The Neurocircuitry Underlying Additive Effects of Safety Instruction on Extinction Learning

Arash Javanbakht[1]*, Lana Ruvolo Grasser[1], Shantanu Madaboosi[1], Asadur Chowdury[1], Israel Liberzon[2] and Vaibhav A. Diwadkar[1]

[1] Department of Psychiatry and Behavioral Neurosciences, Wayne State University School of Medicine, Detroit, MI, United States, [2] Department of Psychiatry, Texas A&M University Central Texas, Killeen, TX, United States

Extinction learning is the dominant laboratory model for exposure therapy, a treatment involving both experience of safety near the feared object, and safety instructions relayed by a therapist. While the experiential aspect of extinction learning is well researched, less is known about instructed extinction learning and its neurocircuitry. Here, in 14 healthy participants we examined the neural correlates of, and the network interactions evoked by instructed extinction learning. Following fear conditioning to two CS+ stimuli, participants were instructed about the absence of the aversive unconditioned stimulus (US) for one of the CS+s (instructed CS; CS+I) but not the second CS+ (uninstructed CS+; CS+U). Early during extinction learning, greater activation was observed for the CS+I > CS+U contrast in regions including the vmPFC, dmPFC, vlPFC, and right parahippocampus. Subsequently, psychophysiological interaction (PPI) was applied to investigate functional connectivity of a seed in the vmPFC. This analyses revealed significant modulation of the dmPFC, parahippocampus, amygdala, and insula. Our findings suggest that the addition of cognitive instruction yields greater activation of emotion regulation and reappraisal networks during extinction learning. This work is a step in advancing laboratory paradigms that more accurately model exposure therapy and identifies regions which may be potential targets for neuromodulation to enhance psychotherapy effects.

Keywords: instructed extinction, extinction learning, fear conditioning, fear extinction, informed extinction, fMRI, PPI

INTRODUCTION

Fear conditioning is the established laboratory model for emotional learning (Armony et al., 1997; Craske et al., 2014). During classical fear conditioning, a neutral cue (conditioned stimulus, CS+) is repeatedly paired with an aversive stimulus (unconditioned stimulus, US), leading to the development of a fear response to the CS+ (even in the absence of the US) termed the conditioned response (CR). The CR may be defined as increased startle response, elevated electrodermal activity, and/or increased self-reported fear/anxiety/expectation of the CS-US pairing (Grasser and Jovanovic, 2020). This self-reported knowledge that an individual may have regarding the CS predicting the US is called contingency awareness (Craske et al., 2008), and contingency awareness in turn is a strong correlate of the conditioned response (Purkis and Lipp, 2007). Extinction learning is a complement to fear conditioning in which a new competing memory that is formed indicates

that the CS+ is no longer predictive of the US. This is achieved via repeated presentation of the CS+ without the US, leading to a decay in the CR (Milad and Quirk, 2012). Extinction learning is a form of safety learning (Grasser and Jovanovic, 2020) and a key mechanism (and thus a dominant laboratory model) for exposure therapy (Craske et al., 2008). Notably, abnormalities in extinction learning have been linked to fear related disorders such as phobias and post-traumatic stress disorder (Norrholm and Jovanovic, 2018).

Fear and safety learning in humans occur through a combination of direct experience (Pavlovian conditioning), observation (of others being exposed to adverse and safe events), and instruction (Olsson and Phelps, 2004). Instruction-mediated conditioning and learning is highly compelling from the perspective of top-down cortical mechanisms and relevant from the perspective of exposure therapy (Javanbakht et al., 2017). Behavioral and psychophysiology studies have demonstrated that in addition to fear conditioning solely through direct experience (as would most likely happen naturalistically), explicitly instructing participants to expect the US to follow the CS+ also leads to the development of a fear response to the CS+ (Olsson and Phelps, 2004; Raes et al., 2014; Cameron et al., 2016). The aforementioned studies have found that Pavlovian, instructed, and observational methods of fear conditioning result in similar levels of learning. In a similar vein, participants can be instructed about the *absence* of previously established contingency relationships between the CS+ and the US, thereby motivating extinction learning solely through instruction. For example, when instructed that they will no longer receive an aversive US after a previously conditioned CS+, participants show either immediate extinction (Mower, 1938; Rowles et al., 2012; Sevenster et al., 2012), or faster decay of the fear response (Koenig and Henriksen, 2005). These studies have provided valuable information regarding different learning mechanisms through which fear can be extinguished, but they have been limited by (a) between-subjects designs and (b) lack of neuroimaging data.

Complementary "bottom-up" and "top-down" mechanisms mediate brain network interactions underpinning higher-level processes such as self-referential processing (Frewen et al., 2020). "Bottom-up" mechanisms may primarily engage the salience network [insula, anterior cingulate cortex (ACC), and amygdala—see below] to alert individuals to threatening and rewarding stimuli in the environment without conscious knowledge/action. "Top-down" mechanisms may primarily engage the prefrontal cortex to consciously select and attend to stimuli in the environment based on conscious will/effort, and to regulate reactive "bottom-up" responses. Both are presumed to in part underpin psychotherapeutic efficacy (Malejko et al., 2017). Thus, studying their combined effects can inform neuroscientific theories about fear and safety, and enhance the clinical relevance of laboratory models (Javanbakht et al., 2017). In this work, we experimentally manipulated experience and instruction-based extinction learning, using an event-related fMRI design with the aim of identifying the evoked brain network profiles.

A comprehensive overview of the known neurocircuitry of fear learning is beyond our scope (Maren, 2001). Nevertheless, we succinctly sample from a wealth of neuroimaging data to provide a brief summary of key regions herein. The regions in this circuit include the dorsal anterior insula, dorsal ACC, and the amygdala (Adolphs, 2013; Yin et al., 2018), all of which comprise the salience network. The amygdala also playing a significant role in the development of extinction learning, disseminating signals across cortex, including throughout prefrontal regions, to orient and alert the brain to salient stimuli in the environment. Moreover, the structure activates both brainstem and hypothalamus to mount behavioral responses via the sympathetic adrenal medullary axis and the hypothalamic pituitary adrenal axis (Ross et al., 2017). The hippocampus and prefrontal cortex are involved in the formation and recall of extinction memories via context processing, consolidation, and retrieval (Greco and Liberzon, 2016), and the ventromedial prefrontal cortex (vmPFC) has a specific regulatory function regarding the inhibition of the amygdala and the fear response during extinction learning (Kim et al., 2003; Urry et al., 2006; Milad et al., 2007b; Ganella et al., 2017). Both the hippocampus and vmPFC are anatomically connected to the amygdala (Corches et al., 2019). Other areas involved in extinction learning include the anterior cingulate and dorsolateral prefrontal cortex (Greco and Liberzon, 2016). Preclinical studies (Fendt, 1998, 2000; Meyer et al., 2019) corroborate the findings from *in vivo* neuroimaging work in humans.

Functional imaging studies of *instructed* extinction learning are absent, though some studies have attempted to understand the neuronal bases of observational learning. For example, the ventromedial prefrontal cortex (vmPFC) is involved in vicarious learning of safety via observing others (Golkar et al., 2016) and, along with the hippocampus, context processing during extinction learning (Ahs et al., 2015; Hermann et al., 2017). *Context* plays a key role in extinction learning (Bouton, 2002; Bouton et al., 2006). Most studies focus primarily on *physical* context, but in fact context covers a broader spectrum including internal, temporal, and social/cognitive components (Bouton, 2002; Bouton et al., 2006; Liberzon and Sripada, 2008; Javanbakht, 2018). In this spectrum, *cognitive* context (operationalized as explicit instruction) may recruit the hippocampus and prefrontal cortex to judge the relevance of stimuli to memories and strategic goals (Liberzon and Sripada, 2008). Interestingly, instructed reappraisal, another method of emotion regulation, also activates the vmPFC as well as the dmPFC, the dlPFC, and the ACC (Pico-Perez et al., 2019).

In a previous behavioral study (Javanbakht et al., 2017) we examined the additive effects of safety instruction and experience on extinction learning. In the conditioning phase, participants were conditioned to two CS+ stimuli. Then, prior to extinction, they were instructed that the contingency pairing would be removed from one CS+, US contingency. The study revealed that instruction elicited a smaller fear response (measured by skin conductance) compared to the uninstructed CS+ during the *early* phase of extinction learning. These findings were the first to demonstrate the salient effects of instruction in facilitating extinction learning using a within-subject design comparing instruction + experience, vs. experience alone, but it

did not investigate the induced functional interactions within the neurocircuitry of fear.

In the present investigation and analyses, we explore how the combined effects of safety instruction and experiential learning modulated the (a) activation of brain regions and (b) network profiles in the brain's fear circuit. Notably, we hypothesized that the additive effect of instruction with experiential extinction learning would recruit the vmPFC and hippocampus, brain areas involved in emotion regulation and context processing. Psychophysiological interaction (PPI) was used as a simple framework for estimating directional (i.e., from seed-to-target) functional connectivity between a seed region (which in our case was the vmPFC) and its potential functional targets (O'Reilly et al., 2012; Silverstein et al., 2016).

MATERIALS AND METHODS

Participants

Thirty-seven (37) healthy male ($n = 17$) and female ($n = 20$) participants between the ages of 18 and 45 ($\overline{X}_{age} = 26.18$, $SD = 4.61$) were recruited for this study using approved flyers and university forum posts. Oral consent and initial eligibility screening were completed via phone interview. Six participants were excluded due to excess motion (> 4 mm) during MR imaging, leaving 31 (male = 16, female = 15, $\overline{X}_{age} = 26.18$, $SD = 4.78$) participants with usable data to include for analysis. The Institutional Review Board at WSU approved the protocol and all procedures herein. All participants gave informed consent to participate in the study and were able to tolerate small, enclosed spaces associated with fMRI data acquisition without anxiety. Exclusion criteria were: (1) lifetime psychiatric diagnoses (with the exception of history of substance related disorders more than 1 year prior), (2) serious medical or neurological illness that could compromise brain function, (3) history of significant closed head injury, (4) metal, implants, or metallic substances in the body, and/or (5) pregnant or trying to become pregnant. The Mini-International Neuropsychiatric Interview (MINI) was used to rule out psychiatric diagnoses. Consent and administration of the MINI were concurrent, on a separate screening day from the neuroimaging scans (and up to 2 weeks in advance of MRI acquisition). At the time of the screening, participants were exposed to a single trial of the US, a white noise burst described below, to ensure that they could tolerate the sound.

fMRI

Multiband gradient echo EPI fMRI was conducted on a 3T Siemens Verio system using a 32-channel volume head coil (310 vol, TR = 2 s, TE = 29 ms, multiband factor = 3, FOV = 256 × 256 × 144 mm^3, acquisition matrix = 128 × 128, 72 axial slices, pixel resolution = 2 × 2 × 2 mm^3, 10:48 mts). A high-resolution (1 mm^3) structural T$_1$-weighted MRI image was also collected. A scout image in each plane was acquired followed by a 3D T$_1$-weighted anatomical MRI image [3D Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence, TR = 2,150 ms, TE = 3.53 ms, TI-1,100 ms, flip-angle = 8°, FOV = 256 ×

256 × 160 mm^3, 160 axial slices of thickness = 1 mm, pixel resolution = 1 × 1 × 1 mm^3, 4:59 mts].
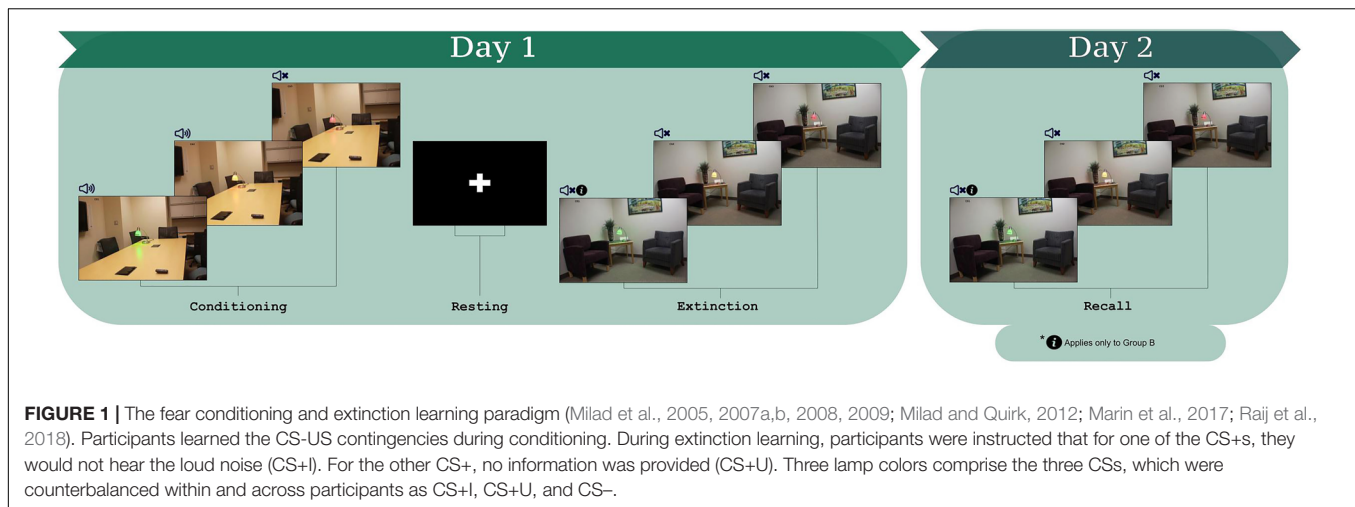
Fear Conditioning and Extinction Learning Paradigm

The employed paradigm was identical to one used previously in a study of instructed fear and extinction learning (Javanbakht et al., 2017) and has been generally applied to study fear and extinction learning (Milad et al., 2005, 2007a,b, 2008, 2009; Milad and Quirk, 2012; Marin et al., 2017; Raij et al., 2018). During the task, stimuli were presented via an MR compatible projector with the paradigm controlled using ePrime 2.0 (Schneider et al., 2002). CSs included pictures of a lamp with three different colors (red, green, and yellow) placed either in an office or in a conference room (stimuli presented in **Figure 1**). The different rooms provided context such that one room was presented consistently during conditioning and the other room during extinction (Milad et al., 2005; Marin et al., 2017; Raij et al., 2018). The assignment of lamp colors and context images to participants was randomized and counterbalanced (randomly differed across participants). The US was a 95 db white noise burst (presented for 500 ms) (Sperl et al., 2016) conveyed through MRI compatible noise canceling headphones. The headphones were worn during conditioning and extinction, signaling the possibility of the US being administered throughout all phases.

The protocol consisted of habituation, conditioning, and extinction learning (protocol schematic in **Figure 1**; Milad et al., 2005, 2007a,b, 2008, 2009; Milad and Quirk, 2012; Javanbakht et al., 2017; Marin et al., 2017; Raij et al., 2018). For habituation, participants were informed that they would see a series of images, none of which would be followed by a loud noise. The purpose of the habituation phase is to remove the effects of novelty of the images that will be presented during the task. Participants saw each of the three CSs, once in the fear conditioning context and once in the extinction context, over four trials, and never heard the loud noise. Fear conditioning immediately followed. Participants were told that they would see the previous images, some of which would be followed by a loud noise. Each image was presented for 4 s with jittered inter-trial intervals between 6 and 12 s. Each CS was presented 15 times; the two CS+ stimuli were paired with the US for 10 trials (i.e., 66% reinforcement). The US was presented 3.5 s after CS+ onset. The third CS (CS-) was never paired with the US.

The session concluded with extinction learning, 10 min after fear conditioning. Prior to extinction learning, all participants were presented with instructions stating that they would *not* hear the loud noise when they saw one of the two CS+s (Instructed CS, CS+I, "You will not hear the loud noise with the red light"). The other CS+ was defined as the uninstructed CS+, CS+U. For extinction learning, each CS was presented 12 times for 4 s with jittered inter-trial intervals between 6 and 12 s, without the US.

At the end of the conditioning phase and again at the end of the extinction phase, participants were asked to verbally rate how much they expected to hear the loud noise when

FIGURE 1 | The fear conditioning and extinction learning paradigm (Milad et al., 2005, 2007a,b, 2008, 2009; Milad and Quirk, 2012; Marin et al., 2017; Raij et al., 2018). Participants learned the CS-US contingencies during conditioning. During extinction learning, participants were instructed that for one of the CS+s, they would not hear the loud noise (CS+I). For the other CS+, no information was provided (CS+U). Three lamp colors comprise the three CSs, which were counterbalanced within and across participants as CS+I, CS+U, and CS−.

presented with each CS on a scale of 1–5, with 1 being "Not at All" and 5 being "Very Much So" (5-point likert scale). This verbally reported expectancy data was acquired for both CS+s and the CS−. Contingency awareness was defined based on expectancy data from immediately following the end of conditioning as the average of the two ratings for CS+I and CS+U being greater than that of the rating for the CS− (Tabbert et al., 2006; Javanbakht et al., 2017). Average rating for the CS+s equal to or less than the rating for the CS− was the threshold for exclusion/classification of being "unaware." Those who were not classified as "aware" of the CS-US contingency ("unaware" participants) were subsequently removed from the present analysis (Tabbert et al., 2006; Javanbakht et al., 2017). Participants for which there was no evidence of learning the CS-US contingency ("unaware") merited exclusion, given that conditioned fear must be established in order for extinction learning to be possible, and the goal of this study was to examine extinction learning.

Data Analysis

Self-Reported Expectancy

As described above, participants were asked to verbally rate how much they expected to hear the loud noise when presented with each CS on a 5-point likert scale at the conclusion of each phase—conditioning and extinction. After conditioning phase, this self-report measure of contingency/expectancy awareness was used to determine whether participants learned the CS-US contingency. We also obtained psychophysiological recordings (skin conductance response) from participants while in the scanner, but were unable to collect a sufficient amount of viable data. Measuring skin conductance in a 3T static magnetic field is a challenge both due to interference with the MRI RF. The lack of viable skin conductance data is a common problem and inherent limitation, as such data are susceptible to motion and high levels of noise (Bjorkstrand, 1990). Self-report contingency awareness data allowed us to deem participants to be "aware" of the CS-US contingency (see above).

All data were checked for out-of-range values, normality, linearity, homoscedasticity, and outliers. To compare self-reported expectancy of CS' by phase (conditioning and extinction) and CS type (CS+I, CS+U, and CS−), a 2 × 3 within-subjects repeated measures ANOVA was performed in SPSS. Mauchly's test indicated the assumption of sphericity had been met. Main effects of phase, CS type, and phase × CS type were all compared using Bonferroni corrected pairwise comparisons with 95% confidence intervals.

fMRI Preprocessing and Statistical Modeling

fMRI data were preprocessed and analyzed using standard methods in MATLAB R2013b with the Statistical Parametric Mapping toolbox (SPM12). For spatial pre-processing, the structural images were manually oriented to the AC-PC line with the reorientation vector applied across EPI image sets. Structural images were then realigned to a reference image to correct for head movement and subsequently co-registered to the structural image. The high resolution T_1 image was segmented and normalized to the MNI template, with the resultant deformations applied to the EPI image set. Low frequency components (scanner drift, physiological noises, etc.) were removed using a high-pass filter (128 s), and the EPI images were spatially smoothed using a Gaussian filter (8 mm full-width half maximum). Because of the previously observed difference between response to CS+I and CS+U during early extinction, and that at the end of the extinction CRs to both CSs fully extinguished (Javanbakht et al., 2017), we separated early extinction (first six trials of each CS+) from late extinction, as at the beginning of the extinction phase, individuals are learning that the CS+ is no longer associated with the US and forming a new, competing safety memory. At the end of the extinction phase, the assumption is (or at least the goal is to ensure) that a competing safety memory has been formed and the CR is extinguished. In first level analyses, events were modeled as regressors (duration 4 s box cars convolved with a canonical HRF) representing the CS+ (Early and Late), CS− (Early and Late)

and inter-trial intervals (ITI). Per convention, an autoregressive AR(1) model was used to account for serial correlation, and the six motion parameters (three for translation and three for rotation) were included as effects of no interest. To examine the additive effects of instruction on activation profiles of extinction learning, regressors representing early phases of CS+I > CS+U were forwarded to a second level random effects model.

A region of interest (ROI) approach was employed, restricting analyses to regions involved in extinction learning and emotion processing (Greco and Liberzon, 2016) (see section "Introduction"). The ROI approach was implemented using deterministic masks in stereotactic space (Maldjian et al., 2003). The *a priori* set of regions included the bilateral ACC and ventral PFC, bilateral amygdala, bilateral dorsal PFC, bilateral hippocampus, bilateral insula, and parahippocampal gyrus. This hypothesis driven *a priori* ROI-based approach was motivated by prior knowledge of fear circuitry (see section "Introduction").

Across all analyses, significant clusters were identified estimating the minimum cluster extent for activated voxels to be rejected as false positive (noise-only) clusters (Woodcock et al., 2016; Friedman et al., 2017). This approach performs a Monte Carlo alpha probability simulation, computing the probability of a random field of noise (after accounting for the spatial correlations of voxels based on the image smoothness within each region of interest estimated directly from the data set) to produce a cluster of a given size, after the noise is thresholded at a given level. Thus, instead of using the individual voxel probability (height) threshold alone in achieving the desired overall significance level, the method uses a combination of both probability and minimum cluster size thresholding.

RESULTS

Fifteen participants did not show evidence of the CS-US contingency awareness after conditioning (i.e., those who equally self-reported expected the US when presented with the CS+s and CS−, and those who indicated greater self-reported expectancy of the US when presented with the CS- than when presented with the CS+s). Because these participants were deemed to have failed the objective behavioral criteria for the experiment, they were excluded from the neuroimaging analyses. An additional six participants had to excluded on account of excess motion during the acquisition. Thus the final analyses included only those participants who did develop contingency awareness after conditioning *and* with viable neuroimaging data ($n = 14$).

Our subsequently presented results are organized as follows: (1) First, we present self-reported contingency awareness data for 14 participants, that reflect participant learning and task effects; (2) Next, we describe activation-based fMRI findings for each of the task conditions and comparisons (see section "Materials and Methods"); (3) These activation-based results motivated further analyses related to network effects, conducted using basic models of directional functional connectivity based on PPI (Friston et al., 1997; O'Reilly et al., 2012; Silverstein et al., 2016).

Self-Reported Contingency Awareness

As noted, self-reported expectancy information was used as a measure of contingency awareness and successful conditioning (given the absence of physiological data). Psychophysiological measures (e.g., skin conductance) are limited by large inter-individual variability and the reality that some participants do not show measurable levels of SCR during the task (Rabinak et al., 2017). Self-reported expectancy ratings of CS-US contingencies have sufficient face, diagnostic, predictive, and constructive validity, warranting them apt to assess contingency awareness (Boddez et al., 2013; Rabinak et al., 2017).

Test statistics for the repeated measure ANOVA are reported in **Table 1**. The within subjects effects of phase, CS type, and phase × CS type were all significant ($ps < 0.001$). Bonferroni-corrected pairwise comparisons indicated that expectancies across CS type were significantly greater following conditioning compared to extinction ($p < 0.001$). Bonferroni-corrected pairwise comparisons also indicated that expectancies across phase were significantly greater for CS+I compared to CS− and CS+U compared to CS- (p's $= 0.001$ and 0.003), but not CS+I compared to CS+U, $p = 1.00$. Therefore, the expectancy data indicated that participants had a greater expectancy of the CS-US contingency in after conditioning compared to after extinction learning and for both CS+I and CS+U compared to CS−. Participants' self-reported expectancy data indicates that they conditioned to the CS+s equally but not the CS-during conditioning, and extinguished conditioned fear to both CS+s by the end of the extinction phase (see **Figure 2**).
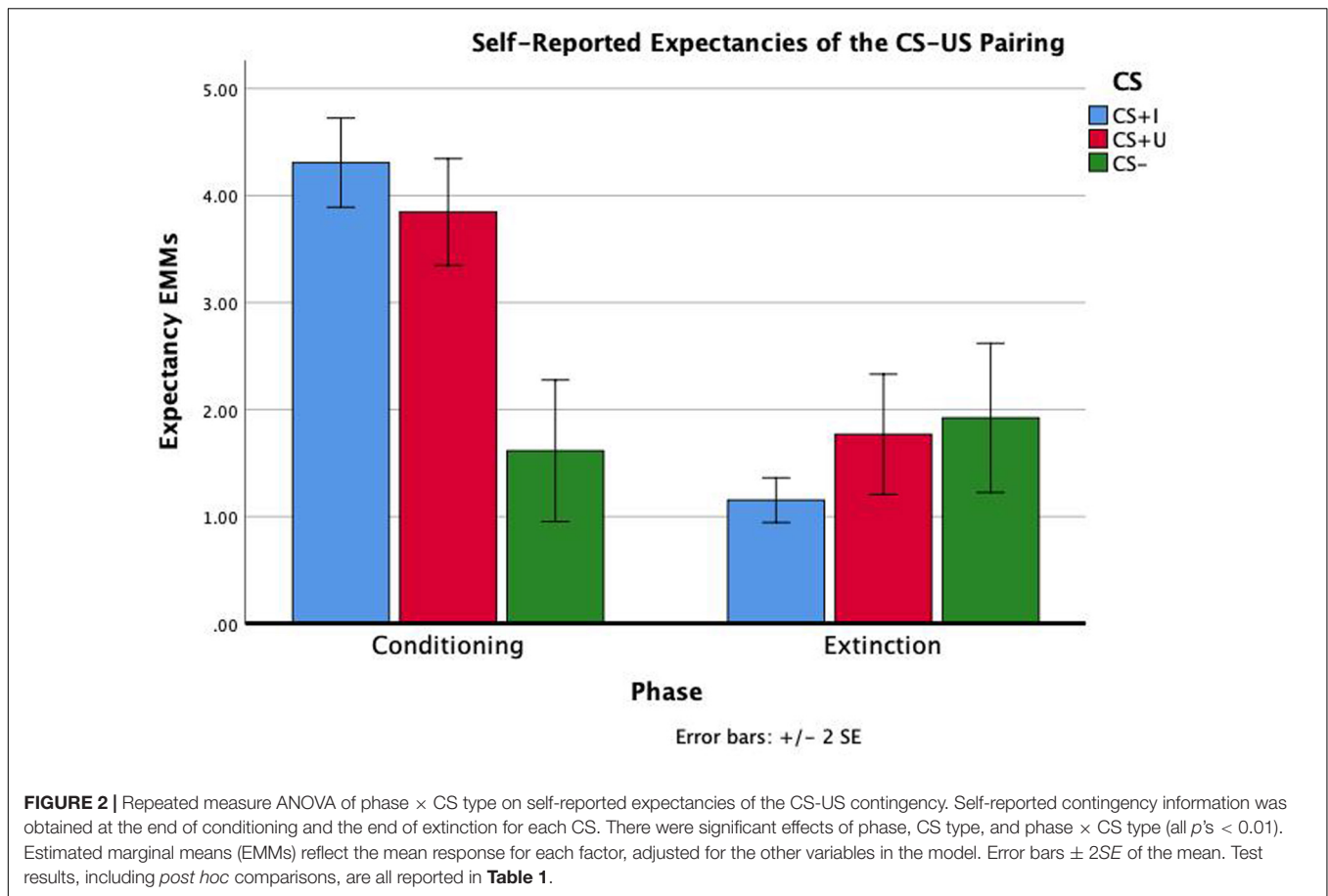
Regional Activation Differences

Markov chain Monte Carlo (MCMC) minimum cluster thresholds can be found in **Table 2** for all bilateral masks,

TABLE 1 | Test statistics for repeated measures ANOVA of self-reported expectancy of CS-US contingencies after conditioning and after extinction.

Main effects	*F*	*df*	Bonferroni corrected 95% CIs (LLCI, ULCI)	*P*-values
Phase	58.284	1,13	1.161, 2.077	<0.001*
CS type	14.980	2,26		<0.001*
CS+I v. CS+U			−0.567, 0.424	1.00
CS+I v. CS−			0.398, 1.530	0.001*
CS+U v. CS−			0.369, 1.703	0.003*
Phase × CS Type	29.922	2, 26		<0.001
Conditioning, CS+I			3.751, 4.677	*
Conditioning, CS+U			3.270, 4.301	*
Conditioning, CS−			0.903, 2.240	*
Extinction, CS+I			0.933, 1.353	*
Extinction, CS+U			1.140, 2.288	*
Extinction, CS−			1.146, 2.568	*

*For significance interpretation of Phase × CS Type interaction effects, the test of a significant interaction effect was significant, p < 0.001, and Bonferroni corrected 95% confidence intervals (CIs) are reported for specific effects where CIs that do not cross 0 are considered to represent significant effects, which are marked with an *.*

FIGURE 2 | Repeated measure ANOVA of phase × CS type on self-reported expectancies of the CS-US contingency. Self-reported contingency information was obtained at the end of conditioning and the end of extinction for each CS. There were significant effects of phase, CS type, and phase × CS type (all p's < 0.01). Estimated marginal means (EMMs) reflect the mean response for each factor, adjusted for the other variables in the model. Error bars ± 2SE of the mean. Test results, including *post hoc* comparisons, are all reported in **Table 1**.

including those for which non-significant findings were not identified. During early *extinction* (first six trials), significant activation (CS+I > CS+U) was induced in the left vmPFC, left dmPFC, bilateral vlPFC, and right parahippocampus (**Figure 3**; test statistics available in **Table 3**), though no

significant differences were observed in the converse contrast of CS+U > CS+I. Based on previous studies, our hypothesis, and our finding of involvement of vmPFC, this node was employed as an *a priori* seed in subsequent exploratory PPI analyses.

Exploratory Psycho-Physiological Interactions (PPI)

Psychophysiological interaction (PPI) (Friston et al., 1997; Friedman et al., 2017) a basic model of functional connectivity was used to explore network profiles of the functionally defined vmPFC seed during extinction learning. PPI models the response of target brain regions in terms of an interaction between a linear convolution of the physiological response of the *a priori* determined seed region and the psychological contrast of interest (e.g., CS+I > CS+U).

MCMC minimum cluster thresholds can be found in **Table 2** for all bilateral masks, including those for which non-significant findings were not identified. The left vmPFC cluster derived from the activation results (CS+I > CS+U) and centered at the significance peak (coordinates: $x = -10$, $y = 50$, $z = -1$) was defined as the seed region For each participant, time series were extracted using the first eigenvariate of the weighted means of the modeled effects within a sphere (radius = 4 mm, F contrast, effects of interest, $p < 0.99$). This time series was convolved with the contrast of interest (CS+I > CS+U). The

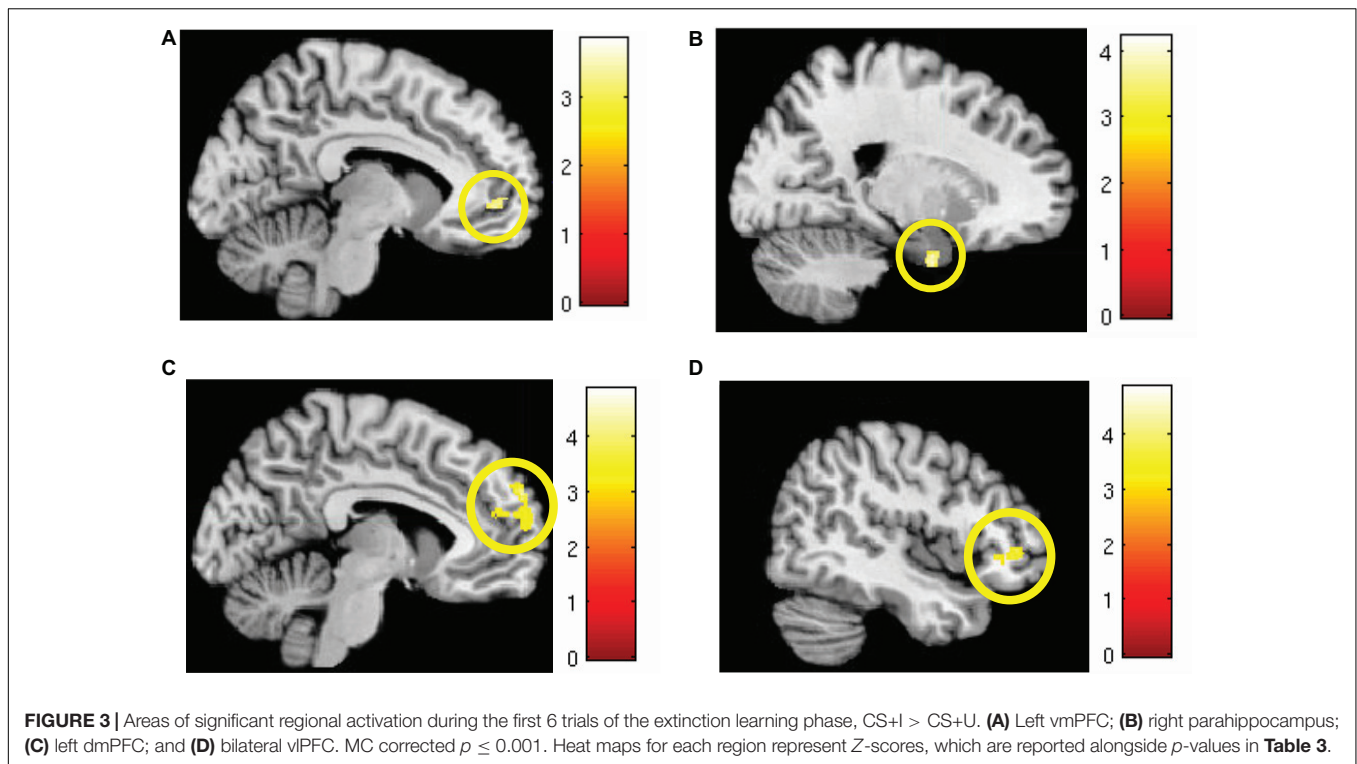**TABLE 2 |** Minimum cluster thresholds by mask for Monte Carlo corrections for multiple comparisons (MCMC) simulations.

Early extinction, CS+I > CS+U	MCMC minimum cluster threshold (regional activation; $p \leq 0.005$)	MCMC minimum cluster threshold (vmPFC interaction; $p \leq 0.05$)
Anterior cingulate cortex and ventral prefrontal cortex	55.6	20.5
Amgydala	15.2	34.9
Dorsal prefrontal cortex	173	87.3
Hippocampus	33.4	42.5
Insula	38.7	44.2
Parahippocampus	28.9	35.4

Minimum cluster thresholds represent the minimum required number of voxels showing significant activation greater than that due to noise at the a priori established p-value of 0.005 for the ROI-based analysis and 0.05 for PPI with the left vmPFC as the seed. All masks, including those for which there were no significant findings, are presented. Masks are bilateral; contrast is CS+I > CS+U for the early phase (first six trials) of extinction.

FIGURE 3 | Areas of significant regional activation during the first 6 trials of the extinction learning phase, CS+I > CS+U. **(A)** Left vmPFC; **(B)** right parahippocampus; **(C)** left dmPFC; and **(D)** bilateral vlPFC. MC corrected $p \leq 0.001$. Heat maps for each region represent Z-scores, which are reported alongside p-values in **Table 3**.

resultant PPI interaction terms models the effects of the vmPFC on any potential targets in the context of this psychological context (i.e., the contrast). Thus, each participant, contributed one first-level PPI map to a second level random effects analyses, to identify the modulatory effects of the vmPFC seed on the *a priori* network of regions (amygdala, insula, dACC, hippocampus, dmPFC, dlPFC, vlPFC, and vmPFC). Significant clusters (identified using a one-sample *t*-test) were observed in the dmPFC, parahippocampus, amygdala, and insula (**Figure 4**, statistical information in **Table 4**).

DISCUSSION

We open the Discussion by reiterating that this is the first study of the neurocircuitry evoked by the combined effects of instruction and experience on extinction learning. We reprise

our results here. In assessing activation profiles for concurrently acquired fMRI data, we found increased activation in the vmPFC during the early phase of extinction learning when comparing the CS+I with CS+U, as well as increased dmPFC, vlPFC, and parahippocampal activity. PPI analysis motivated by our hypothesis and the activation results showed that the vmPFC seed significantly modulated the dmPFC, amygdala, insula, and parahippocampus during extinction learning. In the remainder of the paper, we discuss the importance of the results from the perspective of how extinction learning is represented in brain regions and networks, and conclude with implications of these results for clinical intervention.
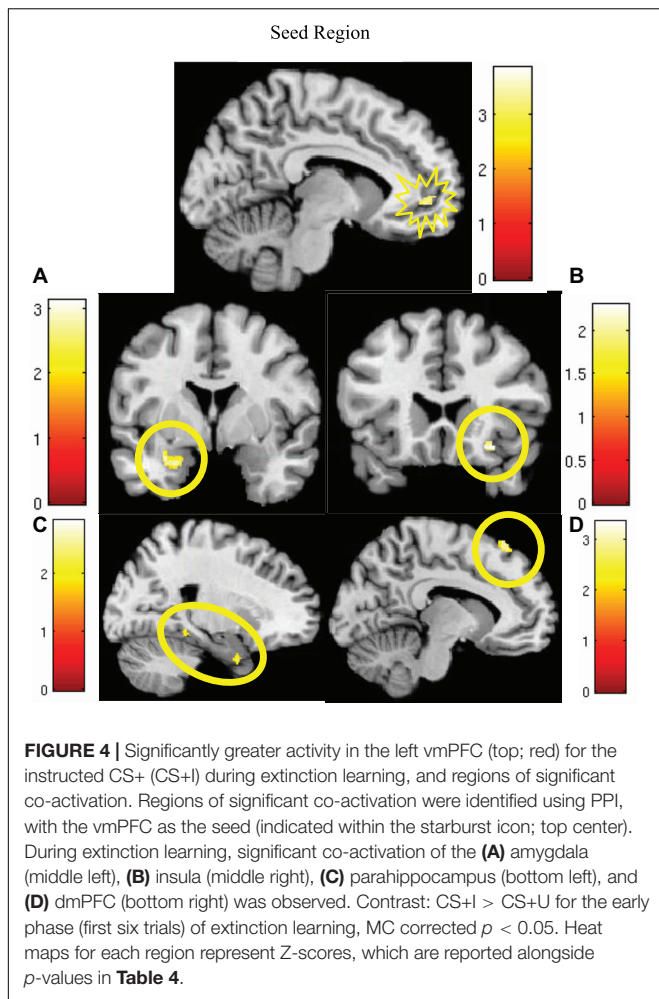
fMRI Effects: Activation and Network Profiles

The notably observed activation of the vmPFC is consistent with its purported role during extinction learning, social learning of safety, and cognitive reappraisal related emotion regulation. Previous studies have frequently reported vmPFC activation during extinction learning and its recall, and in general in response to safety signals (Milad et al., 2007b; Fullana et al., 2018). The vmPFC is also implicated in fear reversal studies when a CS previously associated with threat, is now safe. vmPFC is also suggested to have an inhibitory role on conditioned fear response during early extinction (for a review see Greco and Liberzon, 2016). Other works have shown involvement of the vmPFC in vicarious safety learning via social observation (Golkar et al., 2016) and reappraisal-related emotion regulation (Pico-Perez et al., 2019). A meta-analysis of a diverse group of emotion regulation tasks found vmPFC activation to be

TABLE 3 | Significant regions of interest activated during extinction learning.

Early extinction, CS+I > CS+U	Cluster size corrected in voxels	Coordinates	P-values (peak)	Z-scores
Left vmPFC	81	−10, 50, −1	0.001	3.09
Right parahippocampus	54	20, −1, −30	0.001	3.29
Left dmPFC	432	−9, 60, 29	<0.001	3.61
Bilateral vlPFC	184	−38, 58, −6	0.001	3.19
	190	50, 40, −1	0.001	3.07

p-values threshold: p < 0.005.

FIGURE 4 | Significantly greater activity in the left vmPFC (top; red) for the instructed CS+ (CS+I) during extinction learning, and regions of significant co-activation. Regions of significant co-activation were identified using PPI, with the vmPFC as the seed (indicated within the starburst icon; top center). During extinction learning, significant co-activation of the **(A)** amygdala (middle left), **(B)** insula (middle right), **(C)** parahippocampus (bottom left), and **(D)** dmPFC (bottom right) was observed. Contrast: CS+I > CS+U for the early phase (first six trials) of extinction learning, MC corrected $p < 0.05$. Heat maps for each region represent Z-scores, which are reported alongside p-values in **Table 4**.

the most consistent element of negative emotion regulation whether via extinction learning, cognitive emotion regulation (reappraisal studies), or placebo effect (Diekhof et al., 2011). Previous work has shown that instruction about the absence of CS-US contingency leads to either immediate extinction (Mower, 1938; Rowles et al., 2012; Sevenster et al., 2012), or faster decay of the fear response (Koenig and Henriksen, 2005). Thus, our findings suggest that this facilitation may happen through increased activation in the vmPFC in conjunction with other areas involved in emotion regulation (dmPFC, parahippocampus, insula, and amygdala). Moreover, the PPI analyses suggest that the vmPFC exerts network modulation that reflects the network-based signatures of instruction. In this way, instruction-mediated emotion regulation may be the meta-process that underpins extinction learning. The application of PPI allows for our resultant findings to support this theory, given that PPI implies directionality.

Other related mPFC regions, vlPFC and dmPFC showed activation during instructed extinction learning. The dmPFC also showed significant coactivation with the vmPFC. Both the vlPFC and dmPFC are involved in emotional regulation, subsequent behavioral responding, and are also activated during threat

appraisal (Milad et al., 2007a) and reappraisal (Buhle et al., 2014; Helion et al., 2019), which is a relevant mechanism involved here as a function of the instruction.

The parahippocampal cortex is both anatomically and functionally connected with the medial prefrontal cortex (Baldassano et al., 2013), and in addition to the hippocampus and vmPFC, is a key brain region involved in processing contextual associations (Aminoff et al., 2013). As previously noted, context plays an important role in signaling safety of the previously conditioned cues in the environment linked with extinction learning (Maren and Quirk, 2004). Activation of the parahippocampus, and its coactivation with the vmPFC, suggests a function of relaying safety instructions as a component of the cognitive context to indicate absence of the CS-US contingency (Javanbakht et al., 2017). Previous work has suggested that similar to the physical component of the context, instruction serves as a "cognitive context" that guides reactions when the conditions of the instruction are available (e.g., red light indicating absence of the loud noise) (Maren et al., 2013). Hippocampus and prefrontal cortex that are involved in context processing, are suggested to have a role in processing cognitive information during extinction learning (Garfinkel et al., 2014). In other words, cognitive context may be manipulated by the presentation of social cues or verbal instruction (Phelps et al., 2001). By changing the expectation, this cognitive manipulation can affect fear responses during extinction learning (Hugdahl and Ohman, 1977; Olsson and Phelps, 2004). This differs from cognitive "reappraisal," as the participant is not instructed to *change/reappraise* salience of a cue, but rather is *informed* about the salience.

Clinical Implications

These findings, albeit in healthy participants who were tested in an experimental lab setting, have implications for clinical practice. The finding of increased vmPFC activity during instructed extinction learning suggests that instruction may have an additive benefit in engaging this crucial structure associated with emotion and emotion regulation. Thus, therapeutic methods that utilize instruction may enhance efficacy in patients with anxiety disorder and PTSD by modulating the activity and network profiles of the vmPFC. For instance, evidence supports efficacy of both exposure therapy (based on experiential

TABLE 4 | Significant regions of co-activation with left vmPFC seed during PPI analysis, derived from ROI-based analyses.

Early extinction, CS+I > CS+U	Cluster size corrected in voxels	Coordinates	p-values (peak)	Z-scores
Bilateral parahippocampus	93	−10, −4, −19	0.006	2.51
	64	20, −38, −13	0.009	2.73
	36	21, 4, −31	0.012	2.57
Right insula	46	30, 22, −14	0.029	2.07
Left dmPFC	155	−12, 30, 59	0.003	2.79
Left amygdala	165	−22, −1, −24	0.004	2.65

p-values threshold: p < 0.05.

extinction learning), and cognitive processing therapy (focused on cognitive manipulation of patient's perspective on the trauma leading to reduced fear) in treatment of PTSD (Cusack et al., 2016). However, there is a difference between laboratory models of fear conditioning and extinction learning and real clinical practice: Here fear learning happens in the laboratory and participants may give more credence to instruction provided by the experimenter, while in clinical practice fear learning has happened prior to engaging the therapist. While most laboratory models of exposure therapy are based on experiential extinction learning, such therapy involves a combination of experience and instruction. In the clinic a therapist is always signaling safety of the feared object via instruction. This is less applicable during generalization of extinction learning with self-practice outside of the clinic where the therapist is not available, although the memory of instruction is often still present. To realistically model the neurobiology of psychotherapy, we need to understand the combined role of the human social safety cue along with the experiential learning of safety. In this sense, our paradigm adds to the current laboratory models, to be more representative of the complexities of clinical work. Finally, understanding how instruction enhances extinction learning can help in individualized treatment and identifying those who may benefit from potential utilization of neuromodulation methods targeting these deficits in vmPFC to enhance response to therapy. Neuromodulation may be helpful as adjunct to therapy of conditions which have repeatedly shown deficits in vmPFC function, and which treatment involves cognitive behavioral therapy.

We also note some study limitations. The sample size for the fMRI analyses was relatively small (for reasons detailed in the section "Materials and Methods"), and the lack of viable psychophysiological data which is a general challenge (Bjorkstrand, 1990) also affected us. We were compelled to rely on self-reported expectancy as a measure of contingency awareness and successful conditioning, while noting that these measures have been shown to be associated with psychophysiological data (Indovina et al., 2011; Pohlack et al., 2012a,b). The "awareness" of fear is a clinically relevant patient experience, suggesting that contingency awareness may be an experimental surrogate of a clinically relevant "phenotype." Additionally, multiple studies note that conditioned psychophysiological responses can be acquired even in the absence of cognitive awareness of contingency awareness (Ohman and Soares, 1994; Schultz and Helmstetter, 2010; Raio et al., 2012). Therefore, while our lack of viable psychophysiology data is certainly a limitation of this, our self-report expectancy and neuroimaging data still provide a valuable and meaningful contribution to the literature. As contingency awareness data was obtained at the end of the extinction learning phase, and not after the first six trials—early phase of extinction—we lack behavioral evidence of differences between CS+I and CS+U. Future studies should place an emphasis on gathering quality psychophysiological recordings and probe contingency awareness immediately after the early phase of extinction learning (first six trials), as well as at the end. Our neuroimaging data indicate greater recruitment of prefrontal and hippocampal areas when presented with CS+I compared to CS+U in the early

phase of extinction, and these regions are typically associated with inhibition of conditioned fear responses. Therefore, we would expect to see differences in self-reported expectancy data after the first six trials, such that participants would report less expectancy of the US to follow the CS+I compared to the CS+U. However, this remains to be tested. Thus these data do not capture potential differences between CS+I and CS+U, suggesting that future studies should emphasize collection of physiological recordings *and* probe contingency awareness immediately after the early phase of extinction learning, and at the conclusion. Finally, while the formal bases of PPI permits the inference of directional (and putatively asymmetric) interactions between seeds and targets (Stephan, 2004), recovering the true bases of directionality in the brain is a fundamentally challenging question (Friston et al., 2012). The challenge is (a) empirical, given that the fMRI signal is limited by the hemodynamic filter that is stages removed from neuronal processes and (b) philosophical, given that discovering directionality is in part based on discovering "causality" (Mannino and Bressler, 2015). We submit that our claims of directionality refer to previous understanding of the involved networks, and in the narrow sense to the nature of inference based on PPI, and that we cannot make strong claims about general directionality in the brain.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available although we may share the data on a case by case basis. Requests to access the datasets should be directed to AJ, ajavanba@med.wayne.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Wayne State University IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AJ has contributed to the design of the study, data analysis and interpretation, and writing the results. LG, SM, and AC contributed to the data collection, analysis, and interpretation, and wrote the results. IL contributed to the design of the study, data analysis, and interpretation, and wrote the results. VD contributed to the data analysis and interpretation, and wrote the results. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Adolphs, R. (2013). The biology of fear. *Curr. Biol.* 23, R79–R93. doi: 10.1016/j.cub.2012.11.055

Ahs, F., Kragel, P. A., Zielinski, D. J., Brady, R., and LaBar, K. S. (2015). Medial prefrontal pathways for the contextual regulation of extinguished fear in humans. *Neuroimage* 122, 262–271. doi: 10.1016/j.neuroimage.2015.07.051

Aminoff, E. M., Kveraga, K., and Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends Cogn. Sci.* 17, 379–390. doi: 10.1016/j.tics.2013.06.009

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., and Ledoux, J. E. (1997). Computational modeling of emotion: explorations through the anatomy and physiology of fear conditioning. *Trends Cogn. Sci.* 1, 28–34. doi: 10.1016/S1364-6613(97)01007-3

Baldassano, C., Beck, D. M., and Fei-Fei, L. (2013). Differential connectivity within the Parahippocampal place area. *NeuroImage* 75, 228–237. doi: 10.1016/j.neuroimage.2013.02.073

Bjorkstrand, P. A. (1990). Effects of conditioned stimulus pre-exposure on human electrodermal conditioning to fear-relevant and fear-irrelevant stimuli. *Biol. Psychol.* 30, 35–50. doi: 10.1016/0301-0511(90)90089-f

Boddez, Y., Baeyens, F., Luyten, L., Vansteenwegen, D., Hermans, D., and Beckers, T. (2013). Rating data are underrated: validity of US expectancy in human fear conditioning. *J. Behav. Ther. Exp. Psychiatry* 44, 201–206. doi: 10.1016/j.jbtep.2012.08.003

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biol. Psychiatry* 52, 976–986. doi: 10.1016/s0006-3223(02)01546-9

Bouton, M. E., Westbrook, R. F., Corcoran, K. A., and Maren, S. (2006). Contextual and temporal modulation of extinction: behavioral and biological mechanisms. *Biol. Psychiatry* 60, 352–360. doi: 10.1016/j.biopsych.2005.12.015

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., et al. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cereb. Cortex* 24, 2981–2990. doi: 10.1093/cercor/bht154

Cameron, G., Roche, B., Schlund, M. W., and Dymond, S. (2016). Learned, instructed and observed pathways to fear and avoidance. *J. Behav. Ther. Exp. Psychiatry* 50, 106–112. doi: 10.1016/j.jbtep.2015.06.003

Corches, A., Hiroto, A., Bailey, T. W., Speigel, J. H. III, Pastore, J., Mayford, M., et al. (2019). Differential fear conditioning generates prefrontal neural ensembles of safety signals. *Behav. Brain Res.* 360, 169–184. doi: 10.1016/j.bbr.2018.11.042

Craske, M. G., Kircanski, K., Zelikowsky, M., Mystkowski, J., Chowdhury, N., and Baker, A. (2008). Optimizing inhibitory learning during exposure therapy. *Behav. Res. Ther.* 46, 5–27. doi: 10.1016/j.brat.2007.10.003

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., and Vervliet, B. (2014). Maximizing exposure therapy: an inhibitory learning approach. *Behav. Res. Ther.* 58, 10–23. doi: 10.1016/j.brat.2014.04.006

Cusack, K., Jonas, D. E., Forneris, C. A., Wines, C., Sonis, J., Middleton, J. C., et al. (2016). Psychological treatments for adults with posttraumatic stress disorder: a systematic review and meta-analysis. *Clin. Psychol. Rev.* 43, 128–141. doi: 10.1016/j.cpr.2015.10.003

Diekhof, E. K., Geier, K., Falkai, P., and Gruber, O. (2011). Fear is only as deep as the mind allows: a coordinate-based meta-analysis of neuroimaging studies on the regulation of negative affect. *Neuroimage* 58, 275–285. doi: 10.1016/j.neuroimage.2011.05.073

Fendt, M. (1998). Different regions of the periaqueductal grey are involved differently in the expression and conditioned inhibition of fear-potentiated startle. *Eur. J. Neurosci.* 10, 3876–3884. doi: 10.1046/j.1460-9568.1998.00395.x

Fendt, M. (2000). Expression and conditioned inhibition of fear-potentiated startle after stimulation and blockade of AMPA/Kainate and GABA(A) receptors in the dorsal periaqueductal gray. *Brain Res.* 880, 1–10. doi: 10.1016/s0006-8993(00)02665-2

Frewen, P., Schroeter, M. L., Riva, G., Cipresso, P., Fairfield, B., Padulo, C., et al. (2020). Neuroimaging the consciousness of self: review, and conceptual-methodological framework. *Neurosci. Biobehav. Rev.* 112, 164–212. doi: 10.1016/j.neubiorev.2020.01.023

Friedman, A. L., Burgess, A., Ramaseshan, K., Easter, P., Khatib, D., Chowdury, A., et al. (2017). Brain network dysfunction in youth with obsessive-compulsive disorder induced by simple uni-manual behavior: the role of the dorsal anterior cingulate cortex. *Psychiatry Res. Neuroimaging* 260, 6–15. doi: 10.1016/j.pscychresns.2016.12.005

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., and Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* 6, 218–229. doi: 10.1006/nimg.1997.0291

Friston, K. J., Rabinovich, M. I., and Varona, P. (2012). "Principles of brain dynamics: global state interactions," in *From Computational Neuroscience Series*, Cambridge, MA, MIT Press.

Fullana, M. A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., et al. (2018). Fear extinction in the human brain: a meta-analysis of fMRI studies in healthy participants. *Neurosci. Biobehav. Rev.* 88, 16–25. doi: 10.1016/j.neubiorev.2018.03.002

Ganella, D. E., Barendse, M. E. A., Kim, J. H., and Whittle, S. (2017). Prefrontal-amygdala connectivity and state anxiety during fear extinction recall in adolescents. *Front. Hum. Neurosci.* 11:587. doi: 10.3389/fnhum.2017.00587

Garfinkel, S. N., Abelson, J. L., King, A. P., Sripada, R. K., Wang, X., Gaines, L. M., et al. (2014). Impaired contextual modulation of memories in PTSD: an fMRI and psychophysiological study of extinction retention and fear renewal. *J. Neurosci.* 34, 13435–13443. doi: 10.1523/jneurosci.4287-13.2014

Golkar, A., Haaker, J., Selbing, I., and Olsson, A. (2016). Neural signals of vicarious extinction learning. *Soc. Cogn. Affect. Neurosci.* 11, 1541–1549. doi: 10.1093/scan/nsw068

Grasser, L. R., and Jovanovic, T. (2020). Safety learning during development: implications for development of psychopathology. *Behav. Brain Res.*

Greco, J. A., and Liberzon, I. (2016). Neuroimaging of fear-associated learning. *Neuropsychopharmacology* 41, 320–334. doi: 10.1038/npp.2015.255

Helion, C., Krueger, S. M., and Ochsner, K. N. (2019). "Chapter 14 – emotion regulation across the life span," in *Handbook of Clinical Neurology*, Vol. 163, eds M. D'Esposito and J. H. Grafman (Amsterdam: Elsevier), 257–280. doi: 10.1016/b978-0-12-804281-6.00014-8

Hermann, A., Stark, R., Blecker, C. R., Milad, M. R., and Merz, C. J. (2017). Brain structural connectivity and context-dependent extinction memory. *Hippocampus* 27, 883–889. doi: 10.1002/hipo.22738

Hugdahl, K., and Ohman, A. (1977). Effects of instruction on acquisition and extinction of electrodermal responses to fear-relevant stimuli. *J. Exp. Psychol. Hum. Learn.* 3, 608–618. doi: 10.1037/0278-7393.3.5.608

Indovina, I., Robbins, T. W., Nunez-Elizalde, A. O., Dunn, B. D., and Bishop, S. J. (2011). Fear-conditioning mechanisms associated with trait vulnerability to anxiety in humans. *Neuron* 69, 563–571. doi: 10.1016/j.neuron.2010.12.034

Javanbakht, A. (2018). A theory of everything: overlapping neurobiological mechanisms of psychotherapies of fear and anxiety related disorders. *Front. Behav. Neurosci.* 12:328. doi: 10.3389/fnbeh.2018.00328

Javanbakht, A., Duval, E. R., Cisneros, M. E., Taylor, S. F., Kessler, D., and Liberzon, I. (2017). Instructed fear learning, extinction, and recall: additive effects of cognitive information on emotional learning of fear. *Cogn. Emot.* 31, 980–987. doi: 10.1080/02699931.2016.1169997

Kim, H., Somerville, L. H., Johnstone, T., Alexander, A. L., and Whalen, P. J. (2003). Inverse amygdala and medial prefrontal cortex responses to surprised faces. *Neuroreport* 14, 2317–2322. doi: 10.1097/00001756-200312190-00006

Koenig, K. P., and Henriksen, K. (2005). Cognitive manipulation of GSR extinction: analogues for conditioning therapies. *Integr. Physiol. Behav. Sci.* 40, 129–135. doi: 10.1007/bf03159709

Liberzon, I., and Sripada, C. S. (2008). The functional neuroanatomy of PTSD: a critical review. *Prog. Brain Res.* 167, 151–169. doi: 10.1016/S0079-6123(07)67011-3

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., and Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19, 1233–1239. doi: 10.1016/s1053-8119(03)00169-1

Malejko, K., Abler, B., Plener, P. L., and Straub, J. (2017). Neural correlates of psychotherapeutic treatment of post-traumatic stress disorder: a systematic literature review. *Front. Psychiatry* 8:85. doi: 10.3389/fpsyt.2017.00085

Mannino, M., and Bressler, S. L. (2015). Foundational perspectives on causality in large-scale brain networks. *Phys. Life Rev.* 15, 107–123. doi: 10.1016/j.plrev.2015.09.002

Maren, S. (2001). Neurobiology of Pavlovian fear conditioning. *Annu. Rev. Neurosci.* 24, 897–931. doi: 10.1146/annurev.neuro.24.1.897

Maren, S., Phan, K. L., and Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nat. Rev. Neurosci.* 14, 417–428. doi: 10.1038/nrn3492

Maren, S., and Quirk, G. J. (2004). Neuronal signalling of fear memory. *Nat. Rev. Neurosci.* 5, 844–852. doi: 10.1038/nrn1535

Marin, M.-F., Zsido, R. G., Song, H., Lasko, N. B., Killgore, W. D. S., Rauch, S. L., et al. (2017). Skin conductance responses and neural activations during fear conditioning and extinction recall across anxiety disorders. *JAMA Psychiatry* 74, 622–631. doi: 10.1001/jamapsychiatry.2017.0329

Meyer, H. C., Odriozola, P., Cohodes, E. M., Mandell, J. D., Li, A., Yang, R., et al. (2019). Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans. *Proc. Natl. Acad. Sci. U.S.A.* 116, 26970–26979. doi: 10.1073/pnas.1910481116

Milad, M. R., Orr, S. P., Lasko, N. B., Chang, Y., Rauch, S. L., and Pitman, R. K. (2008). Presence and acquired origin of reduced recall for fear extinction in PTSD: results of a twin study. *J. Psychiatr. Res.* 42, 515–520. doi: 10.1016/j.jpsychires.2008.01.017

Milad, M. R., Orr, S. P., Pitman, R. K., and Rauch, S. L. (2005). Context modulation of memory for fear extinction in humans. *Psychophysiology* 42, 456–464. doi: 10.1111/j.1469-8986.2005.00302.x

Milad, M. R., Pitman, R. K., Ellis, C. B., Gold, A. L., Shin, L. M., Lasko, N. B., et al. (2009). Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biol. Psychiatry* 66, 1075–1082. doi: 10.1016/j.biopsych.2009.06.026

Milad, M. R., and Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: ten years of progress. *Annu. Rev. Psychol.* 63, 129–151. doi: 10.1146/annurev.psych.121208.131631

Milad, M. R., Quirk, G. J., Pitman, R. K., Orr, S. P., Fischl, B., and Rauch, S. L. (2007a). A role for the human dorsal anterior cingulate cortex in fear expression. *Biol. Psychiatry* 62, 1191–1194. doi: 10.1016/j.biopsych.2007.04.032

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

Mower, O. H. (1938). Preparatory set (expectancy) – a determinant in motivation and learning. *Psychol. Rev.* 45, 62–91. doi: 10.1037/h0060829

Norrholm, S. D., and Jovanovic, T. (2018). Fear processing, psychophysiology, and PTSD. *Harv. Rev. Psychiatry* 26, 129–141. doi: 10.1097/HRP.0000000000000189

Ohman, A., and Soares, J. J. (1994). "Unconscious anxiety": phobic responses to masked stimuli. *J. Abnorm. Psychol.* 103, 231–240. doi: 10.1037//0021-843x.103.2.231

Olsson, A., and Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychol. Sci.* 15, 822–828. doi: 10.1111/j.0956-7976.2004.00762.x

O'Reilly, J. X., Woolrich, M. W., Behrens, T. E., Smith, S. M., and Johansen-Berg, H. (2012). Tools of the trade: psychophysiological interactions and functional connectivity. *Soc. Cogn. Affect. Neurosci.* 7, 604–609. doi: 10.1093/scan/nss055

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., and Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nat. Neurosci.* 4, 437–441. doi: 10.1038/86110

Pico-Perez, M., Alemany-Navarro, M., Dunsmoor, J. E., Radua, J., Albajes-Eizagirre, A., Vervliet, B., et al. (2019). Common and distinct neural correlates of fear extinction and cognitive reappraisal: a meta-analysis of fMRI studies. *Neurosci. Biobehav. Rev.* 104, 102–115. doi: 10.1016/j.neubiorev.2019.06.029

Pohlack, S. T., Nees, F., Liebscher, C., Cacciaglia, R., Diener, S. J., Ridder, S., et al. (2012a). Hippocampal but not amygdalar volume affects contextual fear conditioning in humans. *Hum. Brain Mapp.* 33, 478–488. doi: 10.1002/hbm.21224

Pohlack, S. T., Nees, F., Ruttorf, M., Schad, L. R., and Flor, H. (2012b). Activation of the ventral striatum during aversive contextual conditioning in humans. *Biol. Psychol.* 91, 74–80. doi: 10.1016/j.biopsycho.2012.04.004

Purkis, H. M., and Lipp, O. V. (2007). Automatic attention does not equal automatic fear: preferential attention without implicit valence. *Emotion* 7, 314–323. doi: 10.1037/1528-3542.7.2.314

Rabinak, C. A., Mori, S., Lyons, M., Milad, M. R., and Phan, K. L. (2017). Acquisition of CS-US contingencies during Pavlovian fear conditioning and extinction in social anxiety disorder and posttraumatic stress disorder. *J. Affect. Disord.* 207, 76–85. doi: 10.1016/j.jad.2016.09.018

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., and Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLoS One* 9:e84888. doi: 10.1371/journal.pone.0084888

Raij, T., Nummenmaa, A., Marin, M.-F., Porter, D., Furtak, S., Setsompop, K., et al. (2018). Prefrontal cortex stimulation enhances fear extinction memory in humans. *Biol. Psychiatry* 84, 129–137. doi: 10.1016/j.biopsych.2017.10.022

Raio, C. M., Carmel, D., Carrasco, M., and Phelps, E. A. (2012). Nonconscious fear is quickly acquired but swiftly forgotten. *Curr. Biol.* 22, R477–R479. doi: 10.1016/j.cub.2012.04.023

Ross, D. A., Arbuckle, M. R., Travis, M. J., Dwyer, J. B., van Schalkwyk, G. I., and Ressler, K. J. (2017). An integrated neuroscience perspective on formulation and treatment planning for posttraumatic stress disorder: an educational review. *JAMA Psychiatry* 74, 407–415. doi: 10.1001/jamapsychiatry.2016.3325

Rowles, M. E., Lipp, O. V., and Mallan, K. M. (2012). On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology* 49, 375–380. doi: 10.1111/j.1469-8986.2011.01308.x

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Schultz, D. H., and Helmstetter, F. J. (2010). Classical conditioning of autonomic fear responses is independent of contingency awareness. *J. Exp. Psychol. Anim. Behav. Process* 36, 495–500. doi: 10.1037/a0020263

Sevenster, D., Beckers, T., and Kindt, M. (2012). Instructed extinction differentially affects the emotional and cognitive expression of associative fear memory. *Psychophysiology* 49, 1426–1435. doi: 10.1111/j.1469-8986.2012.01450.x

Silverstein, B., Bressler, S., and Diwadkar, V. A. (2016). Inferring the dysconnection syndrome in schizophrenia: interpretational considerations on methods for the network analyses of fMRI data. *Front. Psychiatry* 7:132. doi: 10.3389/fpsyt.2016.00132

Sperl, M. F. J., Panitz, C., Hermann, C., and Mueller, E. M. (2016). A pragmatic comparison of noise burst and electric shock unconditioned stimuli for fear conditioning research with many trials. *Psychophysiology* 53, 1352–1365. doi: 10.1111/psyp.12677

Stephan, K. E. (2004). On the role of general system theory for functional neuroimaging. *J. Anat.* 205, 443–470. doi: 10.1111/j.0021-8782.2004.00359.x

Tabbert, K., Stark, R., Kirsch, P., and Vaitl, D. (2006). Dissociation of neural responses and skin conductance reactions during fear conditioning with and without awareness of stimulus contingencies. *NeuroImage* 32, 761–770. doi: 10.1016/j.neuroimage.2006.03.038

Urry, H. L., van Reekum, C. M., Johnstone, T., Kalin, N. H., Thurow, M. E., Schaefer, H. S., et al. (2006). Amygdala and ventromedial prefrontal cortex are inversely coupled during regulation of negative affect and predict the diurnal pattern of cortisol secretion among older adults. *J. Neurosci.* 26, 4415–4425. doi: 10.1523/JNEUROSCI.3215-05.2006

Woodcock, E. A., Wadehra, S., and Diwadkar, V. A. (2016). Network profiles of the dorsal anterior cingulate and dorsal prefrontal cortex in schizophrenia during hippocampal-based associative memory. *Front. Syst. Neurosci.* 10:32. doi: 10.3389/fnsys.2016.00032

Yin, S., Liu, Y., Petro, N. M., Keil, A., and Ding, M. (2018). Amygdala adaptation and temporal dynamics of the salience network in conditioned fear: a single-trial fMRI study. *eNeuro* 5. doi: 10.1523/ENEURO.0445-17.2018 [Epub ahead of print].