Social Cognitive and Affective Neuroscience, 2016, 1541–1549

doi: 10.1093/scan/nsw068
Advance Access Publication Date: 7 June 2016
Original article

Neural signals of vicarious extinction learning

Armita Golkar,[1,2] Jan Haaker,[1] Ida Selbing,[1] and Andreas Olsson[1]

[1]Department of Clinical Neuroscience, Karolinska Institutet, Nobels Väg 11, Stockholm 17177, Sweden,
[2]Department of Psychology, University of Amsterdam

Correspondence should be addressed to Armita Golkar, Department of Clinical Neuroscience, Karolinska Institutet, Nobels väg 11, 17177, Stockholm, Sweden. E-mail: Armita.golkar@ki.se

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 ($\mathrm{CS}+_{\mathrm{vic\ safety}}$) or aversive reinforcement ($\mathrm{CS}+_{\mathrm{vic\ reinf}}$). During extinction, ventromedial prefrontal cortex (vmPFC) responses to the $\mathrm{CS}+_{\mathrm{vic\ reinf}}$ increased but decreased to the $\mathrm{CS}+_{\mathrm{vic\ safety}}$. 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 $\mathrm{CS}+_{\mathrm{vic\ reinf}}$ 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 influences the persistence of fear memories in individuals suffering from emotional disorders.

Key words: 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 behaviour (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

Received: 31 August 2015; **Revised:** 6 March 2106; **Accepted:** 13 May 2016

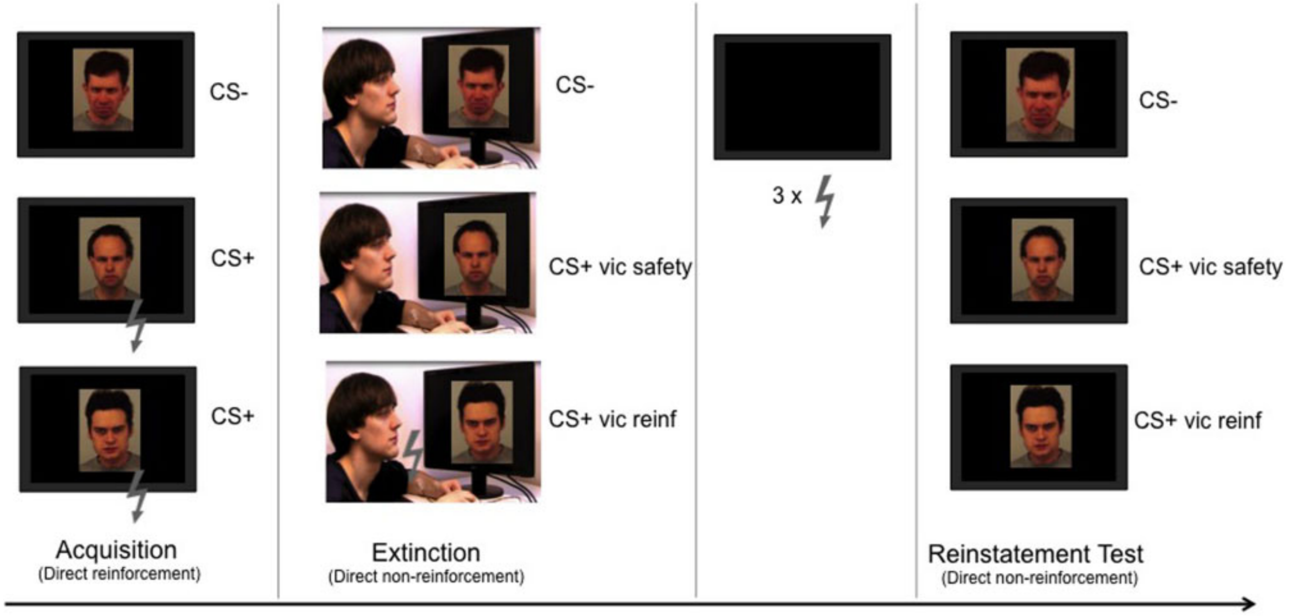

Fig. 1. Experimental design. The experiment was divided into different stages. Within each stage, all CSs were presented eight times each in a pseudorandomized order. During acquisition, two angry faces (CS + s) were repeatedly paired with a mild electric shock (US) given to the participants' wrist (six 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 + $_{\text{vic safety}}$), but received shocks on the presentations of the other CS+ (CS + $_{\text{vic reinf}}$, six 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.

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 unsignalled presentations of the shock (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+$_{\text{vic reinf}}$), 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+$_{\text{vic safety}}$) 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+$_{\text{vic safety}}$ *vs* the CS+$_{\text{vic reinf}}$, 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).

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 participants who were free from self-reported life-time psychiatric or neurological disease and medication. We used the stopping rule that all participants had to complete the functional magnetic resonance (fMRI) task and the questionnaires. Prior to analysis, we excluded two participants because they failed to report the contingency between the CSs and the unconditioned stimulus (US) and one participant with abnormal brain anatomy leaving a final sample of 20 participants with a mean age of 25 years (s.d. = 1.25). All participants gave written informed consent and were paid 350 SEK (∼50 USD) for their participation. Due to technical problems, the skin conductance data were missing for one participant, who was excluded from all statistical analyses of the skin conductance data.

Stimuli

Three pictures depicting angry male faces from the Karolinska Directed Emotional Faces database (Lundqvist et al., 1998) served as CSs (Items AM02ANS; AM04ANS; AM06ANS). During each stage of the experiment, each CS was presented eight times, with a duration of 6 s. The inter-trial interval between each CS was jittered between 11 and 15 s. The US consisted of a 100 ms DC-pulse electric stimulation applied to the participant's right wrist. The coupling between a specific conditioned face stimulus and the US, and the order of presentation of the two CS + s (CSs that were coupled to the US) was counterbalanced between participants. For the extinction stage, we created two movies (counterbalancing the order of the CS+ presentations) using Adobe Premiere Pro CS5.5 that was each 4 min and 18 s in length. The movies showed the demonstrator sitting in front of a computer screen watching the CS presentations. Which face that served as CS+$_{\text{vic safety}}$ and the CS+$_{\text{vic reinf}}$ was counterbalanced between participants. A shock electrode was visibly attached to the demonstrator's right wrist. Apart from the order of CS+ presentations, the movies were identical in terms of content and timing.

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 extinction stage that followed, participants watched a movie (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+$_{\text{vic safety}}$), but received shocks on 75% of the presentations of the other CS+ (CS+$_{\text{vic reinf}}$). The model reacted to the shocks by slightly twitching the arm and blinking. After the end of the 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 unsignalled 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 used to model clinical relapse of anxiety symtoms. During the reinstatement procedure, participants were exposed to a black screen for 30 s, 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

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 micro-Siemens (µS)] in the 0.5 to 4.5 s window following stimulus onset. Responses below .02 µS were scored as zero, and data were 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 eight-channel head coil. Each functional image volume comprised 46 continuous axial slices (2.3 mm thick, no gap) that were acquired using a T2*-sensitive gradient echo-planar imaging sequence (repetition time: 3000 ms; echo time: 31 ms; flip angle: 85°; field of view: 96 × 96 mm, 3 × 3 mm in-plane resolution). To account for T1 equilibrium effects, the first five volumes of each time series were discarded. High-resolution T1-weighted structural images (1 × 1 × 1 mm) were acquired after the experimental session. Pre-processing using Statistical parametric mapping [SPM8 (www.fil.ion.ucl.ac.uk/spm)] running on Matlab2013b (The MathWorks, Natick, MA)] involved realignment, unwarping co-registration and normalization to a sample-specific template, using DARTEL (Ashburner, 2007). Normalized data series were spatially smoothed with a 6 mm FWHM isotropic Gaussian kernel and manually inspected for excessive head movement. Further processing included temporal high-pass filtering (cut-off 128 s) and correction for temporal auto-correlations using first-order autoregressive modelling.

Regions of interest selection

The pre-defined regions of interest (ROIs) included two key structures of fear and safety memory processing in humans: the amygdala (LaBar *et al*., 1998; Phelps *et al*., 2004) and the vmPFC (Phelps *et al*., 2004; Kalisch *et al*., 2006; Milad *et al*., 2007). The amygdala ROI was defined as an anatomical mask derived from the automatic anatomical labelling atlas (Tzourio-Mazoyer *et al*., 2002). The vmPFC ROI was defined as a box (20 × 16 × 16 mm) around the average peak coordinate [xyz (MNI) = 0.41–12] of previous human fMRI studies (Phelps *et al*., 2004; Kalisch *et al*., 2006; Milad *et al*., 2007; Spoormaker *et al*., 2010; Haaker *et al*., 2013; Rabinak *et al*., 2013; Lonsdorf *et al*., 2014) testing for extinction recall.

Statistical analyses

For the SCR data, each stage of the experiment (Acquisition, Extinction and Reinstatement test) was analyzed with separate repeated measures analysis of variance (ANOVA). For the fMRI data, a general linear model 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+$_{\text{vic safety to-be}}$, CS + $_{\text{vic shock to-be}}$, CS − ; Extinction: CS+$_{\text{vic safety}}$, CS+$_{\text{vic reinf}}$, CS − ; Reinstatement: CS + $_{\text{previously vic safety}}$, CS+$_{\text{previously vic shock}}$, 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$_{\text{vic safety to-be}}$, US$_{\text{vic shock to-be}}$). During extinction, we modelled the vicarious US (administered to the model) to CS + $_{\text{vic shock}}$ and the omission of each shock to the CS+$_{\text{vic safety}}$. In addition, two nuisance regressors were included to factor out experimental effects of no interest: one regressor modelled 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

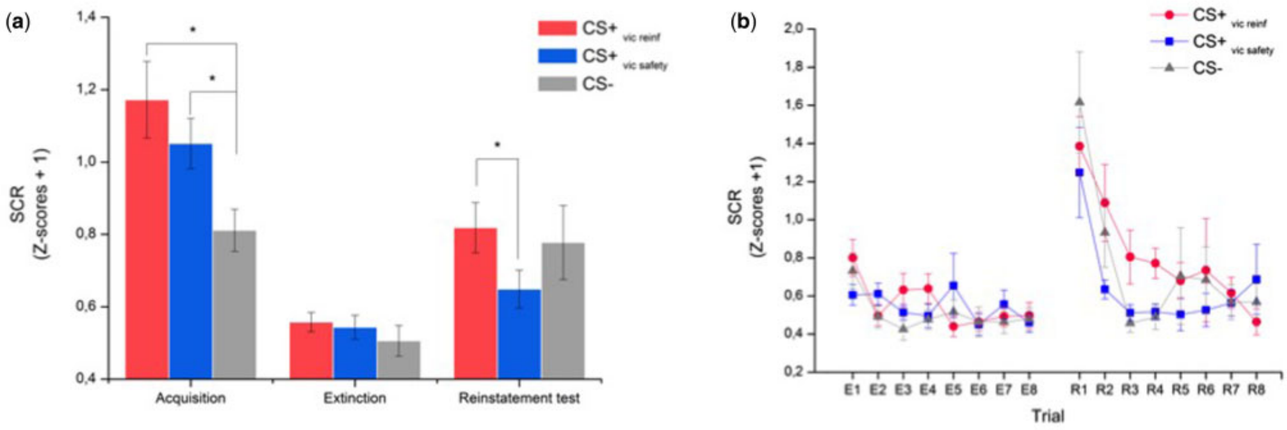

Fig. 2. (a) Mean SCR as a function of experimental stage and conditioned stimulus (CS). (b) Trial-by-trial data for extinction (E1-E) and reinstatement (R1–R8). Note that in order to capture the immediate response to a new context (i.e. the orienting response), the first CS presentation during the reinstatement test was always a CS−. Error bars indicate SEM. *Statistically significant differences ($P < 0.05$).

comparisons between the CS+$_{\text{vic safety}}$ and the CS+$_{\text{vic reinf}}$ for the effect of vicarious extinction. Separate analyses for each session included beta-estimates for each CS (one factor, three levels), derived from individual single subjects general linear modeling. We also included a comparison for the effects of reinstatement between extinction and reinstatement (two factors with three 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 ($P < 0.001$) and a liberal threshold of $k > 5$ voxel are reported for each analysis in Supplementary Table S1. To examine condition-specific functional connectivity during extinction and reinstatement testing, each participant's BOLD signal time-course at the individual peak within the vmPFC ROI (for extinction: from the CS+$_{\text{vic reinf}}$ > CS+$_{\text{vic safety}}$ contrast; for reinstatement: from the CS+$_{\text{vic safety}}$ > CS+$_{\text{vic reinf}}$, thresholded at $P < 005$ uncorrected) was extracted as an eigenvariate. The time course was deconvolved and multiplied with the condition specific onset (e.g. onset of the CS+$_{\text{vic reinf}}$ or CS+$_{\text{vic safety}}$ 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 (Supplementary information). Parameter estimates for each CS condition were then contrasted using one-sample t-test (for extinction: CS+$_{\text{vic reinf}}$ > CS+$_{\text{vic safety}}$; for reinstatement: CS+$_{\text{vic safety}}$ > CS+ $_{\text{vic reinf}}$).

Results

Subjective ratings

On a scale from 1 (not at all) to 7 (very much), participants rated how much discomfort they thought that the model experienced when receiving shocks ($M = 2.28$; s.d. = 1.13), how much discomfort they thought that the model experienced when receiving shocks ($M = 3.5$; s.d. = 1.30), how much they could identify themselves with the person in the movie ($M = 4.17$; s.d. = 1.15), how much empathy they felt for the person in the movie ($M = 3.17$; s.d. = 1.54) and how much they liked the person in the movie on a scale from −3 (disliked) to 3 (liked) ($M = 0.44$; s.d. = 0.07). None of these ratings was significantly related to the extinction or reinstatement data as assessed with correlation analysis (all $P < 0.05$).

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 ($F(2,36) = 13.27$, $P < 0001$; $\eta^2 = .42$), showing that mean SCRs to both CS+s were larger than to the CS− [CS+$_{\text{vic safety to-be:}}$ $t(18) = 4.20$, $P = 0.001$, 95% confidence interval (CI) for the difference between conditions = (0.12–0.36); (CS+$_{\text{vic reinf to-be:}}$ $t(18) = 5.05$, $P < 0.001$; 95% CI = (0.21–0.51)], and that the CS+s did not differ from each other [$t(18) = 1.45$, $P = 0.16$; 95% CI = (−0.05 to 0.30)]. The conditioned fear responses diminished during extinction (no main effect of stimulus: $F(2,36) = 1.39$, $P = 0.26$ and a significant main effect of Trial: $F(7,126) = 3.60$, $P = 0.002$; $\eta^2 = .17$), 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) × time (2) ANOVA that resulted in a significant interaction ($F(2,36) = 3.70$, $P = 0.03$; $\eta^2 = .17$). Follow-up t-tests revealed a marginal increase in SCR to the CS+$_{\text{vic safety}}$ [$t(18) = 2.10$, $P = 0.05$, 95% CI = (−2.11, 0.00)]; CS+$_{\text{vic reinf}}$ [$t(18) = 3.58$, $P = 0.002$, 95% CI = (−0.41, −0.11); CS− $t(18) = 2.90$, $P = 0.009$, 95% CI = (−0.47, −0.08)], and planned comparisons revealed that mean SCRs during the reinstatement test were significantly lower to the CS+$_{\text{vic safety}}$ than to the CS+$_{\text{vic reinf}}$ [$t(18) = 2.58$; $P = 0.019$, 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

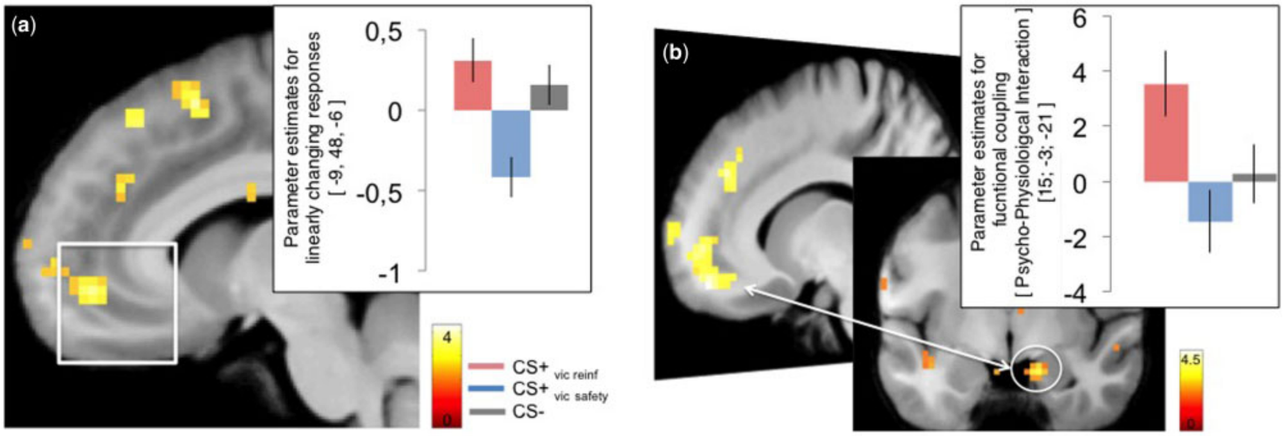
(a)
0,5
0
-0,5
-1
Parameter estimates for linearly changing responses [-9, 48, -6]
4
0
CS+ vic reinf
CS+ vic safety
CS-
(b)
6
4
2
0
-2
-4
Parameter estimates for fucntional coupling [Psycho-Physiologcal Interaction] [15; -3; -21]
4.5
0

Fig. 3. (a) Activity in the vmPFC ROI increased linearly to CS + $_{\text{vic reinf}}$ during extinction learning whereas responses in this region decreased to the CS + $_{\text{vic safety}}$. (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 + $_{\text{vic rein}}$ *vs* the CS + $_{\text{vic safety}}$. *T*-values are superimposed on a normalized average structural image. fMRI display threshold: $P < 0.005$, uncorrected for illustrative purposes. Error bars represent SEM.

($F(2,36) = 4.22$, $P = 0.02$, $\eta^2 = .19$) explained by a significantly higher SCR to the CS+$_{\text{vic reinf}}$ *vs* the CS − [$t(18) = 2.11$, $P = 0.049$, 95% CI = (0.00–0.377)], and no differences between the CS+$_{\text{vic safety}}$ *vs* the CS − [$t(18) = 0.36$, $P = 0.72$, 95% CI = (−0.134 to 0.095)].

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+$_{\text{vic safety}}$ and the CS+$_{\text{vic reinf}}$ 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 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, $T = 3.11$, $Z = 2.98$; $P(\text{SVC}) = 0.047$, $P(\text{uncorrected}) < 0.001$] (Supplementary Table S2). 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+$_{\text{vic reinf}}$ > CS+$_{\text{vic safety}}$ [*x*, *y*, *z* (MNI) = 9, 45, −9, $Z = 3.21$, $P(\text{SVC}) = 0.064$, $P(\text{uncorrected}) < 0.001$]. To further characterize this difference, we ran a separate model including parametric regressors for each CS modelling linearly increasing (trial-by-trial) responses (Figure 3a). This model revealed that activity within the vmPFC ROI increased to the CS+$_{\text{vic reinf}}$, whereas responses to the CS+$_{\text{vic safety}}$ decreased [*x*, *y*, *z* (MNI) = −9, 48, −6, $Z = 3.75$, $P(\text{SVC}) = 0.01$, $P(\text{uncorrected}) < 0.001$], 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+$_{\text{vic reinf}}$ > CS+$_{\text{vic safety}}$ ($P < 0.05$). We found that the connectivity between the vmPFC and a region located in the lateral amygdala and the anterior hippocampus (Figure 3b, Table 1) was more positive for the CS+ $_{\text{vic reinf}}$ *vs* the CS+ $_{\text{vic safety}}$.

Table 1. Whole brain results of the PPI [$P(\text{uncorr}) < 0.001$ and cluster-size (k) ≥ 5]

Contrast/region	T	Z	Coordinates
CS + $_{\text{vic reinf}}$ > CS + $_{\text{vic safety}}$			
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 + $_{\text{vic safety}}$ > CS + $_{\text{vic reinf}}$			
No cluster above threshold			

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+$_{\text{vic reinf}}$ with the CS − . This analysis revealed a greater activity to the CS+$_{\text{vic reinf}}$ *vs* the CS − in the left amygdala [*x*, *y*, *z* (MNI) = −24, 0, −21, $Z = 2.98$, $P(\text{SVC}) = 0.04$, $P(\text{uncorrected}) < 0.001$], but no difference between the CS+$_{\text{vic safety}}$ and the CS − [no voxel above $P(\text{uncorrected}) < 0.01$ in the ROI], demonstrating that vicarious extinction blocked the reinstatement of defensive responses. In fact, amygdala responses to the CS+$_{\text{vic safety}}$ was intermediate between the amygdala responses to the CS − and to the CS+$_{\text{vic reinf}}$, and there were no significant difference between the amygdala response between the CS + s [no voxel above $P(\text{uncorrected}) < 0.01$ in the ROI], (Figure 4a). Additionally, we observed that the difference in reinstated amygdala response in the CS+$_{\text{vic reinf}}$ > CS+$_{\text{vic safety}}$ contrast was positively correlated with reinstated SCRs (CS+$_{\text{vic reinf}}$ > CS+$_{\text{vic safety}}$) during the

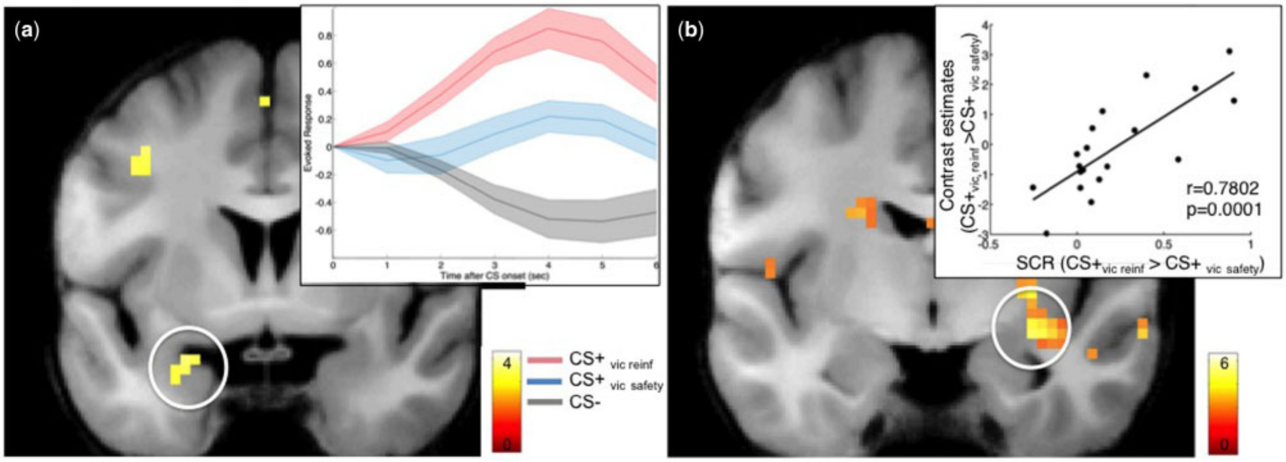

Fig. 4. (a) Mean estimated evoked responses (arbitrary units) within the amygdala during reinstatement testing scaled to the response at onset of each CS. Shaded areas represent SEM. (b) Correlation between amygdala activity during Reinstatement test in the contrast $CS+_{\text{vic reinf}} > CS+_{\text{vic safety}}$ and the reinstated SCRs for $CS+_{\text{vic reinf}} > CS+_{\text{vic safety}}$. *T*-values are superimposed on a normalized average structural image. fMRI display threshold: $P < 0.005$, uncorrected for illustrative purposes.

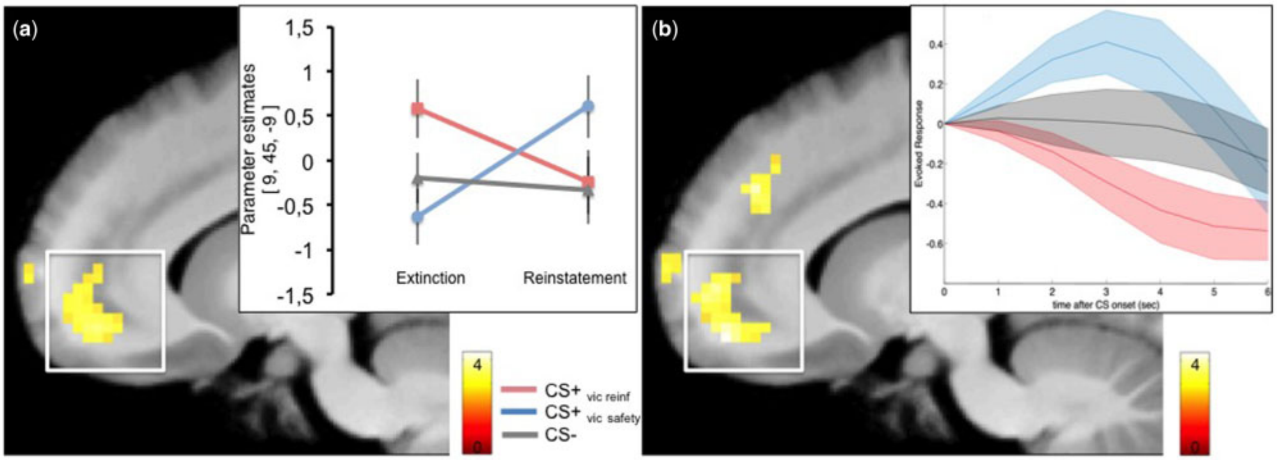

Fig. 5. (a) Change in vmPFC activity from extinction to reinstatement testing displayed for all CSs separately (b). Mean estimated evoked responses (arbitrary units) within the vmPFC during Reinstatement test scaled to the response at onset of each CS. Shaded areas represent SEM. *T*-values are superimposed on a normalized average structural image. fMRI display threshold: $P < 0.005$, uncorrected for illustrative purposes.

reinstatement test ($r = 0.78$, $P < 0.001$; 30, −9, −12; $t = 5.01$, $P(\text{FWE}) = 0.005$) (Figure 4b).

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 $CS+_{\text{vic safety}}$ as compared to the $CS+_{\text{vic reinf}}$ [*x*, *y*, *z* (MNI) = 9, 45, −9, $Z = 3.73$, $P(\text{SVC}) = 0.011$, $P(\text{uncorrected}) < 0.001$], Figure 5a. As predicted, mean activity within the vmPFC during reinstatement was larger to the $CS+_{\text{vic safety}}$ as compared to the $CS+_{\text{vic reinf}}$ [*x*, *y*, *z* (MNI) = −9, 48, −15, $Z = 3.99$, $P(\text{SVC}) = 0.004$, $P(\text{uncorrected}) < 0.001$], but did not differ from the CS− [$P(\text{SVC}) = 0.236$, $Z = 2.51$]. The individual average peak responses within the vmPFC ROI were enhanced to the $CS+_{\text{vic safety}}$, intermediate to the CS− and decreased to the $CS+_{\text{vic reinf}}$ (Figure 5b), echoing the reversed ordinal pattern in the SCR data.

Conditions-specific functional connectivity during reinstatement

Finally, we examined the condition-specific functional connectivity during reinstatement using the vmPFC ROI (from the $CS+_{\text{vic safety}} > CS+_{\text{vic reinf}}$ contrast; $P < 0.05$) as seed. We found that connectivity between the vmPFC and the anterior hippocampus, as well as the inferior temporal gyrus (Table 2) was more positive for the CS+ vic safety vs the CS+ vic reinf. Whole brain results for all stages of the experiment are reported in Supplementary Table S2.

Discussion

Our study demonstrates that socially transmitted safety information prevent previously learned fear responses from recovering, reaffirming the efficiency of vicarious safety learning

Table 2. Whole brain results of the PPI [*P*(uncorr) < 0.001 and cluster-size (k) ≥ 5] during reinstatement

Contrast/region	T	Z	Coordinates
CS $+_{\text{vic safety}} >$ CS $+_{\text{vic reinf}}$			
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 $+_{\text{vic reinf}} >$ CS $+_{\text{vic safety}}$			
No cluster above threshold			

Coordinates are given in MNI space.

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$+_{\text{vic reinf}}$). 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$+_{\text{vic reinf}}$), 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$+_{\text{vic safety}}$) 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$+_{\text{vic reinf}}$ 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$+_{\text{vic safety}}$ 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 reduced activity, and less functional connectivity with region located within the lateral amygdala, in response to the CS$+_{\text{vic safety}}$ compared to the CS$+_{\text{vic reinf}}$. Conversely, during the reinstatement test, activity within the vmPFC increased and showed stronger coupling with the anterior hippocampus, in response to CS$+_{\text{vic safety}}$ compared to the CS$+_{\text{vic reinf}}$. This strengthened coupling between the vmPFC and the amygdala/hippocampus is in line with the proposed role of this network in gating successful recall of context-dependent extinction memory (Kalisch *et al.*, 2006; Milad *et al.*, 2007). Moreover, the diminished vmPFC activity to the CS$+_{\text{vic reinf}}$ during reinstatement is in line with previous studies demonstrating diminished vmPFC activity in post-traumatic stress disorder patient during extinction recall failure (Milad *et al.*, 2009; Garfinkel *et al.*, 2014), and might provide a route through which previously learned fears can be maintained through social reinforcement.

On a more general level, our finding of increased activity in the vmPFC in response to the vicariously extinguished CS+ during reinstatement is consistent with a suggested role of the vmPFC in integrating information from distributed brain regions involved in signaling affective value, episodic memory and social cognition (Roy *et al.*, 2012), and using this information to provide a selective safety signal that indicates which stimuli are safe to ignore (Schiller *et al.*, 2008). In the present study, the vmPFC appears to track the relative cue value, by responding more to the relatively more dangerous cue (CS$+_{\text{vic reinf}}$ *vs* the CS$+_{\text{vic safety}}$) when presented in the safe, extinction, context and conversely, shift to responding more to the relatively safe cue (CS$+_{\text{vic safety}}$ *vs* CS$+_{\text{vic reinf}}$) presented in the dangerous, reinstatement context. Importantly, because direct exposure to the CSs was held constant during extinction, the increased vmPFC activity to the vicariously extinguished CS+ during the reinstatement test is likely to reflect a socially transmitted safety signal beyond what was accomplished through direct exposure only. It is noteworthy that the reinstatement test in our design included a change of context (from extinction context to the original acquisition context), suggesting that vicarious extinction learning results in a context-independent retrieval of extinction memory. This finding is intriguing given that standard extinction procedures typically yield a highly context-dependent decrease in CR that recovers when tested in a context differed than the extinction context (Bouton, 2002). This drawback of standard extinction procedures is parallel to relapse of anxiety in patients after an initially successful exposure treatment. Overcoming this contextual dependency of exposure-based procedures has been suggested to be one of the challenges in finding effective treatment protocols (Vervliet *et al.*, 2013). Taken together with our previous demonstration that vicarious extinction learning enhanced safety memory retrieval by attenuating fear reinstatement compared to a standard extinction procedure (Golkar *et al.*, 2013), the finding that vicarious extinction learning generalized to a new contexts may be of particular relevance for understanding and treating the persistence of fear memories in individuals suffering from emotional disorders. Notwithstanding, an important step in approaching the clinical utility of vicarious extinction learning is to examine its long-term effects on acquired fear memory, optimally after allowing for consolidation of both the acquisition and the extinction memory separately (Haaker *et al.*, 2014).

Noteworthy, we did not find a relationship between empathy and the effects of vicarious extinction, neither did we observe any additional brain regions linked to the processing of social-affective information, such as the anterior cingulate cortex (ACC), the anterior insula (Lamm *et al.*, 2011) and the dorsal medial prefrontal cortex, dmPFC (Zaki and Ochsner, 2012). For example, previous research has implicated the ACC and the anterior insula in the processing of social pain (Eisenberger, 2012) and inactivation of the ACC has been shown to retard vicarious fear learning in mice (Jeon *et al.*, 2010). Interestingly, empathetic appraisals has been shown to enhance vicarious fear learning in humans (Olsson *et al.*, 2016), and both animal (Jeon *et al.*, 2010) and human work (Golkar *et al.*, 2015) indicate that vicarious learning is augmented when learning from an in-group compared to an out-group demonstrator, perhaps reflecting a general tendency to display greater empathic and otherwise pro-social responses to in-group, as compared to out-group, individuals (Xu *et al.*, 2009; Hein *et al.*, 2010). In the present study, the lack of the involvement of brain areas previously linked to processing of social information might be due to the nature of our paradigm, as well as the analytic strategy. Whereas studies describing the involvement of the medial PFC have contained explicit instructions to form impressions of the target's mental states (Ochsner *et al.*, 2004; Amodio and Frith, 2006), our paradigm contained no such instructions. Moreover, unlike studies of empathic processes (Lamm *et al.*, 2011), the statistical contrasts in our paradigm were optimized to capture the underlying associative processes and therefore we analyzed responses that were predictive of (i.e. preceded) the onset of the response to shock (or no shock) to the same individual demonstrator. These factors might have contributed to the lack of significant differences in self-reported social-cognitive measures when contrasting the $CS+_{vic\ safety}$ and $CS+_{vic\ reinf}$ conditions. Critically, however, the effects of vicariously learned safety were observed at the reinstatement test stage, establishing the effects of vicarious extinction learning in the absence of the demonstrator. This finding also suggests that the demonstrator is not merely acting as a conditioned inhibitor that predicts the absence of the US because removing such safety signals typically augments the return of conditioned fear responses (e.g. Craske *et al.*, 2008). Future studies should also investigate whether similar effects are obtained using a mixed gender population.

Taken together, vicarious extinction learning prevented the return of conditioned fear responses during reinstatement testing in a new context. This effect was accompanied by 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.

Acknowledgements

This research was supported by a Riksbankens Jubileumsfond (Grant P11-1017) and an Independent Starting Grant (284366; Emotional Learning in Social Interaction) from the European Research Council to A. Olsson.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Agren, T., Engman, J., Frick, A., *et al.* (2012). Disruption of reconsolidation erases a fear memory trace in the human amygdala. *Science* **337**, 1550–2.

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience* **7**, 268–77.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage* **38**, 95–113.

Bandura, A. (1977) *Social Learning Theory*, Prentice-Hall, NJ: Englewood Cliffs.

Boucsein, W., Fowles, D.C., Grimnes, S., *et al.* (2012). Publication recommendations for electrodermal measurements. *Psychophysiology* **49**, 1017–34.

Bouton, M.E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry* **52**, 976–86.

Craske, M.G., 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, N.I. (2012). The neural bases of social pain: evidence for shared representations with physical pain. *Psychosomatic Medicine* **74**, 126–35.

Garfinkel, S.N., Abelson, J.L., King, A.P., *et al.* (2014). Impaired contextual modulation of memories in PTSD: an fMRI and psychophysiological study of extinction retention and fear renewal. *J Neurosci* **34**, 13435–43.

Golkar, A., Castro, V.B., Olsson, A. (2015) Social learning of fear and safety is determined by the demonstartor's racial group. *Biology Letters*, **11**(1), 20140817.

Golkar, A., Selbing, I., Flygare, O., Öhman, A., Olsson, A. (2013). Other People as Means to a Safe End: Vicarious Extinction Blocks the Return of Learned Fear. *Psychological Science* **24**, 2182–90.

Haaker, J., Golkar, A., Hermans, D., Lonsdorf, T.B. (2014). A review on human reinstatement studies: an overview and methodological challenges. *Learning & Memory* **21**, 424–40.

Haaker, J., Gaburro, S., Sah, A., *et al.* (2013). Single dose of L-dopa makes extinction memories context-independent and prevents the return of fear. *Proc Natl Acad Sci U S A* **110**, E2428–36.

Hartley, C.A., Casey, B.J. (2013). Risk for anxiety and implications for treatment: developmental, environmental, and genetic factors governing fear regulation. *Ann N Y Acad Sci* **1304**, 1–13.

Hein, G., Silani, G., Preuschoff, K., Batson, C.D., Singer, T. (2010). Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping. *Neuron* **68**, 149–60.

Jeon, D., Kim, S., Chetana, M., *et al.* (2010). Observational fear learning involves affective pain system and Ca(v)1.2 Ca2+ channels in ACC. *Nature Neuroscience* **13**, 482. U105.

Kalisch, R., Korenfeld, E., Stephan, K.E., Weiskopf, N., Seymour, B., Dolan, R.J. (2006). Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *Journal of Neuroscience* **26**, 9503–11.

Kim, J.H., Richardson, R. (2010). New findings on extinction of conditioned fear early in development: theoretical and clinical implications. *Biological Psychiatry* **67**, 297–303.

LaBar, K.S., Gatenby, J.C., Gore, J.C., LeDoux, J.E., Phelps, E.A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* **20**, 937–45.

Laland, K.N. (2004). Social learning strategies. *Learning & Behavior* **32**, 4–14.

Lamm, C., Decety, J., Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* **54**, 2492–502.

Lonsdorf, T.B., Haaker, J., Fadai, T., Kalisch, R. (2014). No evidence for enhanced extinction memory consolidation through noradrenergic reuptake inhibition-delayed memory test and reinstatement in human fMRI. *Psychopharmacology* **231**, 1949–62.

Lundqvist, D., Flykt, A., Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF. In: CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Maren, S., Phan, K.L., Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience* **14**, 417–28.

Milad, M.R., Quirk, G.J. (2002). Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* **420**, 70–4.

Milad, M.R., Wright, C.I., Orr, S.P., Pitman, R.K., Quirk, G.J., Rauch, S.L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry* **62**, 446–54.

Milad, M.R., Pitman, R.K., Ellis, C.B., *et al.* (2009). Neurobiological Basis of Failure to Recall Extinction Memory in Posttraumatic Stress Disorder. *Biological Psychiatry* **66**, 1075–82.

Ochsner, K.N., Knierim, K., Ludlow, D.H., *et al.* (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience* **16**, 1746–72.

Olsson, A., McMahon, K., Papenberg, G., Zaki, J., Bolger, N., Ochsner, K.N. (2016). Vicarious fear learning depends on empathic appraisals and trait empathy. *Psychological ScienceJanuary 2016* **27**, 25–33.

Olsson, A., Nearing, K.I., Phelps, E.A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience* **2**, 3–11.

Phelps, E.A., Delgado, M.R., Nearing, K.I., LeDoux, J.E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron* **43**, 897–905.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* **10**, 59–63.

Rabinak, C.A., Angstadt, M., Sripada, C.S., *et al.* (2013). Cannabinoid facilitation of fear extinction memory recall in humans. *Neuropharmacology* **64**, 396–402.

Roy, M., Shohamy, D., Wager, T.D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences* **16**, 147–56.

Schiller, D., Kanen, J.W., LeDoux, J.E., Monfils, M.H., Phelps, E.A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proc Natl Acad Sci U S A* **110**, 20040–5.

Schiller, D., Levy, I., Niv, Y., LeDoux, J.E., Phelps, E.A. (2008). From Fear to Safety and Back: Reversal of Fear in the Human Brain. *Journal of Neuroscience* **28**, 11517–25.

Schiller, D., Monfils, M.H., Raio, C.M., Johnson, D.C., LeDoux, J.E., Phelps, E.A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **463**, 49. U51.

Spoormaker, V.I., Sturm, A., Andrade, K.C., *et al.* (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**, 1121–8.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., *et al.* (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–89.

Vervliet, B., Craske, M.G., Hermans, D. (2013). Fear extinction and relapse: state of the art. *Annu Rev Clin Psychol* **9**, 215–48.

Xu, X.J., Zuo, X.Y., Wang, X.Y., Han, S.H. (2009). Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses. *Journal of Neuroscience* **29**, 8525–9.

Zaki, J., Ochsner, K.N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nature Neuroscience* **15**, 675–80.