OXFORD

Social Cognitive and Affective Neuroscience, 2016, 612–620

doi: 10.1093/scan/nsv142
Advance Access Publication Date: 25 November 2015
Original article

Neural correlates of and processes underlying generalized and differential return of fear

Robert Scharfenort and Tina B. Lonsdorf

Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Correspondence should be addressed to Robert Scharfenort, Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20251 Hamburg, Germany. E-mail: r.scharfenort@uke.de.

Abstract

Relapse represents a major limitation to long-term remission of psychopathology (anxiety, addiction). Relapse of anxiety can be modeled in the laboratory as return of fear (ROF) following un-signaled re-presentation of the aversive event (reinstatement, RI) after extinction. In humans, response enhancement to both the CS+ and CS− (generalized RI) or specifically to the CS+ (differential RI) has been described following RI. The (psychological) mechanisms and boundary conditions underlying these different RI qualities were investigated in 76 healthy participants using autonomic measures and functional magnetic resonance imaging. Our results suggest that both processes reflect distinct albeit intertwined (psychological) processes which are reflected in different neural activation patterns. Differential RI was linked to CS+ related hippocampal activation and CS− related disinhibition of the ventromedial prefrontal cortex (vmPFC). The latter likely contributes to robust generalized RI which was mirrored in thalamic and visual areas (as well as the bed nucleus of the striatum and inusula) possibly indicating generally facilitated salience processing. In addition, we also present data on experimental boundary conditions of RI (trial sequence effects, time stability). Taken together, this first comprehensive analysis of RI-induced ROF aids not only experimental research on ROF but also understanding of factors promoting clinical relapse and the role of the vmPFC.

Key words: reinstatement; fMRI; hippocampus; vmPFC; generalization; discrimination

Introduction

Despite the existence of effective psychological and pharmacological interventions for anxiety disorders, relapse following initial treatment success represents a major limitation to long-term remission. Relapse prevention by means of pharmacological or behavioral interventions has thus been a major focus of research during the past years (Kindt *et al.*, 2009; Schiller *et al.*, 2010; Haaker *et al.*, 2013; Fitzgerald *et al.*, 2014).

Relapse can be studied in the laboratory in classical conditioning paradigms through the induction of return of fear (ROF) following extinction training. During initial fear acquisition, one stimulus (CS+) is paired with an aversive event (US) whereas a second stimulus (CS−) is not. Consequently, after a number of CS–US pairings, the mere presentation of the CS+ is sufficient to elicit fearful responding (conditioned reaction, CR). During subsequent extinction, both CSs are presented without reinforcement by the US, leading to a gradual waning of the CR. Importantly, extinction does not erase fear memories, but is thought to generate competing and co-existing inhibitory extinction memories (Bouton, 2002; Myers and Davis, 2007). Consequently, insufficient expression of extinction memories at a later time promotes ROF, that is, clinical relapse, after successful extinction/exposure treatment (Bouton, 2002).

In the laboratory, ROF can be triggered by experimental manipulations (for an overview, in humans see Vervliet *et al.*, 2013b; in animals see Bouton, 2004) that include the mere passage of time (spontaneous recovery), the induction of contextual change (renewal) or exposure to un-signaled USs [reinstatement (RI)]. The RI phenomenon has been well characterized in rodents already decades ago (Bouton and Bolles, 1979; Bouton and King, 1983; Bouton, 1984) and has been implicated in relapse of anxiety as well as addiction in humans (e.g. Mantsch *et al.*, 2015). RI-induced ROF in humans however has only been studied more recently and mainly served as an

Received: 14 April 2015; **Revised:** 19 November 2015; **Accepted:** 20 November 2015

outcome measure for manipulations of extinction memory consolidation (for a review, see Haaker *et al*., 2014). Nonetheless the experimental boundary conditions have not been targeted yet, rendering comparability and interpretation of different findings problematic. Although some studies report a general response enhancement for both the CS+ and CS− following RI (generalized RI), others observe a rather specific response enhancement for the CS+ (differential RI; see Haaker *et al.* 2014, Table 3 for an overview). The underlying mechanisms behind these different qualities in RI remain however unexplored and cannot be resolved by consulting rodent literature. Although nearly all rodent studies (for an exception, see Dirikx *et al*., 2007) use single-cue paradigms which employ only a single predictor of the US (the CS+), human studies mostly use differential conditioning protocols, which also include a non-paired control CS (CS−). Consequently, single-cue paradigms cannot generate different qualities of RI. In addition, only differential designs allow controlling for the effect of processes such as orienting and sensitization because these processes affect the CS+ and the CS− in a similar vein. In contrast, genuine associative processes are not expected to affect both CSs similarly even though it has been discussed that generalized ROF may also result from associative learning to the CS− (Vervliet *et al*., 2013b).

To date, it remains unresolved whether generalized ROF following RI reflects association-based processes or if it is merely attributable to sensitization/orienting effects to uncertainty (Haaker *et al*., 2014). Whether ROF is differential or generalized is however of critical clinical importance, as the ability to discriminate safety from threat cues is negatively associated with pathological anxiety (Lissek *et al*., 2005; Duits *et al*., 2015) and predictive of resilient responding to life stress (Craske *et al*., 2013). Consequently, the ability to maintain discrimination under aversive circumstances might critically underlie long-term remission and/or resilience.

Consequently, neuro-scientific methods such as functional magnetic resonance imaging (fMRI) may shed light on the (psychological) process underlying different qualities of RI (generalized *vs* differential). Thereby, differential RI is expected to reflect remainders of the fear memory to the CS+ and thus an association-based process. Consequentially, differential RI is expected to be mirrored in activation of brain areas observed during fear conditioning (Fullana *et al*., 2015) such as the anterior insula (AI)/frontal operculum, the anterior cingulate cortex (ACC)/dorsomedial prefrontal area (dmPFC), hippocampus and the amygdala and reduced activation in fear-inhibitory areas such as the ventromedial prefrontal cortex (vmPFC). Generalized RI in turn is expected to (partly) reflect sensitization and elevated salience processing to any CS presentation following RI. Therefore, we expect activation of the thalamus and primary sensory (visual) areas, which are involved in the processing of salient stimuli. As a secondary exploratory research question we investigated experimental boundary conditions of RI.

Taken together, the current study uses psychophysiological measures and fMRI to both explore the psychological processes as well as boundary conditions of RI in humans.

Materials and methods

Participants

Eighty-four right-handed [assessed by the Edinburgh Handedness Inventory (Oldfield, 1971)] participants [41 female, mean age(s.d.): 25(3.5) years] with normal or corrected-to-normal vision were recruited from a pool of participants ($N = 390$) from an ongoing data collection based on their history of stressful life events (SLEs), as accessed via a modified version of the Life events checklist (Canli *et al*., 2006). The effects of SLEs on fear conditioning, extinction and ROF processes are beyond the scope of this study and will be published elsewhere (see also Supplementary Method S1 and Table S1 for details). Exclusion criteria included any known current or prior psychiatric or neurological disorders and the abuse of illegal drugs. Prior to enrollment, all participants provided written informed consent to the protocol approved by the ethics committee of the General Medical Council (Ärztekammer Hamburg). The study was conducted in accordance with the Declaration of Helsinki and participants received 50€ for their participation. Three participants had to be excluded from the study, two due to technical issues on day 1 and one due to pathological anatomy. Additional five participants had to be excluded from day 2 (one drop-out, four due to technical issues), leaving $N = 76$ for analyses.

Stimulus material

Visual material. Two fractals (grey: RGB [230,230,230], 340 × 320 pixel, resolution: 1024 × 768) served as CSs (Supplementary Figure S1). Allocation to CS+/CS− was counterbalanced over all subjects. A white cross served as inter-trial interval (ITI). All stimuli (CSs and ITI) were during all experimental phases presented on a grey background screen (Supplementary Figure S1) to induce a general 'context' and avoid confounding renewal effects (for a discussion and recommendation, see Haaker *et al*., 2014).

Electrotactile US. The US consisted of a train of three electrotactile stimuli (interval 50 ms, duration 10 ms) administered through a surface electrode on the right dorsal hand via a DS7A electrical stimulator (Digitimer, Elwyn Garden City, UK). Intensity was calibrated individually to a maximum tolerable level of pain [mean intensity(s.d.): 6.9(4.9) mA].

Procedure

Participants performed a differential, 2-day delayed, fear conditioning and extinction paradigm within the MR-Scanner (as in Lonsdorf *et al*., 2014). In all phases of the experiment, CSs were presented in a pseudo-randomized order for 6–8 s (mean: 7 s; this jitter was introduced to enable separation of the CS and US at the neural level) whereas ITI duration was 10–16 s (mean: 13 s).

On day 1, the procedure included the attachment of two skin conductance recording and one stimulation electrode after participants were positioned inside the scanner as well as subsequent US intensity calibration. During the initial habituation phase, each of the two CSs was presented explicitly unreinforced for seven times. In the un-instructed acquisition phase both CSs were presented 14 times, whereby one of the CSs (CS+) was reinforced (100% reinforcement rate) with the US, while the second was not (CS−).

Approximately 24 h after conditioning, participants returned to the laboratory. Stimulation and recording electrodes were re-attached (but not re-calibrated) as before. During extinction, each CS was presented 14 times without reinforcement. Participants were not informed beforehand about any changes in the CS–US contingencies. Thirty seconds after the last extinction trial, participants received three (RI) USs while the grey background that served as the 'general context' during the

experiment (as during CS and ITI presentations) without any CS/ITI was presented. This RI phase lasted for 20 s. Thirteen seconds after the last RI-US the RI-test phase followed which consisted of seven presentations of each CS. Thereby, the CS-type (CS+/CS−) of the first RI-test trial was counterbalanced between participants.

Subjective and autonomic measures

Skin conductance responses (SCRs) were recorded using a BIOPAC MP-100 amplifier (BIOPAC Systems Inc., Goleta, CA, USA) with Ag/AgCl electrodes placed on the palmar side of the left hand on the distal and proximal hypothenar. Data were down-sampled to 10 Hz. The phasic SCRs to the CS onsets were semi-manually scored off-line using a custom-made software. SCR amplitudes (in μS) were scored as the largest response initiating 0.9–4.0 s after CS onset (Boucsein *et al.*, 2012). To normalize the distribution, the SCRs were logarithmized (Venables and Christie, 1980) and range-corrected to account for inter-individual variability (Lykken and Venables, 1971). Due to technical difficulties, SCR data from some participants were excluded from the analysis ($N_{Day1} = 1$; $N_{Day2} = 5$). Missing data points were excluded from the analysis. Ratings were provided once per CS type after each experimental phase (as stress/fear/tension elicited by the CS+ and CS−, respectively) on a 25-stepped visual analog scale (anchored at 0 and 100). After the RI-test both CSs were rated retrospectively referring to their first presentation following the RI-US presentation (retrospectively) as well as their last presentation. For participants who failed to give a valid retrospective rating (i.e. when the rating was not provided within a given time period of 10 s or not confirmed by pressing 'enter'), we replaced the missing data point by the mean value derived from all valid responses for this specific rating across participants ($N_{\text{one replacement}} = 16$, $N_{\text{two replacements}} = 2$).

Statistical analyses of subjective and autonomic data

For SCRs and fear ratings, separate repeated measure analyses of covariance (ANCOVAs) were performed with the within-subject factors stimulus (CS+/CS−) and time. Thereby, for ratings, time refers to the last rating provided after extinction and the rating referring to the first presentation after RI for each CS type. For SCRs, the factor time refers to the last extinction trial and the first RI-test trial for each CS type. Single trials were used for each CS, as the RI effect has been shown to be transient (Haaker *et al.*, 2014). Significant findings were further followed-up using one-factorial ANCOVAs. A main effect of time thereby is indicative of 'generalized RI' (i.e. stronger responding to both CS types after as compared to before RI) whereas a CS type*time interaction (i.e. stronger responding to the CS+ as compared to the CS− after RI as compared to before) is indicative of 'differential RI'.

For an explorative investigation of trial sequence effects, the sample was divided based on whether the CS + ($N = 40$) or the CS − ($N = 36$) was presented as the first RI-test trial. A repeated measure ANCOVA with stimulus type (CS+/CS−) and time (1st, 2nd and 3rd presentation) as within-subject factors and subgroup as between-subject variable was performed. For an explorative investigation of time stability, SCR analyses were also performed with blocks of two (last two extinction trials and first two trials after RI) and three (last three extinction trials and first three trials after RI) trials (see Supplementary Material for results).

Data were analyzed using SPSS 22 for Windows (IBM Corp., Armonk, New York). An α-level of $P < 0.05$ was considered significant, and Greenhouse–Geisser-corrected degrees of freedom were used when appropriate. As all participants were invited depending on their history of SLEs, SLE group was used as a covariate in all analyses.

fMRI data acquisition, preprocessing and statistical analyses

MRI data were acquired on a 3 Tesla MR-scanner (MAGNETOM trio, Siemens, Germany) using a 32-channel head coil. Functional data were obtained using an echo planar images sequence (TR = 2460 ms, TE = 26 ms). For each volume, 40 slices with a voxel size of $2 \times 2 \times 2$ mm (1 mm gap) were acquired sequentially. Structural images were obtained by using a T1 MPRAGE sequence.

fMRI data were analyzed using SPM8 (Welcome Trust Centre for Neuroimaging, UCL, London, UK). Preprocessing included, coregistration to the individual structural image, realignment, normalization to group-specific templates created via the DARTEL-algorithm (Ashburner, 2007) as well as smoothing (6 mm FWHM).

At the first level, four effects-of-interest regressors were built (i.e. last extinction and first RI-test trial for CS+ and CS−) as well as eight nuisance regressors (RI-USs; ratings; six movement parameters derived from realignment). All regressors of interest were modeled as stick function and time locked to stimulus (CS/US/rating) onset for volumes of interest (onset of the first regressor of interest-TR until the onset of the last regressor of interest + 3 × TR). Regression coefficients (beta values) for the regressor in each voxel were computed via the general linear model. In order to closely mirror SCR analyses, contrasts for the generalized RI effect ($Ext_{CS+/CS-} < RI_{CS+/CS-}$) and differential RI ($Ext_{CS+>CS-} < RI_{CS+>CS-}$) as well as for completeness their inverse ($Ext_{CS+/CS} > RI_{CS+/CS-}$; $Ext_{CS+>CS} > RI_{CS+>CS-}$) were estimated on the first level and taken into the second level analysis employing one-sample *t*-tests (generalized RI effect) or paired *t*-tests (differential RI effect), respectively. SLE group was used as covariate. These analyses were initially performed using single-trials but were for exploratory reasons also performed for trial-blocks of two and three single-trials per CS-type (see Supplementary Material for results) to track the temporal transience of the RI-effect (Haaker *et al.*, 2014). ROI analyses were based on the amygdala, ACC, thalamus and hippocampus masks derived from the Harvard–Oxford cortical and subcortical structural atlases (Desikan *et al.*, 2006; threshold: 0.7). As AI and vmPFC masks are not provided by the Harvard–Oxford atlases, an AI mask consisting of the dorsal and ventral AI was derived from Deen *et al.* (2011) for each hemisphere separately and for the vmPFC, a 10 mm sphere centered on coordinates derived from our previous (independent) study on RI was employed (*x*, *y*, *z*: 0, 40, −12; Lonsdorf *et al.*, 2014). For all fMRI analyses a peak voxel, FWE corrected, threshold at $P < 0.05$ (cluster size: $k \geq 10$) was considered as significant.

Exploratory, one-sided Pearson correlations [SPSS 22 for Windows (IBM Corp., Armonk, New York)] between fMRI estimates extracted from peak voxels activations derived from the above-described fMRI analyses and SCR responses were calculated for the generalized (first RI-test $trial_{CS+/CS-}$ − last Ext $trial_{CS+/CS-}$) and differential [(first RI-test $trial_{CS+\ -\ CS-}$) − (last Ext $trial_{CS+\ -\ CS-}$)] contrasts. For further explorations of the

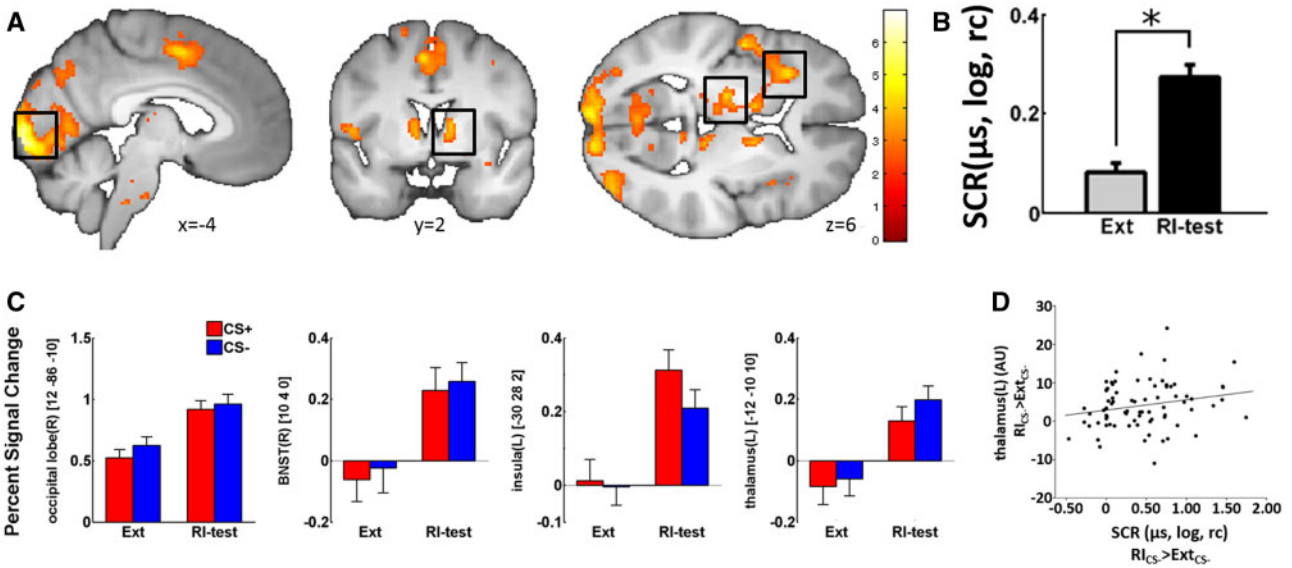

Fig. 1. (A) Neural activation reflecting the generalized RI effect ($Ext_{CS+/CS-} < RI_{CS+/CS-}$) on a visualization threshold of $P < 0.01$uc. (B) Logarithmized (log) and range corrected (rc) mean SCRs responses (in µs) (irrespective of CS type) for the last extinction (Ext) and first RI-test trial (RI-test). (C) Extracted betas values for areas observed in (A). (D) Correlation between generalized RI effect of SCR and the corresponding fMRI estimates of the thalamus peak voxel (AU, arbitrary units). Error bars represent the standard error of the mean.

differential contrast, correlations were also performed separately for both CS-types.

Results

As our analyses focused on ROF, results for the preceding acquisition and extinction phases will only be described briefly here. During acquisition the CS+ and the CS− were clearly discriminated in SCRs ($P = 0.003$) and fear ratings ($P < 0.001$) (Supplementary Table S2 and Figure S2 for details) which was on a neuro-functional level reflected in enhanced activation of areas typically observed in fear conditioning such as the dmPFC/dmACC, insula/frontal operculum, ventral striatum, thalamus and the amygdala (e.g. Sehlmeyer *et al.*, 2009; Fullana *et al.*, 2015; Supplementary Figure S3).

Importantly, CS+/CS− discrimination was no longer observable in SCRs ($P = 0.731$) or fear ratings ($P = 0.416$) or any of our ROI's at the end of extinction (Supplementary Table S2 and Figure S3).

Behavioral and physiological data

After RI (as compared to before), a general increase in SCRs was observed (generalized RI; Figure 1B), as indicated by a significant main effect of time [$F(1,71) = 13.57$; $P < 0.001$, $\eta^2 = 0.16$]. In addition, a time*stimulus interaction [$F(1,71) = 5.22$; $P = 0.025$; $\eta^2 = 0.07$], indicating differential RI, was driven by significantly stronger increase in SCR responding to the CS+ [main effect of time: $F(1,71) = 14.81$, $P < 0.001$, $\eta^2 = 0.17$] as compared to the CS− [main effect of time: $F(1,71) = 3.40$, $P = 0.07$; Figure 2B]. In contrast to these pronounced RI effects in SCRs, no significant effect involving the factor time (last extinction rating *vs* first RI rating) was observed for retrospectively provided subjective fear ratings.

Functional data

Differential RI ($Ext_{CS+>CS} < RI_{CS+>CS-}$) elicited significant activation in the vmPFC [$P_{FWE(svc)} = 0.011$] and the left hippocampus [$P_{FWE(svc)} = 0.015$], as well as, at a more lenient exploratory threshold of $P_{uc} < 0.001$, activation clusters within bilateral rectal gyrus, the left parietal operculum and the left dorsal inferior temporal lobe (Table 1 and Figure 2A). Of note, analyses for both CS types separately revealed that the hippocampal activation was driven by an increase of CS + related activation from the last extinction trial to the first RI-test trial [$P_{FWE(svc)} = 0.006$] whereas no CS − related activation was observed in this ROI. In contrast, vmPFC activation was mainly driven by a significant decrease of CS − related activation from the last extinction trial to the first RI-test trial [$P_{FWE(svc)} < 0.001$], whereas the increase in CS + related activation was only marginally significant [$P_{FWE(svc)} = 0.075$]. In line with this, extracted parameters estimates from the vmPFC peak activation cluster for differential RI (first RI-test trial−last Ext trial) for the CS−, correlated significantly negative with the SCR amplitude for the CS− contrast (first RI-test trial − last Ext trial, $r = -0.193$; $P = 0.044$) (Figure 2D) whereas no correlation was observed between extracted parameter estimates for the CS+ related vmPFC activation and the corresponding SCR contrast for the CS + ($r = -0.103$; $P = 0.183$). Parameter estimates for the hippocampus did not correlate significantly with SCRs responses (all $r < 0.185$; all $P > 0.05$).

Generalized RI ($Ext_{CS+/CS-} < RI_{CS+/CS-}$) in turn was reflected in significantly higher activation of the left thalamus [$P_{FWE(svc)} = 0.002$; Table 1 and Figure 1A] and the left insula [$P_{FWE(svc)} = 0.007$] as well as, at a more lenient exploratory threshold of $P_{uc} < 0.001$ within the occipital lobe [$P_{FWE} < 0.001$], the left parietal operculum, the left inferior parietal lobe, the left supplementary motor area, the left cuneus, the bilateral cerebellum and the bilateral BNST (Table 1). In addition, a significantly positive correlation between parameter estimates extracted from the peak voxel of thalamic activation for the generalized RI contrast ($Ext_{CS+/CS-} < RI_{CS+/CS-}$) correlated significantly positive with the corresponding SCRs differences ($r = -0.217$; $P = 0.029$) (Figure 1D), whereas no significant correlations were observed for the other ROIs (all $r < 0.188$, all $P > 0.05$).

Exploratory analyses (boundary conditions)

Further analyses exploring the time stability of both RI effects were performed using trial-blocks of two and three single trials

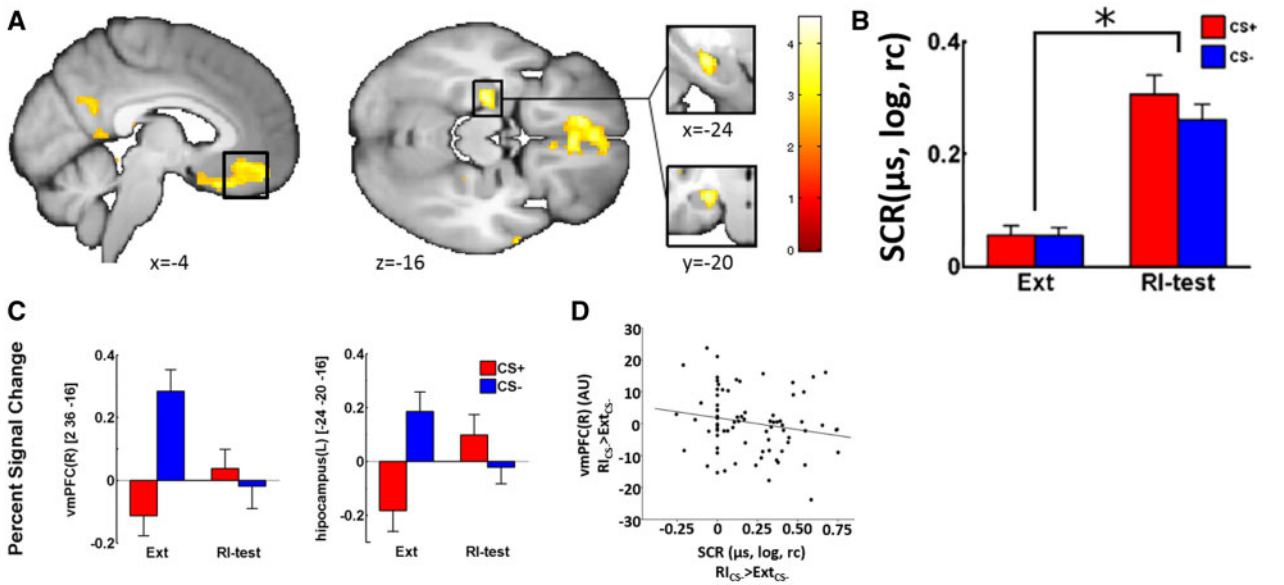

Fig. 2. (A) Neural activation reflecting the differential RI effect ($Ext_{CS+>CS-} < RI_{CS+/>CS-}$) on a visualization threshold of $P < 0.01$uc. (B) Logarithmized (log) and range corrected (rc) mean SCRs responses (in µs) per CS type for the last extinction (Ext) and first RI-test trial (RI-test). (C) Extracted betas values for areas observed in (A). (D) Correlation between CS− differences (first RI-test trial − last Ext trial) of the SCR and the corresponding fMRI estimates for the vmPFC peak voxel (AU, arbitrary units). Error bars represent the standard error of the mean.

(per CS type) and revealed comparable and consequently time stable results to the aforementioned single-trial analyses (Supplementary Figures S4 and S5).

Furthermore, with respect to possible trial sequence effects on the quality of ROF, SCR amplitudes to the CS+and CS− did not differ significantly depending on presentation sequence (CS type × time; $F < 1$; Figure 3).

Discussion

Our study provides an in-depth investigation of the neural correlates of and processes underlying generalized and differential ROF following RI and presents several key findings which substantially extend previous work and provide answers to hitherto unanswered questions. First, our results lend support from functional neuroimaging as well as autonomic measures for the idea of a distinction between the (psychological) processes underlying generalized and differential RI. Second, both processes occur in parallel and seem inherently intertwined. Finally, we provide evidence for the time-stability as well as experimental boundary conditions of the RI phenomenon. These findings as well as their implications and clinical relevance will be discussed in detail in the following.

First, our data suggest that generalized (i.e. CS unspecific) and differential (i.e. CS + specific) RI effects may in fact reflect distinct albeit intertwined processes. It has been discussed whether generalized RI, that is equal response enhancement to both the CS+ and the CS− following RI, is reflective of genuine association-based ROF processes or may merely arise through sensitization or orienting effects (e.g. Haaker *et al.*, 2014; Vervliet *et al.*, 2013a). In this context it is tempting to speculate that significant activation of thalamic and occipital regions, as observed for generalized RI, may reflect generally facilitated salience processing of incoming information (thalamus: Robinson and Petersen, 1992; Snow *et al.*, 2009; primary visual areas: Khayat *et al.*, 2006), which may support a contribution of non-associative processes to generalized RI. Interestingly a significant correlation between peak parameter estimates and SCR responses, as a measure related to arousal, was only observed for thalamic activations. It can be speculated that this might be related to the thalamus function as a gate to higher cognitive areas (Sherman and Guillery, 2002). However, it is also possible that occipital and particularly thalamic activation may (partly) reflect both the input and the output of associative processes. In addition, activation of regions also observed for generalized RI such as the insula/operculum, the BNST as well as the supplementary motor area (but also the thalamus) are linked to fear-processing and are typically observed for the CS+ > CS− contrast during fear conditioning (Fullana *et al.*, 2015). Consequently, it is tempting to interpret their activation as fear memory re-activation and possibly as RI-induced dominance of the conditioning memory trace over the extinction memory trace (Bouton, 2004; Myers and Davis, 2007). Activation of these areas during generalized RI was however unexpected, as a re-activation of fear memory can only occur to previously feared stimuli (i.e. the CS+). The CS− in turn serves as a safety signal during preceding fear conditioning which is mirrored by activation in fear inhibitory areas such as the vmPFC during conditioning (data not shown) as recently shown in a meta-analysis (Fullana *et al.*, 2015). Consequently, activation of the aforementioned areas linked to fear-processing might reflect a final common pathway of different underlying processes for both CS types (i.e. CS+ and CS−). In support of this interpretation, activation of the vmPFC and the hippocampus was observed when probing differential RI [i.e. CS+ > CS−(RI) > CS+ > CS − (EXT)] instead of the re-activation of the fear network. More precisely, vmPFC activation resulted from reduced activation to the CS− during RI as compared to the end of extinction CS− > CS + (RI) < CS− > CS + (EXT). As the vmPFC is well recognized for its role in fear inhibition during extinction recall (Kalisch *et al.*, 2006; Milad *et al.*, 2007; Lonsdorf *et al.*, 2014) and safety processing (Milad and Quirk, 2002), a reduction of vmPFC activation in our study might be indicative of RI-induced 'release from inhibition' specifically for the CS−. This is also supported by the CS− specific negative correlation between peak parameter estimates and SCR responses. This

Table 1. Neural activation reflecting the generalized and differentiated RI effects

Contrast	Brain area	x	y	z	T	P(uc)	$P(svc_{FWE})/P_{FWE}$[a]
Differential RI	vmPFC	2	36	−16	3.64	<0.001	0.011
$RI_{CS+>CS-} > Ext_{CS+>CS-}$		−4	42	−14	3.11	0.001	0.031
	Hippocampus(L)	−24	−20	−16	3.94	<0.001	0.015
	Rectal gyrus(R)	6	30	−22	4.49	<0.001	—
	Rectal gyrus(L)	−10	38	−18	3.87	<0.001	—
		−6	26	−24	3.70	<0.001	—
	Precuneus(L)	−12	−42	4	3.91	<0.001	—
	Parietal operculum(L)	−54	−6	10	4.31	<0.001	—
	Dorsal inferior temporal lobe(R)	64	0	−20	4.31	<0.001	—
	WM	22	−30	38	4.41	<0.001	—
		10	−34	−2	3.40	<0.001	—
$RI_{CS+>CS-} < Ext_{CS+>CS}$	Insula(R)	32	20	−4	3.70	<0.001	0.038
	Frontal inferior operculum(R)	56	16	6	4.06	<0.001	—
	Cerebellum(L)	−14	−50	−34	3.79	<0.001	—
	Cerebellum(R)	12	−60	−38	3.75	<0.001	—
Generalized RI	Occipital lobe(R)	12	−86	−10	6.93	<0.001	<0.001[a]
RI > Ext	Thalamus(L)	−12	−10	10	4.77	<0.001	0.002
		−14	−28	8	3.90	<0.001	0.031
	Parietal operculum(L)	−50	−4	8	4.58	<0.001	—
	Insula(L)	−32	24	6	4.25	<0.001	0.007
		−42	10	−2	3.44	<0.001	0.072
	Inferior parietal lobe(L)	−48	−50	44	4.23	<0.001	—
		−30	−50	38	3.54	<0.001	—
	Sup. motor area(L)	−4	2	50	4.22	<0.001	—
	Cuneus(L)	−12	−74	34	4.20	<0.001	—
	Cerebellum(R)	32	−48	−50	4.06	<0.001	—
		38	−46	−36	3.71	<0.001	—
		48	−54	−38	3.6	<0.001	—
	BNST(R)	10	4	0	4.02	<0.001	—
		12	2	8	3.86	<0.001	—
	BNST(L)	−10	4	8	3.92	<0.001	—
	Cerebellum(L)	−36	−54	−24	3.85	<0.001	—
RI < Ext	ACC(L)	−6	36	−4	3.71	<0.001	0.033
	ACC(R)	2	34	−8	3.48	<0.001	0.077
	Medial orbitofrontal cortex(R)	4	54	−8	4.11	<0.001	—
	WM	−26	−42	10	4.59	<0.001	—
		−34	−38	−12	4.03	<0.001	—
		−6	−18	24	3.79	<0.001	—
		10	−14	26	3.75	<0.001	—
		24	−36	12	3.60	<0.001	—

Note: WM, white matter; R, right; L, left; n.s., no suprathreshold clusters.
[a]Whole brain $P_{FWE} = 0.05$ corrected.

interpretation would be in line with the observation of strong generalized RI and our data can thus be taken to suggest that generalized RI is at least partly reflective of and may represent the output of association-based processes (i.e. response enhancement to the CS− through release from inhibition). As such, both qualities of RI seem inherently inter-twined and dependent on each other. In contrast to vmPFC activation during differential RI, hippocampus activation was largely driven by enhanced CS+ related response enhancement, rather mirroring genuine fear memory re-activation, a finding that matches previous reports of hippocampal involvement in RI-induced ROF (humans: LaBar and Phelps, 2005; rodents: Laurent and Westbrook, 2010; Lonsdorf *et al.*, 2014). In sum, hippocampus activation following RI seems to reflect ROF to the CS+ whereas attenuated vmPFC activation seems to be reflective of release from inhibition to the CS−. Of note, enhanced vmPFC activation following RI has already been observed previously for cue conditioning (as opposed to context conditioning; Lonsdorf *et al.*, 2014). Interestingly, with respect to generalization *vs* discrimination, others have reported opposing generalization gradients in areas that are in this study linked to generalized and differential RI, respectively. More specifically, positive generalization gradients (i.e. increasing activation with increasing similarity to the CS+) were observed in the insula, whereas negative generalization gradients (i.e. decreasing activation with increasing similarity to the CS+) were observed in the vmPFC and the hippocampus (Lissek *et al.*, 2010; Greenberg *et al.*, 2013). This again, may be taken to support association-based generalization processes to be involved in the RI-effects (both differential and generalized) reported here.

Finally, the generalized RI effect is very robust whereas differential ROF is more subtle within the current experimental

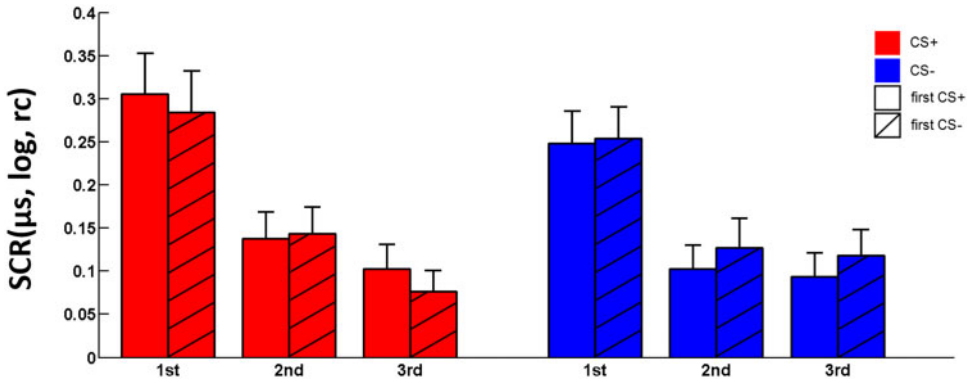

Fig. 3. Logarithmized (log) and range corrected (rc) mean SCR responses (in µs) for the 1st, 2nd and 3rd CS+ (red) and CS− (blue) trial following RI. Plain bars represent participants which were presented with the CS+ as first CS after RI. Dashed bars represent participants which were presented with the CS− as first CS after RI. Error bars represent the standard error of the mean.

design. Hence, we are tempted to speculate that both processes might be differentially vulnerable to experimental boundary conditions, which may partly explain heterogeneous findings in the (human) literature. For instance (see exploratory analyses) due to the transience of the RI effect both on a behavioral autonomic as well as on a neural level, trial sequence effects (i.e. the CS type of the first RI-test trial) have been discussed as possible experimental boundary conditions that may contribute to differentiability of the RI effect. Contrary to our expectations however, our experimental data show that the quality of RI (i.e. SCR amplitude to CS+ and CS−) was not affected by trial sequence (i.e. whether the CS+ or the CS− was presented as the first CS during RI-test), providing evidence for the experimental robustness of RI. Consequently methodological artifacts due to trial-sequence effects are unlikely to contribute to divergent findings with respect to differentiability of RI in the literature (see Haaker *et al.*, 2014), but should not be neglected in experimental designs. In addition, the RI effect was found to be most pronounced in the first trial following RI but still present when investigating blocks of two or three trials.

Future studies specifically targeting other possible individual and experimental boundary conditions through direct experimental manipulations will have to guide our understanding of the RI phenomenon. For instance, due to the crucial role of the context in RI (for a discussion, see Haaker *et al.*, 2014; Lonsdorf *et al.*, 2014) subtle context changes during RI may have substantial impact in particular on the differentiability of subsequent ROF. Comprehensive investigations of such conditions are eagerly awaited given that RI is increasingly used as outcome measure of manipulations of extinction memory consolidation.

While the above-described methodological implications may be of specific relevance only for a specialized audience, a deeper understanding of the factors contributing to differentiability of RI is of crucial clinical relevance. More precisely, the ability to maintain discrimination between aversive and safe cues under aversive circumstances is crucial to long-term remission and/or resilience and consequently for the prevention of relapse. Deficits in discrimination are in fact described as a hallmark of anxiety- and stress-related disorders (Grillon *et al.*, 2008; Lissek *et al.*, 2009, 2013; Jovanovic *et al.*, 2012; Duits *et al.*, 2015) and are predictive of resilient responding to stress (Craske *et al.*, 2013). In support of this, anxiety patients are characterized by more shallow generalization gradients as compared to controls (Duits *et al.*, 2015). To date, it remains however unclear whether patients are also characterized by a stronger tendency to generalize returning (learned) fears or by an impaired ability to inhibit fear to safety cues following an aversive (relapse-inducing) event and such clinical studies are eagerly awaited (Haaker *et al.*, 2014).

In sum, we provide compelling experimental evidence for distinct albeit intertwined underlying (psychological) processes of differential and generalized ROF and generally raise the question about appropriate control conditions in fear conditioning studies, as the CS− seems to be clearly affected by associative processes. In addition, we present initial in-depth characterization of experimental and procedural boundary conditions of ROF following RI which are informative for future studies. We suggest that ROF following RI might represent a promising laboratory intervention not only as outcome measure for fear and extinction memory manipulations but also as a possible tool for clinical applications. It is in fact conceivable that the ability of CS+/CS− discrimination following RI might prove as a predictive factor for individual relapse risk in the anxiety disorders and/or addiction. Thereby, interventions targeting discriminatory abilities may serve as a starting point for the future aim of treatment individualization and relapse prevention.

Acknowledgements

We thank Kathrin Bergholz, Timo Krämer and Katrin Wendt for help with MR data acquisition as well as Johanna Niehaus and Tobias Handtke for study assistance as well as Mareike Düesberg for helpful comments on fMRI analyses.

Funding

This work was funded by the German Research Foundation (DFG) through the SFB TRR 58 (sub-projects B07 (to T.B.L.) and Z02 as well as grant LO1980/1-1 (to T.B.L.).

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, **38**(1), 95–113.

Boucsein, W., Fowles, D.C., Grimnes, S., *et al.* (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, **49**(8), 1017–34.

Bouton, M.E. (1984). Differential control by context in the inflation and reinstatement paradigms. *Journal of Experimental Psychology: Animal Behavior Processes*, **10**(1), 56–74.

Bouton, M.E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry*, **3223**(02). Available: http://www.sciencedirect.com/science/article/pii/S0006322302015469.

Bouton, M.E. (2004). Context and behavioral processes in extinction. *Learning & Memory (Cold Spring Harbor, N.Y.)*, **11**(5), 485–94.

Bouton, M.E., Bolles, R.C. (1979). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology. Animal Behavior Processes*, **5**(4), 368–78.

Bouton, M.E., King, D.A. (1983). Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *Journal of Experimental Psychology. Animal Behavior Processes*, **9**(3), 248–65.

Canli, T., Qiu, M., Omura, K., *et al.* (2006). Neural correlates of epigenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(43). Available: http://www.pnas.org/content/103/43/16033.short.

Craske, M.G., Taylor, K.B.W., Waters, A.M., Epstein, A., Naliboff, B., Ornitz, E. (2013). Elevated responding to safe conditions as a specific risk factor for anxiety versus depressive disorders: evidence from a longitudinal investigation. *Journal of Abnormal Psychology*, **121**(2), 20.

Deen, B., Pitskel, N.B., Pelphrey, K.A. (2011). Three systems of insular functional connectivity identified with cluster analysis. *Cerebral Cortex*, **21**(7), 1498–506.

Desikan, R.S., Ségonne, F., Fischl, B., *et al.* (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, **31**(3), 968–80.

Dirikx, T., Beckers, T., Muyls, C., *et al.* (2007). Differential acquisition, extinction, and reinstatement of conditioned suppression in mice. *Quarterly Journal of Experimental Psychology (2006)*, **60**(10), 1313–20.

Duits, P., Cath, D.C., Lissek, S., *et al.* (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety*, **32**(4), 239–53.

Fitzgerald, P.J., Seemann, J.R., Maren, S. (2014). Can fear extinction be enhanced? A review of pharmacological and behavioral findings. *Brain Research Bulletin*. Available: http://doi.org/10.1016/j.brainresbull.2013.12.007.

Fullana, M.A., Harrison, B.J., Soriano-Mas, C., *et al.* (2015). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*. Available: http://doi.org/10.1038/mp.2015.88.

Greenberg, T., Carlson, J.M., Cha, J., Hajcak, G., Mujica-Parodi, L.R. (2013). Neural reactivity tracks fear generalization gradients. *Biological Psychology*, **92**(1), 2–8.

Grillon, C., Lissek, S., Rabin, S., McDowell, D., Dvir, S., Pine, D.S. (2008). Increased anxiety during anticipation of unpredictable but not predictable aversive stimuli as a psychophysiologic marker of panic disorder. *American Journal of Psychiatry*, **165**, 898–904.

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. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E2428–36.

Haaker, J., Golkar, A., Hermans, D., Lonsdorf, T.B. (2014). A review on human reinstatement studies: an overview and methodological challenges. *Learning & Memory*, **21**(9), 424–40.

Jovanovic, T., Kazama, A., Bachevalier, J., Davis, M. (2012). Impaired safety signal learning may be a biomarker of PTSD. *Neuropharmacology*, **62**, 695–704.

Kalisch, R., Korenfeld, E., Stephan, K.E., Weiskopf, N., Seymour, B., Dolan, R.J. (2006). Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, **26**(37), 9503–11.

Khayat, P.S., Spekreijse, H., Roelfsema, P.R. (2006). Attention lights up new object representations before the old ones fade away. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, **26**(1), 138–42.

Kindt, M., Soeter, M., Vervliet, B. (2009). Beyond extinction: erasing human fear responses and preventing the return of fear. *Nature Neuroscience*, **12**, 256–8.

LaBar, K.S., Phelps, E.A. (2005). Reinstatement of conditioned fear in humans is context dependent and impaired in amnesia. *Behavioral Neuroscience*, **119**(3), 677–86.

Laurent, V., Westbrook, R.F. (2010). Role of the basolateral amygdala in the reinstatement and extinction of fear responses to a previously extinguished conditioned stimulus. *Learning & Memory*, **17**(2), 86–96.

Lissek, S., Kaczkurkin, A.N., Rabin, S., Geraci, M., Pine, D.S., Grillon, C. (2013). Generalized anxiety disorder is associated with overgeneralization of classically conditioned fear. *Biological Psychiatry*. Available: http://doi.org/10.1016/j.biopsych.2013.07.025.

Lissek, S., Powers, A.S., McClure, E.B., *et al.* (2005). Classical fear conditioning in the anxiety disorders: a meta-analysis. *Behaviour Research and Therapy*, **43**(11), 1391–424.

Lissek, S., Rabin, S., Heller, R.E., *et al.* (2010). Overgeneralization of conditioned fear as a pathogenic marker of panic disorder. *American Journal of Psychiatry*, **167**, 47–55.

Lissek, S., Rabin, S.J., McDowell, D.J., *et al.* (2009). Impaired discriminative fear-conditioning resulting from elevated fear responding to learned safety cues among individuals with panic disorder. *Behaviour Research and Therapy*, **47**, 111–8.

Lonsdorf, T.B., Haaker, J., Kalisch, R. (2014). Long-term expression of human contextual fear and extinction memories involves amygdala, hippocampus and ventromedial prefrontal cortex: a reinstatement study in two independent samples. *Social Cognitive and Affective Neuroscience*. Available: http://doi.org/10.1093/scan/nsu018.

Lykken, D., Venables, P. (1971). Direct measurement of skin conductance: a proposal for standardization. *Psychophysiology*, **8**(5). Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1469–8986.1971.tb00501.x/full.

Mantsch, J.R., Baker, D.A., Funk, D., Lê, A.D., Shaham, Y. (2015). Stress-induced reinstatement of drug seeking: 20 years of progress. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*. Available: http://doi.org/10.1038/npp.2015.142.

Milad, M.R., Quirk, G.J. (2002). Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature*, **420**(6911), 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**(5), 446–54.

Myers, K.M., Davis, M. (2007). Mechanisms of fear extinction. *Molecular Psychiatry*, **12**(2), 120–50.

Oldfield, R. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. Available: http://www.sciencedirect.com/science/article/pii/0028393271900674.

Robinson, D., Petersen, S. (1992). The pulvinar and visual salience. *Trends in Neurosciences*, **15**(4), 411–4.

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**, 49–53.

Sehlmeyer, C., Schöning, S., Zwitserlood, P., *et al*. (2009). Human fear conditioning and extinction in neuroimaging: a systematic review. *PLoS One*, **4**(6), e5865.

Sherman, S.M., Guillery, R.W. (2002). The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, **357**(1428), 1695–708.

Snow, J.C., Allen, H.A, Rafal, R.D., Humphreys, G.W. (2009). Impaired attentional selection following lesions to human pulvinar: evidence for homology between human and monkey. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(10), 4054–9.

Venables, P., Christie, M. (1980). Electrodermal activity. In: Martin, I., Venables, P., editors. *Techniques in Pychophysiology*. Chichester: Wiley.

Vervliet, B., Baeyens, F., Van den Bergh, O., Hermans, D. (2013a). Extinction, generalization, and return of fear: a critical review of renewal research in humans. *Biological Psychology*, **92**(1), 51–8.

Vervliet, B., Craske, M.G., Hermans, D. (2013b). Fear extinction and relapse: state of the art. *Annual Review of Clinical Psychology*, **9**, 215–48.