] during Reinstatement

Contrast / Region	**T**	**Z**	**Coordinates**
$CS+_{vic\ safety} > CS+_{vic\ reinf}$			
dorso-medial PFC	4.81	3.84	-9; 54; 39
left inferior temporal gyrus	3.47	3.02	-48;-54;-18
left anterior hippocampus	3.41	2.97	-21;-9;-24
$CS+_{vic\ reinf} > CS+_{vic\ safety}$			
no cluster above threshold			

Note: Coordinates are given in MNI space

Discussion

Our study demonstrates that socially transmitted safety information prevent previously learned fear responses from recovering, reaffirming the efficiency of vicarious safety learning accomplished through vicarious extinction (Golkar et al., 2013). Vicariously transmitted inhibition of fear during the reinstatement test was associated with enhanced vmPFC activity and a more positive connectivity between the vmPFC and the amygdala/anterior hippocampus, as compared to a vicariously reinforced CS+ ($CS+_{vic\ reinf}$). Vicarious reinforcement of a previous learned fear association, on the other hand, resulted in significant recovery of conditioned fear responses.

During the extinction stage, when participants did not receive any shocks themselves, we observed an increase in vmPFC activity to the vicariously reinforced CS+ ($CS+_{vic\ reinf}$), similar

to what is typically reported during standard extinction learning (Phelps et al., 2004; Milad et al., 2007; Schiller et al., 2013). Interestingly, exposure to the vicariously extinguished CS+ ($CS+_{vic\ safety}$) did not seem to engage this circuitry, suggesting that vicarious extinction learning might bypass the engagement of the vmPFC during extinction learning. Whereas the pattern of the vmPFC to the $CS+_{vic\ reinf}$ is consistent with the suggested role of the vmPFC in fear suppression during direct extinction learning, the reduced engagement of the vmPFC in response to the $CS+_{vic\ safety}$ during extinction was not predicted. Such reduced engagement of the vmPFC during extinction has, however, been reported in other extinction procedures that have resulted in less return of conditioned responding (Kim and Richardson, 2010; Schiller et al., 2013). Most recently, vmPFC activity during extinction training initiated shortly after a reactivation trial (i.e. during reconsolidation) decreased to the reactivated CS+ compared to a non-reactivated CS+, and conditioned responses to the reactivated CS+ did not recover during a subsequent reinstatement test (Schiller et al., 2013). Although similar neural patterns do not imply overlapping mechanisms (i.e. reverse inference, see Poldrack, 2006), the shared experience of safety during vicarious extinction in our study might have reduced the necessity of an inhibitory vmPFC-amygdala circuitry during extinction and enabled a better prevention of the return of defensive responses as compared to what is accomplished by standard extinction only. It is unclear from the present data whether this was accomplished through unlearning of the original CS-US association, strengthening of the extinction association or by neutralizing the affective value of the CS+.

Interestingly, in our data, vicarious modulation of previously learned fear was associated with a temporal shift in the involvement of the vmPFC from extinction and reinstatement test. Accordingly, during extinction learning, the vmPFC displayed a reduced activity, and less functional connectivity with a region located within the lateral amygdala, in response to the

422 $CS+_{vic\ safety}$ compared to the $CS+_{vic\ reinf}$. Conversely, during the reinstatement test, activity
423 within the vmPFC increased, and showed stronger coupling with the anterior hippocampus, in
424 response to $CS+_{vic\ safety}$ compared to the $CS+_{vic\ reinf}$. This strengthened coupling between the
425 vmPFC and the amygdala/hippocampus is in line with the proposed role of this network in
426 gating successful recall of context-dependent extinction memory (Kalisch et al., 2006; Milad
427 et al., 2007). Moreover, the diminished vmPFC activity to the $CS+_{vic\ reinf}$ during reinstatement
428 is in line with previous studies demonstrating diminished vmPFC activity in post-traumatic
429 stress disorder (PTSD) patient during extinction recall failure (Milad et al., 2009; Garfinkel et
430 al., 2014), and might provide a route through which previously learned fears can be
431 maintained through social reinforcement.
432
433 On a more general level, our finding of increased activity in the vmPFC in response to the
434 vicariously extinguished CS+ during reinstatement is consistent with a suggested role of the
435 vmPFC in integrating information from distributed brain regions involved in signaling
436 affective value, episodic memory and social cognition (Roy et al., 2012), and using this
437 information to provide a selective safety signal that indicates which stimuli are safe to ignore
438 (Schiller et al., 2008). In the present study, the vmPFC appears to track the relative cue value,
439 by responding more to the relatively more dangerous cue ($CS+_{vic\ reinf}$ vs the $CS+_{vic\ safety}$) when
440 presented in the safe, extinction, context and conversely, shift to responding more to the
441 relatively safe cue ($CS+_{vic\ safety}$ vs. $CS+_{vic\ reinf}$) presented in the dangerous, reinstatement
442 context. Importantly, because direct exposure to the CSs was held constant during extinction,
443 the increased vmPFC activity to the vicariously extinguished CS+ during the reinstatement
444 test is likely to reflect a socially transmitted safety signal beyond what was accomplished
445 through direct exposure only. It is noteworthy that the reinstatement test in our design
446 included a change of context (from extinction context to the original acquisition context),

447 suggesting that vicarious extinction learning results in a context-independent retrieval of
448 extinction memory. This finding is intriguing given that standard extinction procedures
449 typically yield a highly context-dependent decrease in CR that recovers when tested in a
450 context differed than the extinction context (Bouton, 2002). This drawback of standard
451 extinction procedures is parallel to relapse of anxiety in patients after an initially successful
452 exposure treatment. Overcoming this contextual dependency of exposure-based procedures
453 has been suggested to be one of the challenges in finding effective treatment protocols
454 (Vervliet et al., 2013). Taken together with our previous demonstration that vicarious
455 extinction learning enhanced safety memory retrieval by attenuating fear reinstatement
456 compared to a standard extinction procedure (Golkar et al., 2013), the finding that vicarious
457 extinction learning generalized to a new contexts may be of particular relevance for
458 understanding and treating the persistence of fear memories in individuals suffering from
459 emotional disorders. Notwithstanding, an important step in approaching the clinical utility of
460 vicarious extinction learning is to examine its long-term effects on acquired fear memory,
461 optimally after allowing for consolidation of both the acquisition and extinction memory
462 separately (Haaker et al., 2014).
463
464 Noteworthy, we did not find a relationship between empathy and the effects of vicarious
465 extinction, neither did we observe any additional brain regions linked to the processing of
466 social-affective information, such as the anterior cingulate cortex (ACC), the anterior insula
467 (Lamm et al., 2011), and the dorsal medial prefrontal cortex, dmPFC (Zaki and Ochsner,
468 2012). For example, previous research has implicated the ACC and the anterior insula in the
469 processing of social pain (Eisenberger, 2012) and inactivation of the ACC has been shown to
470 retard vicarious fear learning in mice (Jeon et al., 2010). Interestingly, empathetic appraisals
471 has been shown to enhance vicarious fear learning in humans (Olsson et al., 2015), and both

472 animal (Jeon et al., 2010) and human work (Golkar et al., 2015) indicate that vicarious
473 learning is augmented when learning from an in-group compared to an out-group
474 demonstrator, perhaps reflecting a general tendency to display greater empathic and otherwise
475 pro-social responses to in-group, as compared to out-group, individuals (Xu et al., 2009; Hein
476 et al., 2010). In the present study, the lack of the involvement of brain areas previously linked
477 to processing of social information might be due to the nature of our paradigm, as well as the
478 analytic strategy. Whereas studies describing the involvement of the medial PFC have
479 contained explicit instructions to form impressions of the target's mental states (Ochsner et
480 al., 2004; Amodio and Frith, 2006), our paradigm contained no such instructions. Moreover,
481 unlike studies of empathic processes (Lamm, Decety, & Singer, 2011), the statistical contrasts
482 in our paradigm were optimized to capture the underlying associative processes and therefore
483 we analyzed responses that were predictive of (i.e. *preceded*) the onset of the response to
484 shock (or no shock) to the same individual demonstrator. These factors might have
485 contributed to the lack of significant differences in self-reported social-cognitive measures
486 when contrasting the $CS+_{vic\ safety}$ and $CS+_{vic\ reinf}$ conditions. Critically however, the effects of
487 vicariously learned safety were observed at the reinstatement test stage, establishing the
488 effects of vicarious extinction learning in the absence of the demonstrator. This finding also
489 suggests that the demonstrator is not merely acting as a conditioned inhibitor that predicts the
490 absence of the US, because removing such safety signals typically augments the return of
491 conditioned fear responses (e.g. Craske et al., 2008). Futures studies should also investigate
492 whether similar effects are obtained using a mixed gender population.
493
494 Taken together, vicarious extinction learning prevented the return of conditioned fear
495 responses during reinstatement testing in a new context. This effect was accompanied by
496 enhanced vmPFC activity and functional connectivity with the amygdala/anterior

hippocampus compared to a vicariously reinforced CS+. During extinction, vicarious extinction learning was associated with a decreased engagement of the vmPFC-amygdala circuitry, suggesting that vicarious extinction may reduce the necessity for PFC-mediated inhibition during learning that is typically observed in traditional extinction procedures (e.g. Phelps et al., 2004). Collectively, these patterns of activity are in line with an integrative role of the vmpFC (Roy et al., 2012), in which the vmPFC and its connectivity with subcortical regions represent conceptual information relevant for determining the current cue value. We hope that our novel experimental model will serve to inspire research to further specify the mechanisms underlying vicarious extinction learning, and their applicability in overcoming the return of fear that accompanies traditional exposure-based treatments for anxiety disorders.

REFERENCES

Agren T, Engman J, Frick A, Bjorkstrand J, Larsson EM, Furmark T, Fredrikson M (2012) Disruption of reconsolidation erases a fear memory trace in the human amygdala. Science 337:1550-1552.

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nature reviews Neuroscience 7:268-277.

Ashburner J (2007) A fast diffeomorphic image registration algorithm. NeuroImage 38:95-113.

Bandura A (1977) Social learning theory. NJ: Prentice-Hall: Englewood Cliffs.

Boucsein W, Fowles DC, Grimnes S, Ben-Shakhar G, roth WT, Dawson ME, Filion DL (2012) Publication recommendations for electrodermal measurements. Psychophysiology 49:1017-1034.

Bouton ME (2002) Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. Biological psychiatry 52:976-986.

Craske MG, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A (2008) Optimizing inhibitory learning during exposure therapy. Behaviour Research and Therapy 46:5-27.

Eisenberger NI (2012) The neural bases of social pain: evidence for shared representations with physical pain. Psychosomatic medicine 74:126-135.

529 Garfinkel SN, Abelson JL, King AP, Sripada RK, Wang X, Gaines LM, Liberzon I (2014) Impaired
530 contextual modulation of memories in PTSD: an fMRI and psychophysiological study
531 of extinction retention and fear renewal. J Neurosci 34:13435-13443.
532 Golkar A, Castro VB, Olsson A (2015) Social learning of fear and safety is determined by the
533 demonstartor's racial group. Accepted in Biology letters.
534 Golkar A, Selbing I, Flygare O, Öhman A, Olsson A (2013) Other People as Means to a Safe
535 End: Vicarious Extinction Blocks the Return of Learned Fear. Psychological science
536 24:2182-2190.
537 Haaker J, Golkar A, Hermans D, Lonsdorf TB (2014) A review on human reinstatement
538 studies: an overview and methodological challenges. Learning & Memory 21:424-
539 440.
540 Haaker J, Gaburro S, Sah A, Gartmann N, Lonsdorf TB, Meier K, Singewald N, Pape HC,
541 Morellini F, Kalisch R (2013) Single dose of L-dopa makes extinction memories
542 context-independent and prevents the return of fear. Proc Natl Acad Sci U S A
543 110:E2428-2436.
544 Hartley CA, Casey BJ (2013) Risk for anxiety and implications for treatment: developmental,
545 environmental, and genetic factors governing fear regulation. Ann N Y Acad Sci
546 1304:1-13.
547 Hein G, Silani G, Preuschoff K, Batson CD, Singer T (2010) Neural Responses to Ingroup and
548 Outgroup Members' Suffering Predict Individual Differences in Costly Helping.
549 Neuron 68:149-160.
550 Jeon D, Kim S, Chetana M, Jo D, Ruley HE, Lin SY, Rabah D, Kinet JP, Shin HS (2010)
551 Observational fear learning involves affective pain system and Ca(v)1.2 Ca2+
552 channels in ACC. Nature neuroscience 13:482-U105.
553 Kalisch R, Korenfeld E, Stephan KE, Weiskopf N, Seymour B, Dolan RJ (2006) Context-
554 dependent human extinction memory is mediated by a ventromedial prefrontal and
555 hippocampal network. Journal of Neuroscience 26:9503-9511.
556 Kim JH, Richardson R (2010) New findings on extinction of conditioned fear early in
557 development: theoretical and clinical implications. Biological psychiatry 67:297-303.
558 LaBar KS, Gatenby JC, Gore JC, LeDoux JE, Phelps EA (1998) Human amygdala activation
559 during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron
560 20:937-945.
561 Laland KN (2004) Social learning strategies. Learning & Behavior 32:4-14.
562 Lamm C, Decety J, Singer T (2011) Meta-analytic evidence for common and distinct neural
563 networks associated with directly experienced pain and empathy for pain.
564 NeuroImage 54:2492-2502.
565 Lonsdorf TB, Haaker J, Fadai T, Kalisch R (2014) No evidence for enhanced extinction
566 memory consolidation through noradrenergic reuptake inhibition-delayed memory
567 test and reinstatement in human fMRI. Psychopharmacology 231:1949-1962.
568 Lundqvist D, Flykt, A., & Öhman, A. (1998) The Karolinska Directed Emotional Faces - KDEF.
569 In: CD ROM from Department of Clinical Neuroscience, Psychology section,
570 Karolinska Institutet, ISBN 91-630-7164-9.
571 Maren S, Phan KL, Liberzon I (2013) The contextual brain: implications for fear conditioning,
572 extinction and psychopathology. Nature Reviews Neuroscience 14:417-428.
573 Milad MR, Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear
574 extinction. Nature 420:70-74.

Milad MR, Wright CI, Orr SP, Pitman RK, Quirk GJ, Rauch SL (2007) Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. Biological psychiatry 62:446-454.

Milad MR, Pitman RK, Ellis CB, Gold AL, Shin LM, Lasko NB, Zeidan MA, Handwerger K, Orr SP, Rauch SL (2009) Neurobiological Basis of Failure to Recall Extinction Memory in Posttraumatic Stress Disorder. Biological psychiatry 66:1075-1082.

Ochsner KN, Knierim K, Ludlow DH, Hanelin J, Ramachandran T, Glover G, Mackey SC (2004) Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. Journal of cognitive neuroscience 16:1746-1772.

Olsson A, Nearing KI, Phelps EA (2007) Learning fears by observing others: the neural systems of social fear transmission. Social Cognitive and Affective Neuroscience 2:3-11.

Olsson A, McMahon K, Papenberg G, Zaki J, Bolger N, Ochsner KN (2015) Vicarious Fear Learning Depends on Empathic Appraisals and Trait Empathy. Psychological science.

Phelps EA, Delgado MR, Nearing KI, LeDoux JE (2004) Extinction learning in humans: Role of the amygdala and vmPFC. Neuron 43:897-905.

Poldrack RA (2006) Can cognitive processes be inferred from neuroimaging data? Trends in cognitive sciences 10:59-63.

Rabinak CA, Angstadt M, Sripada CS, Abelson JL, Liberzon I, Milad MR, Phan KL (2013) Cannabinoid facilitation of fear extinction memory recall in humans. Neuropharmacology 64:396-402.

Roy M, Shohamy D, Wager TD (2012) Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends in cognitive sciences 16:147-156.

Schiller D, Levy I, Niv Y, LeDoux JE, Phelps EA (2008) From Fear to Safety and Back: Reversal of Fear in the Human Brain. Journal of Neuroscience 28:11517-11525.

Schiller D, Kanen JW, LeDoux JE, Monfils MH, Phelps EA (2013) Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. Proc Natl Acad Sci U S A 110:20040-20045.

Schiller D, Monfils MH, Raio CM, Johnson DC, LeDoux JE, Phelps EA (2010) Preventing the return of fear in humans using reconsolidation update mechanisms. Nature 463:49-U51.

Spoormaker VI, Sturm A, Andrade KC, Schroter MS, Goya-Maldonado R, Holsboer F, Wetter TC, Samann PC, Czisch M (2010) The neural correlates and temporal sequence of the relationship between shock exposure, disturbed sleep and impaired consolidation of fear extinction. Journal of Psychiatric Research 44:1121-1128.

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

Vervliet B, Craske MG, Hermans D (2013) Fear extinction and relapse: state of the art. Annu Rev Clin Psychol 9:215-248.

Xu XJ, Zuo XY, Wang XY, Han SH (2009) Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses. Journal of Neuroscience 29:8525-8529.

Zaki J, Ochsner KN (2012) The neuroscience of empathy: progress, pitfalls and promise. Nature neuroscience 15:675-680.

622