HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2022 October 05.

Published in final edited form as:

Neuroimage. 2021 December 15; 245: 118643. doi:10.1016/j.neuroimage.2021.118643.

Neural mediators of subjective and autonomic responding during threat learning and regulation

Hannah S. Savage[a,*], **Christopher G. Davey**[b], **Tor D. Wager**[c], **Sarah N. Garfinkel**[d], **Bradford A. Moffat**[e], **Rebecca K. Glarin**[e], **Ben J. Harrison**[a,*]

[a]Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne and Melbourne Health, Melbourne, Victoria 3053 Australia

[b]Department of Psychiatry, The University of Melbourne, Melbourne, Victoria 3053 Australia

[c]Department of Brain and Psychological Sciences, Dartmouth College, Hanover, NH 03755 United States

[d]Institute of Cognitive Neuroscience, University College London, London WC1N 3AZ United Kingdom

[e]Melbourne Biomedical Centre Imaging Unit, Department of Radiology, The University of Melbourne, Victoria 3010, Australia

Abstract

Threat learning elicits robust changes across multiple affective domains, including changes in autonomic indices and subjective reports of fear and anxiety. It has been argued that the underlying causes of such changes may be dissociable at a neural level, but there is currently limited evidence to support this notion. To address this, we examined the neural mediators of trial-by-trial skin conductance responses (SCR), and subjective reports of anxious arousal and valence in participants (n = 27; 17 females) performing a threat reversal task during ultra-high field functional magnetic resonance imaging. This allowed us to identify brain mediators during initial threat learning and subsequent threat reversal. Significant neural mediators of anxious arousal during threat learning included the dorsal anterior cingulate, anterior insula cortex (AIC), and ventromedial prefrontal cortex (vmPFC), subcortical regions including the amygdala, ventral striatum, caudate and putamen, and brain-stem regions including the pons and midbrain. By comparison, autonomic changes (SCR) were mediated by a subset of regions embedded within this broader circuitry that included the caudate, putamen and thalamus, and two distinct clusters

*Corresponding authors. hssavage@student.unimelb.edu.au (H.S. Savage), habj@unimelb.edu.au (B.J. Harrison).

Declaration of Competing Interest

The authors declare no competing financial interests.

Credit authorship contribution statement

Hannah S. Savage: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Christopher G. Davey:** Conceptualization, Supervision, Writing – original draft. **Tor D. Wager:** Software, Writing – original draft. **Sarah N. Garfinkel:** Writing – original draft. **Bradford A. Moffat:** Methodology, Writing – original draft. **Rebecca K. Glarin:** Methodology, Investigation, Writing – original draft. **Ben J. Harrison:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118643.

within the vmPFC. The neural mediators of subjective negative valence showed prominent effects in posterior cortical regions and, with the exception of the AIC, did not overlap with threat learning task effects. During threat reversal, positive mediators of both subjective anxious arousal and valence mapped to the default mode network; this included the vmPFC, posterior cingulate, temporoparietal junction, and angular gyrus. Decreased SCR during threat reversal was positively mediated by regions including the mid cingulate, AIC, two sub-regions of vmPFC, the thalamus, and the hippocampus. Our findings add novel evidence to support distinct underlying neural processes facilitating autonomic and subjective responding during threat learning and threat reversal. The results suggest that the brain systems engaged in threat learning mostly capture the subjective (anxious arousal) nature of the learning process, and that appropriate responding during threat reversal is facilitated by participants engaging self- and valence-based processes. Autonomic changes (SCR) appear to involve distinct facilitatory and regulatory contributions of vmPFC sub-regions.

Keywords

Neural mediators; Skin conductance response (SCR); Subjective ratings; Threat learning; Threat reversal; 7T fMRI

1. Introduction

Threat learning, also known as ‘fear conditioning’, elicits robust changes across multiple affective domains, from sympathetic autonomic responses to subjective reports of fear and anxiety (Lonsdorf et al., 2017). Whether these changes are coordinated by distinct or overlapping neural systems remains an important unanswered question. Recent work has favoured the notion that different neural mechanisms underlie different physiological and experiential channels (Barrett, 2017; Borgomaneri et al., 2020; Eisenbarth et al., 2016; LeDoux and Pine, 2016; Taschereau-Dumouchel et al., 2019), consistent with observations that autonomic and subjective changes evoked during threat learning are generally poorly correlated (Hodgson and Rachman, 1974; LeDoux and Brown, 2017). Other studies, by comparison, have reported common neural correlates of autonomic and subjective changes, particularly amongst brain regions linked to higher level autonomic-interoceptive processes, such as the dorsal anterior cingulate (dACC) and anterior insula cortex (AIC; Fullana et al., 2016; Harrison et al., 2015; Knight et al., 2005; Linnman et al., 2012; Marin et al., 2020; Milad et al., 2007; Phelps et al., 2001; Savage et al., 2020a, 2020b; Schiller and Delgado, 2010). Addressing this question ultimately promises to advance our understanding of the brain’s extended threat learning circuitry - a topic of major importance across basic and clinical neuroscience.

In a noteworthy recent study, Taschereau-Dumouchel et al. (2019) compared brain activity pattern classifiers of skin conductance responses (SCR) and subjectively reported fear to threatening animal images. In the autonomic domain, increased amygdala, insula and ventromedial prefrontal cortex (vmPFC; ~BA11m) activity emerged as a specific classifier of evoked SCR. This finding is broadly consistent with studies linking amygdala and insula reactivity to increased SCR during threat learning (Linnman et al., 2012; Phelps et al., 2001;

Savage et al., 2020a), but challenges other studies that have reported vmPFC activity to be anti-correlated with SCR (Nagai et al., 2004; Schiller et al., 2008). In the subjective domain, Taschereau-Dumouchel et al. (2019) reported that fear ratings were best classified by a pattern of distributed cortical activity, including the dorsomedial and dorsolateral prefrontal cortex. These results are consistent with prior studies of negative affect (Chang et al., 2015; Cunningham et al., 2004; Kim and Hamann, 2007; Lewis et al., 2007) and also specific threat learning studies which have linked dorsomedial frontal activity to subjective fear and anxiety (Alvarez et al., 2015; Harrison et al., 2015; Savage et al., 2020a). In Harrison et al. (2015), for example, cingulo-frontal cortex activity was specifically linked to the appraisal and experience of bodily anxiety sensations during threat learning, supporting its hypothesized role in the cognitive processing of autonomic-interoceptive signals.

In this study, we set out to further investigate the neural basis of threat-learning evoked autonomic and subjective changes. To do so, we combined a Pavlovian threat learning and reversal task (Savage et al., 2020a) with ultra-high field functional magnetic resonance imaging (UHF fMRI; 7 Tesla) and whole-brain 'multi-level mediation' modelling (Koban et al., 2019; Wager et al., 2009). This analysis enabled us to characterize, on a voxel-wise basis, brain regions engaged during the task that also satisfied criteria for mediators in a standard three variable path-modelling framework. It was our strong expectation that the enhanced sensitivity of UHF fMRI would deliver appropriate statistical power to characterize neural mediators of corresponding changes in SCR, anxious arousal and valence with high anatomical specificity (Cai et al., 2021; Morris et al., 2019; Tak et al., 2018). Using this task, we were able to additionally examine these relationships across a second learning process, threat reversal. Threat reversal requires participants to inhibit threat responses when the former threat now signals safety. This has previously been shown to be facilitated via safety signal processing (Savage et al., 2020a). Our general hypothesis was that responses in both subjective and autonomic domains would be characterised by common neural mediators across cingulo-frontal, insular and subcortical-brainstem regions commonly ascribed to the brain's extended threat learning circuit. We expected that changes in subjective domains (both anxious arousal and valence) would engage additional cortical mediators linked to subjective appraisal and regulatory mechanisms.

2. Materials and methods

2.1. Participants

Forty participants were recruited to the study. All participants met the following eligibility criteria: (i) they were aged between 18 and 25 years; (ii) had no current or past diagnosis of mental illness (iii) were competent English speakers, (iv) were not taking any psychoactive medication, and (v) had no contraindications to MRI, including pregnancy. All participants had normal or corrected-to-normal vision and provided written informed consent, following a complete description of the study protocol, which was approved by The University of Melbourne Human Research Ethics Committee. Of the initial sample, 7 participants did not complete scanning (3 participants due to equipment failure; 2 due to machine stoppage; 2 participants aborted the scan), and a further 6 participants were excluded due to excessive

head motion (see Image pre-processing). The final sample consisted of 27 participants (17 female) with a mean age of 21.93 years (± 2.28 years).

2.2. Materials and procedures

2.2.1. Experimental design—Participants completed a differential threat reversal learning task as previously reported (Savage et al., 2020a, 2020b). For consistency, we largely reproduce the same task description here. Briefly, a blue and a yellow sphere, presented for 2 s against a black background, were used as the conditioned threat (CS+) and safety stimuli (CS−). The unconditioned stimulus (US) was an aversive auditory (white noise) burst (50 ms) presented at ~90 dB. The task had 3 phases: *baseline, threat learning* (i.e., *conditioning)* and *threat reversal.* The phases were acquired in a single experimental run. During *baseline*, each coloured sphere was presented 5 times and the US did not occur.

During *threat learning*, the US co-terminated with one of the CS (forming a CS+) and not with the other (forming a CS−). The colour of the CS+ was counterbalanced across subjects and the CS-US pairing occurred one third of the time, enabling the classification of CS+unpaired trials and the subsequent analysis of CS+ responses without US confounding. During *reversal*, the pairing of the US and CS was switched (un-signalled), such that the conditioning phase CS+ became the 'new CS−', and the conditioning phase CS− became the 'new CS+' (with US pairing). 10 presentations of the CS+ unpaired, 5 of the CS+ paired (33% reinforcement rate) and 10 presentations of the CS− occurred during both *threat learning* and *threat reversal* with no more than two consecutive trials of the same stimuli. The first reinforced CS+ during *threat learning* was the second CS+ presentation, and upon reversal the first presentation of the new CS+ was reinforced. Across all phases, the inter-stimulus interval (ISI) between CS trials was 12 s during which a white visual fixation cross was presented.

At the conclusion of each phase, and as a natural progression of the task, participants were asked to rate each CS on five point Likert scales (Self-Assessment Manikins, SAM; Bradley and Lang, 1994) of affective valence and anxious arousal. To measure valence, participants responded to the question: "How unpleasant/pleasant did you find the [blue or yellow] sphere?", with responses ranging from 1=very unpleasant' to 5='very pleasant'. For anxious arousal, participants responded to the question: "How anxious did the [blue or yellow] sphere make you feel?", with responses ranging from 1='not anxious' to 5='very anxious'. Subjective ratings were unavailable for one participant. At the conclusion of the task, participants were asked to rate how unpleasant they found the noise stimulus on a scale from 1 to 10; 1='not unpleasant' to '10=very unpleasant' (mean rating 7.65 ± 1.76; range 1–10; unavailable for one participant). Participants were then asked the following multiple-choice questions to evaluate contingency awareness: "In part [2 or 3], did the noise stimulus usually occur in association with: (a) the blue sphere, (b) the yellow sphere, (c) randomly, or (d) you don't know? 21/27 participants were aware of the contingency change, and contingency awareness was unavailable for five participants.

The task was programmed in Presentation (Neurobehavioral Systems, Inc.) and was presented on a 32″ LCD BOLD screen (Cambridge Research Systems) visible via a reverse mirror mounted to the participants' head coil. Noise bursts (US) were delivered

via Sensimetrics Insert Earphones (S15 model, Sensimetrics Corp.), which also provided passive noise cancellation (~30 dB). Participants' responses were registered with a 2-button LS-PAIR Lumina response pad (Cedrus Corporation, USA), which they were familiarized with prior to scanning.

2.2.2. Task training—Prior to entering the scanner, participants were given brief instructions on the format and goals of the task. They were informed there were three parts to the task, referred to as 'Part 1, Part 2 and Part 3′. They were told that during the task they would see blue and yellow spheres presented in different orders, and that during Part 1 (*baseline*) no noise stimuli would occur. For Parts 2 and 3 (*threat learning* and *threat reversal*), they were told that their job was to try to understand the relationship between the spheres and the noise stimulus. During training, they were exposed to one instance of the noise stimulus to mitigate novelty effects. They were also instructed on how to complete the SAM ratings of anxious arousal and valence when in the scanner. Immediately prior to commencing the scan, participants were reminded of the general task instructions.

2.2.3. Subjective ratings—Subjective ratings were collected as above (see Experimental design). Paired t-tests were performed to identify whether significant differential learning had occurred within and across experimental phases (*threat learning*: CS+ vs. CS−; *threat reversal*: new CS− vs, CS+, performed in SPSS v.24).

2.2.4. Psychophysiology

Psychophysiology collection.: SCR were recorded using MRI-compatible finger electrodes (silver/silver chloride) fitted with conductance gel to the distal phalanges of the index and middle finger of participants' non-dominant hand. Fingers were cleaned using alcohol wipes prior to the attachment of electrodes and participants were instructed to keep their hand as still as possible for the duration of the scan to decrease contamination of the trace by motion. The signal was amplified and sampled at 1000 Hz using PowerLab v8.0 (ADInstruments, Dunedin, NZ) and recording was scanner-triggered concurrently with the task presentation.

Psychophysiology pre-processing and data quality check.: Taking a conservative approach to individual data screening and noise filtering, we applied a median filter (3 consecutive data points) and a bidirectional Butterworth filter (1.5 Hz low pass; 0.5 Hz high pass) to the raw SCR trace. The data was subsequently down-sampled to 10 Hz and visualised to check for evidence of responding during the *threat learning* phase of the task. An in-house MATLAB script confirmed participants were 'responders' by checking that the average peak value within a 12 s period after the onset of the reinforced-CS + stimulus during the *threat learning* phase was at least 0.02μS. No participants were excluded as a result of these criteria, although two participants were missing SCR data. Participants' (n = 25) filtered time-series were imported to the Psycho-Physiological modelling toolbox (PsPM; Bach and Friston, 2013) for further analysis run in MATLAB R 2017b (The MathWorks Inc.).

Psychophysiology analysis.: Trial onsets for each condition (*baseline* CS+, *baseline* CS−, CS+, CS+ paired, CS−, new CS+, new CS+ paired, new CS−) were specified and convolved

with a canonical skin conductance response function (Bach et al., 2009, 2013). Participants' absolute modelled peak amplitude SCR (μS) per condition were exported to SPSS (v.24) where paired t-tests were performed as for subjective ratings. To facilitate our mediation analyses of trial-by-trial SCR (see below), a separate GLM model was created. The onset of each trial of interest (CS+, CS−, new CS+, new CS−) was specified separately (total of 40 trials specified) and convolved as above. Absolute modelled peak amplitude SCR (μS) per trial were estimated for each participant.

2.2.5. fMRI acquisition and preprocessing

Image acquisition.: Imaging was performed on a 7T research scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 32-channel head-coil (Nova Medical Inc., Wilmington MA, USA). The functional sequence consisted of a multi-band (6 times) and grappa (2 times) accelerated GE-EPI sequence in the steady state (repetition time = 800 ms; echo time = 22.2 ms; and pulse/flip angle = 45°) in a 20.8 cm field-of-view, with a 130 × 130-pixel matrix and a slice thickness of 1.6 mm (no gap) (Setsompop et al., 2012). Eighty-four interleaved slices were acquired parallel to the anterior–posterior commissure line. The total sequence time was 16 min and 10 s, corresponding to 1202 whole-brain EPI volumes. A T1-weighted high-resolution anatomical image (MP2RAGE; Marques et al., 2010) was acquired for each participant to assist with functional time series coregistration (224 contiguous sagittal slices; repetition time = 5 s, echo time = 3.06 ms, inversion times = 700/2700 ms, pulse/flip angle = 13°; in a 24 cm field of view, with a 256 × 256-pixel matrix and a slice thickness of 0.73 mm). To assist with head immobility, foam-padding inserts were placed either side of the participants' head. Respiration was recorded at 50 Hz using a Siemens (Bluetooth) compatible piezoelectric respiration belt applied above the diaphragm, to be later used for physiological noise correction.

Image preprocessing.: Imaging data was transferred to a Unix-based platform that ran MATLAB R 2017b (The MathWorks Inc.) and Statistical Parametric Mapping (SPM) Version 12 (Wellcome Trust Centre for Neuroimaging). Motion correction was performed by aligning each participant's time series to the first image using least-squares minimization and a six-parameter rigid-body spatial transformation. The SPM motion fingerprint toolbox (Wilke, 2012) was then used to quantify scan-to-scan head motion on the basis of the SPM motion parameters. Participants were excluded if movement exceeded ~1.6 mm (~1 native voxel) total displacement. These realigned functional images were co-registered to each participant's respective T1 anatomical scans, which were segmented and spatially normalized to the International Consortium for Brain Mapping template using the unified segmentation approach. The functional images were smoothed with a 3.2-mm full-width-at-half-maximum (FWHM) gaussian filter. Physiological noise correction (respiration) was performed utilizing model based retrospective image correction (RETROICOR; Glover et al., 2000) in the SPM compatible PhysIO toolbox (Kasper et al., 2017).

fMRI analysis: whole-brain multi-level mediation analyses.: The M3 Mediation toolbox (https://github.com/canlab/MediationToolbox) was used to perform whole-brain multi-level mediation analyses (Koban et al., 2019; Wager et al., 2009). This approach allowed us to identify, on a voxel-wise basis, brain regions engaged during the task phases (threat

learning [CS+ > CS−] or threat reversal [new CS− > CS+]) that also satisfied criteria for mediators in a standard three variable path-modelling framework (Fig. 1). Unlike parametric correlations, which illustrate the extent to which two variables are linearly related, mediation models assess whether a mediating variable influences the relationship between two variables, supporting inferences of causality. In the case of our study, for example, we see that presentation of the CS+ (the independent variable) is associated with high ratings of anxious arousal (the dependant/outcome variable; *Path c*). The mediation model therefore tests whether brain regional activity (e.g., vmPFC activity) explains this relationship (*Path a*b*); by regressing the dependant variable on both the mediator and the independent variable we can determine whether the strength of the relationship between the CS+ and the ratings of anxious arousal is reduced (*Path c*' is the effect when we control for *Path a*b*; MacKinnon et al., 2000). The multi-level mediation estimates each path (*Path a, Path b and Path a*b*) within each subject, before testing the significance of the path coefficients across participants, using bootstrapping, where between-subject effects are treated as a random effect (Wager et al., 2009). Performing this mediation analysis at the whole-brain level and across participants, generates a map of all voxels that meet this criterion, supporting inferences of their causal role in generating/regulating the magnitude of the outcome variable (*Path a*b*).

In our models, *Path a* signified the relationship between task phase (X) and brain response (M). *Path b* signified the relationship between brain response (M) and evoked change in SCR or subjective anxious arousal or valence (Y), controlling for X. *Path a*b* signified the formal test of mediation, that is, whether the direct X–Y relationships were significantly reduced by including M in the path model. We performed six mediation analyses to identify neural mediators of subjective changes (anxious arousal and valence) and SCR during threat learning (models 1–3); and similarly, during threat reversal (models 4–6). For schematics of mediation models 1–6, see Supplementary Fig. 1. Regional path effects from these analyses are reported as significant if surviving FDR corrected $q < 0.05$ threshold ($k = 5$), which controlled for all component tests (*Path a, Path b,* and *Path a*b*) within a single mediation model, including first-level and second-level images. Statistical significance was computed via a bootstrap test (10,000 permutations), as described in Wager et al. (2008). Anatomical labelling was referenced against the CANlab_2018_combined atlas (https://sites.google.com/dartmouth.edu/canlab-brainpatterns/brain-atlases-and-parcellations/2018-combined-atlas).

Subjective ratings mediation models.: The initial variable (X) in these path models was the effects code for threat learning ([CS+, CS−, new CS+, new CS−] coded 1, −1, 0 and 0 respectively) or threat reversal ([CS+, CS−, new CS+, new CS−] coded −1, 0, 0 and 1 respectively). The outcome variable (Y) was participants' anxious arousal or valence ratings of each CS. Overall (CS+, CS−, new CS+, new CS, trial-averaged) contrast images were included as the mediating variable (M). To generate these images, each participant's preprocessed fMRI time-series was initially included in a first-level general linear model (GLM) analysis in SPM12, which specified the onsets of each CS event-type in each task phase to be convolved with canonical hemodynamic response function. The period of time when participants were making the subjective ratings was not included in the modelling of

brain responses. The fixation-cross ISI periods throughout whole task served as the implicit baseline. A high-pass filter (1/128 s) accounted for low-frequency noise, while temporal autocorrelations were estimated using a first-order autoregressive model. The subjective rating mediation models therefore allowed us to identify brain regions whose activity explained threat-evoked increases in subjective anxious arousal and negative valence during threat learning, while the threat reversal models slowed us to identify brain regions whose activity explained safety-evoked decreases in anxious arousal and increases in positive valence.

SCR mediation models.: The initial variable (X) in these path models was the trial-by-trial effects code for threat learning ([CS+ > CS−] coded 1 and −1 respectively) or threat reversal ([new CS− > CS+] coded 1 and −1 respectively). The outcome variable (Y) was participants' corresponding series of single trial SCR amplitudes. Primary contrast images were estimated for each CS trial of interest (10 trials per CS+, CS−, new CS+, new CS−) and trial-by-trial contrast images were included as the mediating variable (M). The SCR mediation model for threat learning allowed us to identify brain regions whose activity explained threat-evoked increases in SCR, while the threat reversal model allowed us to identify brain regions whose activity explained safety-evoked decreases in SCR (SCR to the 'new CS –' being smaller in threat reversal as compared to the CS+ during threat learning; see below).

3. Results

3.1. Threat learning

Participants demonstrated significant differential subjective and SCR responses during threat learning, consistent with expectations. Specifically, they demonstrated higher ratings of anxious arousal ($t_{25} = 4.59$, $p < 0.001$, Cohen's $d = 1.1$, paired t-test), more negative valence ratings ($t_{25} = -8.94$, $p < 0.001$, Cohen's $d = 2.4$, paired t-test) and larger mean SCR amplitudes ($t_{24} = 3.44$, $p = 0.002$, Cohen's $d = 0.6$, paired t-test) to the CS+ compared to the CS−. Importantly for our planned mediation analyses, these results confirmed the existence of a direct relationship between the task stimuli and outcome measures (direct X–Y relationship), allowing us to search for their neural mediators.

3.1.1. Task modulation of brain activity (Path a, X–M; see Table 1)—

Positive *Path a* effects, representing greater activation to the CS+ vs. CS− were broadly replicative of prior threat learning studies (Fullana et al., 2016), strongly implicating prominent involvement of midline cingulo-frontal cortex (incl. dACC, pre-supplementary motor area (pre-SMA), dorsomedial prefrontal cortex (PFC)) and bilateral AIC. Significant *Path a* effects also encompassed several subcortical areas including the central amygdala, subregions of the caudate nucleus, putamen, thalamus and hippocampus, as well as regions of the brain stem, including the superior colliculus, red nucleus and periaqueductal grey (PAG), and the cerebellum. Negative *Path a* effects, representing greater activation to the CS− vs. CS+, included distinct clusters across the vmPFC, as well as the bilateral lateral orbitofrontal cortex (OFC; ~BA8/9), posterior cingulate cortex extending to precuneus, temporoparietal junction, bilateral angular gyrus, and lateral visual association areas. These findings also broadly replicate prior threat learning studies comparing differential safety and

threat signals (Fullana et al., 2016; Harrison et al., 2017; Savage et al., 2020a). Results from *Path a*, which were mostly equivalent across all threat learning models are illustrated for the anxious arousal model (Fig. 2a).

3.1.2. Brain mediators of subjective responses and SCR during threat learning (Path a*b; X-M-Y; see Table 1)

Anxious arousal: The relationship between threat learning and subjective ratings of anxious arousal was positively mediated by many regions activated during threat learning; most strongly by the cingulo-frontal cortex (incl. dACC, pre-SMA, dorsomedial PFC), but also to a lesser extent by the bilateral AIC, several subcortical regions including subregions of the caudate, putamen, ventral striatum, hippocampus and thalamus, as well as multiple regions of the brainstem (red nucleus, subregions of the PAG, the superior and inferior colliculus) and the cerebellum (Fig. 2b). Some of the strongest positive mediation effects, were observed in regions that were deactivated by the task. These included the anterior vmPFC (~BA10/11) extending to the right lateral OFC (~BA8/9), the posterior cingulate extending into the precuneus, the temporoparietal junction, fusiform gyrus, and angular gyrus. There was one small cluster identified within the left visual association area that showed a negative mediation effect.

Valence: The relationship between threat learning and subjective ratings of negative valence was most strongly positively mediated by the primary visual cortex and visual association areas, and less so by sub-regions of the cerebellum, caudate and putamen (Fig. 2c). Conversely, negative mediation effects were found in the right AIC, which was activated during threat learning, and the left lateral OFC (~BA8), right superior temporal gyrus, the angular gyrus, and superior precuneus, which were deactivated during threat learning.

SCR: The relationship between threat learning and SCR was positively mediated by a small cluster of the pre-SMA, and the subregions of the caudate, putamen, thalamus, and cerebellum; regions activated during threat learning (Fig. 2d). One of the strongest positive mediators was located within a vmPFC sub-region with a distinct ventral focus (~BA11), which was deactivated during threat learning. Additional positive mediation effects were observed in the right lateral OFC (~BA11) and the right primary visual cortex. Conversely, the sub-genual ACC (~BA25), the tip of the dorsal temporal pole, bilateral temporoparietal junction, the precuneus, supramarginal gyrus and the right visual association area, areas deactivated during threat learning, were identified as negative mediators.

3.2. Threat reversal

We next confirmed that participants' ratings of subjective anxious arousal and valence and SCR reversed to reflect the change in task contingencies during threat reversal. Participants' subjective ratings reflected a decrease in anxious arousal ($t_{25} = -5.93$, $p < 0.001$, Cohen's $d = 1.3$, paired samples t-test), an increase in positive valence ($t_{25} = 8.14$, $p < 0.001$, Cohen's $d = 1.9$, paired samples t-test) and significantly decreased SCR to the new CS− compared to when it was a CS+ ($t_{24} = -3.62$, $p = 0.001$, Cohen's $d = 0.7$, paired samples t-test), as shown previously (Savage et al., 2020a). We then investigated which brain regions mediated

this relationship: that is, which regions mediated the reduction in subjective anxious arousal and SCR amplitude, and increase in positive valence ratings.

3.2.1. Task-evoked modulation of brain activity (Path a; X-M; see Table 2)—

Adding to previous work (Savage et al., 2020a; Schiller et al., 2008), threat reversal mapped onto significant activation of the dorsal vmPFC (BA10), lateral OFC (~BA8/9), the posterior cingulate cortex extending to the medial precuneus, angular gyrus, ventral striatum and hippocampal-dentate, the cerebellum (seventh and eight lobes), the primary visual cortex and visual association areas. Conversely, activity within the mid-cingulate cortex, pre-SMA, bilateral AIC, subcortical sub-regions within the precuneus, caudate, putamen and thalamus, as well the cerebellum (sixth lobe) was greater to the former threat. Results from *path a* are illustrated for the anxious arousal model (see Fig. 3a).

3.2.2. Brain mediators of task-evoked subjective changes and SCR (Path a*b; X-M-Y)

Anxious arousal: The decrease in anxious arousal ratings during threat reversal was positively mediated by many of the regions activated during threat reversal. This included two areas within the vmPFC (a dorsal cluster (~BA10), and a cluster on the genu of the dACC (~BA25)), the supramarginal gyrus, medial temporal lobe and temporal pole, fusiform gyrus, right nucleus accumbens, ventral striatum, cerebellum, primary visual cortex and visual association areas. Positive mediation effects were also observed within the mid cingulate cortex/preSMA, subcortical subregions within the caudate, putamen, and thalamus, areas of the brainstem (red nucleus, pontine subregions), and the cerebellum (sixth lobe); regions deactivated during threat reversal. Conversely, two small clusters in the superior dmPFC (~BA6), activated during threat reversal, showed a negative mediation effect (see Fig. 3b).

Valence: The increase in positive valence ratings during threat reversal was most strongly positively mediated by a cluster within the dorsal vmPFC (~BA10) and another in the posterior cingulate extending into the precuneus. Additional positive mediators were observed in the lateral PFC (~BA8), superior medial PFC (~BA6), angular gyrus, the left temporoparietal junction and the right fusiform gyrus. Regions that were deactivated during threat reversal that showed positive mediation effects, included the mid cingulate cortex, the bilateral AIC, the right anterior caudate, the mediodorsal extending to anterior ventral thalamus, and the cerebellum. Negative effects were found within the inferior parietal lobule, primary visual cortex and visual association areas (see Fig. 3c).

SCR: Decreased SCR during threat reversal was most strongly positively mediated by a three vmPFC subregions (a lateral region of the vmPFC (~BA10), a ventral vmPFC (~BA11) region, and the pregenual ACC (~BA25). Other regions activated during threat reversal that showed positive mediation effects, included a region of the dmPFC (~BA8), the fusiform gyrus, intralaminar thalamus and hippocampus, primary visual cortex and visual association areas. The mid cingulate cortex, SMA, AIC, and subregions of the pallidum and cerebellum, which were deactivated during threat reversal, also showed positive mediation effects. No negative mediation effects met our corrected threshold (see Fig. 3d).

Supplementary materials and online data: As they are not central to the aims or hypotheses of this study, the results for Path b effects (i.e. brain activity predicting non-specific changes in subjective ratings and SCR; M-Y) can be found in the supplementary materials. The data required to run this analysis have been uploaded to figshare, and are freely accessible here: http://doi.org/10.26188/14540004. The M3 Mediation toolbox used to run these analyses can be accessed here: https://github.com/canlab/MediationToolbox. The complete output from the mediation toolbox, including detailed results tables (with individual cluster coordinates, volume and effect size) at both FDR $q<0.05$ and uncorrected $p<0.01$ thresholds, as well as comprehensive figures of each mediation model's paths are also available online (http://doi.org/10.26188/14540004).

4. Discussion

Using ultra-high field functional magnetic resonance imaging (UHF fMRI) and multi-level mediation analyses, we identified both common and distinct neural mediators of autonomic (SCR) and subjective responses during threat learning and threat reversal. While our general hypotheses regarding the nature of common and distinct mediators were supported, our results also emphasise the relationship between the function of this circuitry and subjective affect, particularly threat-evoked anxious arousal.

4.1. The extended neural circuitry of human threat learning mapped with UHF fMRI

Replicating and extending past studies of threat-safety learning tasks, including threat- and safety-reversal learning (Fullana et al., 2016; Savage et al., 2020a), we have provided one of the most anatomically comprehensive mappings of human threat learning neural circuitry. We detail a rich profile of subcortical and brainstem activation that features the central amygdala, and structurally connected output regions including the striatum, PAG and medulla (Fanselow, 1991). While widely implicated in the expression of threat behaviours in animal studies (Bittencourt et al., 2005; Fanselow, 1991; Kalin et al., 2004; Keifer Jr et al., 2015), these regions have correspondingly been inconsistently implicated, and minimally studied during threat responding in humans (Fullana et al., 2016; Hermans et al., 2013; Mobbs et al., 2007; Satpute et al., 2013; Wang et al., 2020). By comparison, regions that decreased activity during threat learning remained primarily cortical in nature, including the vmPFC, and other regions that are broadly considered to encompass the 'default mode network' (DMN; Harrison et al., 2017; Tashjian et al., 2021).

4.2. Neural mediators of subjective experience within the extended threat learning circuitry

The strongest positive neural mediators of heightened subjective anxious arousal *(Path a*b)* broadly encapsulated regions that were both activated and deactivated during threat learning (*Path a*); in particular, activation within the cingulo-frontal cortex and AIC, and deactivation of the vmPFC. This result extends previous studies linking activity within the cingulo-frontal cortex and AIC to ratings of subjective arousal (Alvarez et al., 2015; Harrison et al., 2015; Phan et al., 2003; Savage et al., 2020a, 2020b). Our results within the cingulo-frontal cortex and AIC are interesting to consider with regards to prominent models of interoception and brain function, which emphasize their role in appraising prediction error signals, potentially

generated by positive mediators within the caudate, putamen, and cerebellum (den Ouden et al., 2012; Ernst et al., 2019; Schiller et al., 2008), for the genesis of subjective affect (Barrett and Simmons, 2015; Paulus and Stein, 2006). The strong positive mediation effect observed within the vmPFC indicates participants who had less suppression in this region experienced heightened anxious arousal. Consistent with this, studies that included individuals with threat-related disorders reported correlations between maladaptive hyperactivation of the vmPFC and increased subjective arousal (Apergis-Schoute et al., 2017; Cha et al., 2014; Jovanovic et al., 2012; Via et al., 2018). Compared to subjective anxious arousal, subjective ratings of negative valence were most strongly mediated by regions that did not overlap with *Path a* effects, instead showing prominent effects in posterior cortical regions. This result suggests that the extended circuitry of threat learning, and the subjective experience it manifests, is best captured by participants ratings of anxious arousal.

4.3. Neural mediators of SCR during threat learning are overlapping but discrete

Compared to the more general threat-safety circuitry that mediated subjective arousal, the significant neural mediators of SCR were located in a discrete subset of cortical and subcortical regions. While the central amygdala did emerge in *Path a* effects, and previous studies of threat learning have found amygdala activity to correlate with evoked SCR (Knight et al., 2005; Labar et al., 1998; MacNamara et al., 2015; Phelps et al., 2001), the amygdala did not emerge as a significant mediator of SCR during threat learning (*Path a* b*); we do, however, acknowledge that our stringent FDR correction may have increased the risk of Type 2 errors. The positive mediators we identified are, however, consistent with prior studies investigating the neural correlates of SCR during tasks and at rest (Critchley et al., 2000; Patterson II et al., 2002; Zhang et al., 2012). Within the vmPFC, two distinct clusters emerged as neural mediators of increased SCR during threat learning; a positive mediator located posteriorly and a negative mediator in the subgenual ACC (see Fig. 2). Previous studies have suggested the vmPFC inhibits threat-responsive regions, as the subgenual negative mediator would suggest (Phelps et al., 2004). The positive mediator we identified, however, adds to recent work that indicates the ventral region of the vmPFC (and the lateral OFC) facilitates and regulates threat-related SCR (Battaglia et al., 2020; Taschereau-Dumouchel et al., 2019). Future studies should probe the dissociation of these functional sub-regions further, and investigate their role within the context of other mediators of autonomic responding.

4.4. Threat reversal requires engagement of default mode regions and suppression of brain regions that generate subjective anxious arousal

Many of the regions activated during threat reversal (new CS− > CS+) broadly map to the DMN, and showed equally strong positive mediation effects across both subjective domains. The strongest effects were observed within the dorsal vmPFC (BA10) together with posterior midline regions, replicating past work (Savage et al., 2020a; Schiller et al., 2008). Activity was also implicated in additional cortical and subcortical regions that had not previously been implicated in this flexible process. Activation in DMN regions likely reflects successful safety learning (Harrison et al., 2017), and the refinement of affective predictions that are known facilitate the learning and embedding of contextual safety cues (Barrett and Bar, 2009; Marstaller et al., 2017). Relatedly, *Path b* effects demonstrated

stronger and more extensive involvement of DMN regions in generating positive valence, compared to anxious arousal, suggesting subjective revaluation during threat reversal may be dominated by valence-based processing.

Positive mediation effects within the cingulo-frontal cortex and AIC, regions previously activated during threat learning and now deactivated during threat reversal, were common to both subjective domains (though slightly stronger for valence). This suggests sustained or greater activation within these regions during threat reversal may facilitate appropriate and successful subjective revaluation; people who showed greater deactivation within cingulo-frontal regions during threat reversal showed less flexible and more conserved threat responding. This is consistent with the aforementioned idea that positive mediators in the thalamus, caudate and the cerebellum may generate prediction error signals to the 'new CS−' (that used to signal threat), that are appraised by the cingulo-frontal cortex to construct subjective affect.

During threat reversal a sub-genual vmPFC cluster emerged as a positive mediator of SCR; greater activity in this region during the safety condition caused a smaller SCR. This is consistent with the region's negative mediation effect observed during threat learning and lends further evidence to the inhibitory hypothesis. The ventral vmPFC (~BA11) that was a positive mediator during threat learning, appeared as a positive mediator of SCR during threat reversal; less activity, or hypoactivation, in this region during threat reversal resulted in a greater SCR during threat reversal, supporting the involvement of this sub-region in the generation of SCR (Battaglia et al., 2020).

4.5. Limitations and future directions

Our task asked participants 'how anxious the [blue/yellow] sphere made them feel'. The use of the term 'anxious', which would most commonly be perceived as a negative experience, and as opposed to more neutral terms such as 'alert' or 'aroused', may to a certain extent, have conflated arousal and negative valence. The structure of affect, and whether arousal and valence are truly independent constructs that can be reproducibly parcellated by lay participants, remains a topic of debate (Terracciano et al., 2003). Nevertheless, our results suggest that we were able to capture unique variance, as indicated by a number of distinct neural mediators across these subjective domains. When designing future experiments, researchers should take care to ensure that subjective prompts use neutral language so as not to conflate affective domains.

Our results suggest there may be distinct mediation effects for each domain, however this should not be taken to imply that they are entirely separable. Future studies should more directly compare threat learning and reversal mediation effects both within and between response domains. Our study took advantage of modelling individual trials to facilitate a sensitive characterisation of the neural drivers of SCR, however our subjective reports were made at a single time point at the conclusion of each phase. This meant it was not possible to construct integrated mediation models, where one domain was adjusted for the other in order to control for or explore shared variance or potentially different habituation rates. Future work integrating trial-by-trial subjective ratings (e.g. expectancy ratings) may overcome this limitation, though recording subjective ratings more frequently has recently been shown to

influence participant's learning and alter autonomic responses (Lonsdorf et al., 2017; Ryan et al., 2021). In addition, while the additional statistical power afforded by UHF imaging aids in reducing the sample size necessary for robust mediation analysis, future studies should address this directly and employ sensitivity analyses (VanderWeele, 2016) to clarify the degree of confounding by unmeasured variables.

While threat learning studies most commonly record SCR, other autonomic measures, such as cardiac responses, are also informative (Lonsdorf et al., 2017). As recent work supports a distinction between these autonomic domains, such as between heart rate and SCR (Eisenbarth et al., 2016), future studies might compare the neural mediators of other autonomic domains to subjective reports. The relationship between autonomic responding and subjective experience is likely altered by individual differences in interoceptive abilities, such as a participants' metacognitive awareness of their interoceptive signals (Garfinkel et al., 2015). Future studies should aim to determine whether different afferent autonomic signals modulate interoceptive domains and subjective affect uniquely.

5. Conclusion

Taken together, this study provides three key insights that advance our understanding of the neural mechanisms that underlie our multifaceted response to threat. First, our findings suggest that the brain systems frequently reported as being engaged in threat learning (Fullana et al., 2016), including the dACC, AIC, vmPFC and subcortical regions, mostly capture the subjective anxious arousal experienced during the threat learning process. While secondly, highlighting the likely influence of different underlying mechanisms, autonomic mediators (SCR) appear to be a subset of regions embedded within this broader circuitry that includes distinct facilitatory and regulatory contributions of vmPFC sub-regions. Thirdly, and by comparison, threat reversal appears to rely on active engagement of default-mode and valence-related regions (including the vmPFC, dorsomedial and posterior midline PFC regions) to appropriately update and revaluate previously conditioned stimuli. These findings may help to inform potential domain-specific alterations that may characterise the pathophysiology of common threat-related disorders, including anxiety disorders.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by a National Health and Medical Research Council of Australia (NHMRC) Project Grant (1161897) to BJH. HSS was supported by an Australian Government Research Training Program (RTP) Scholarship. BJH was supported by a NHMRC Career Development Fellowship (1124472). TDW was supported by NIH R01MH116026 and R01MH076136. The authors thank Cristian Stella and Lisa Incerti for their contributions to data collection, and the participants for their involvement in the study. We acknowledge the facilities, and the scientific and technical assistance of the Australian National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability, at the Melbourne Brain Centre Imaging Unit (MBCIU), The University of Melbourne. The multiband fMRI sequence was generously supported by a research collaboration agreement with CMRR, University of Minnesota and the MP2RAGE works in progress sequence was provided by Siemens Healthineers (Germany) as advanced works in progress package.

References

Alvarez RP, Kirlic N, Misaki M, Bodurka J, Rhudy JL, Paulus MP, Drevets WC, 2015. Increased anterior insula activity in anxious individuals is linked to diminished perceived control. Transl. Psychiatry 5, e591–e599. [PubMed: 26125154]

Apergis-Schoute AM, Gillan CM, Fineberg NA, Fernandez-Egea E, Sahakian BJ, Robbins TW, 2017. Neural basis of impaired safety signaling in obsessive compulsive disorder. Proc. Natl. Acad. Sci. USA 114, 3216–3221. [PubMed: 28265059]

Bach DR, Flandin G, Friston KJ, Dolan RJ , 2009. Time-series analysis for rapid event-related skin conductance responses. J. Neurosci. Methods 184, 224–234. [PubMed: 19686778]

Bach DR , Friston KJ, 2013. Model-based analysis of skin conductance responses: towards causal models in psychophysiology. Psychophysiology 50, 15–22. [PubMed: 23094650]

Bach DR, Friston KJ, Dolan RJ, 2013. An improved algorithm for model-based analysis of evoked skin conductance responses. Biol. Psychol 94, 490–497. [PubMed: 24063955]

Barrett LF, 2017. The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci 12, 1–23. [PubMed: 27798257]

Barrett LF, Bar M, 2009. See it with feeling: affective predictions during object perception. Philos. Trans. R. Soc. B Biol. Sci 364, 1325–1334.

Barrett LF, Simmons WK, 2015. Interoceptive predictions in the brain. Nat. Rev. Neurosci 16, 419–429. [PubMed: 26016744]

Battaglia S, Garofalo S, di Pellegrino G, Starita F, 2020. Revaluing the role of vmPFC in the acquisition of Pavlovian threat conditioning in humans. J. Neurosci 40, 8491–8500. [PubMed: 33020217]

Bittencourt A, Nakamura-Palacios E, Mauad H, Tufik S, Schenberg L, 2005. Organization of electrically and chemically evoked defensive behaviors within the deeper collicular layers as compared to the periaqueductal gray matter of the rat. Neuroscience 133, 873–892. [PubMed: 15916856]

Borgomaneri S, Battaglia S, Garofalo S, Tortora F, Avenanti A, di Pellegrino G, 2020. State-dependent TMS over prefrontal cortex disrupts fear-memory reconsolidation and prevents the return of fear. Curr. Biol 30, 3672–3679 e3674. [PubMed: 32735813]

Bradley MM, Lang PJ, 1994. Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry 25, 49–59. [PubMed: 7962581]

Cai Y, Hofstetter S, van der Zwaag W, Zuiderbaan W, Dumoulin SO, 2021. Individualized cognitive neuroscience needs 7T: comparing numerosity maps at 3T and 7T MRI. Neuroimage 237, 118184. [PubMed: 34023448]

Cha J, Greenberg T, Carlson JM, DeDora DJ, Hajcak G, Mujica-Parodi LR, 2014. Circuit-wide structural and functional measures predict ventromedial prefrontal cortex fear generalization: implications for generalized anxiety disorder. J. Neurosci 34, 4043–4054. [PubMed: 24623781]

Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD, 2015. A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol. 13, e1002180–e1002208. [PubMed: 26098873]

Critchley HD, Elliott R, Mathias CJ, Dolan RJ, 2000. Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study. J. Neurosci 20, 3033–3040. [PubMed: 10751455]

Cunningham WA, Raye CL, Johnson MK, 2004. Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. J. Cogn. Neurosci 16, 1717–1729. [PubMed: 15701224]

den Ouden HEM, Kok P, de Lange FP, 2012. How prediction errors shape perception, attention, and motivation. Front. Psychol 3, 1–12. [PubMed: 22279440]

Eisenbarth H, Chang LJ, Wager TD, 2016. Multivariate brain prediction of heart rate and skin conductance responses to social threat. J. Neurosci 36, 11987–11998. [PubMed: 27881783]

Ernst TM, Brol AE, Gratz M, Ritter C, Bingel U, Schlamann M, Maderwald S, Quick HH, Merz CJ, Timmann D, 2019. The cerebellum is involved in processing of predictions and prediction errors in a fear conditioning paradigm. eLife 8, e46831–e46857. [PubMed: 31464686]

Fanselow MS, Antoine Depaulis RB, 1991. The midbrain periaqueductal gray as a coordinator of action in response to fear and anxiety. In: The Midbrain Periaqueductal Gray Matter. Springer, pp. 151–173.

Fullana MA, Harrison BJ, Soriano-Mas C, Vervliet B, Cardoner N, Àvila-Parcet A, Radua J, 2016. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. Mol. Psychiatry 21, 500–508. [PubMed: 26122585]

Garfinkel SN, Seth AK, Barrett AB, Suzuki K, Critchley HD, 2015. Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. Biol. Psychol 104, 65–74. [PubMed: 25451381]

Glover GH, Li TQ, Ress D, 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med 44, 162–167. [PubMed: 10893535]

Harrison BJ, Fullana MA, Soriano-Mas C, Via E, Pujol J, Martínez-Zalacaín I, Tinoco-Gonzalez D, Davey CG, López-Solà M, Pérez SV, Menchón JM, Cardoner N, 2015. A neural mediator of human anxiety sensitivity. Hum. Brain Mapp 36, 3950–3958. [PubMed: 26147233]

Harrison BJ, Fullana MA, Via E, Soriano-Mas C, Vervliet B, Martinez-Zalacain I, Pujol J, Davey CG, Kircher T, Straube B, Cardoner N, 2017. Human ventromedial prefrontal cortex and the positive affective processing of safety signals. Neuroimage 152, 12–18. [PubMed: 28254509]

Hermans EJ, Henckens MJAG, Roelofs K, Fernández G, 2013. Fear bradycardia and activation of the human periaqueductal grey. Neuroimage 66, 278–287. [PubMed: 23110885]

Hodgson R, Rachman S, 1974. II. Desynchrony in measures of fear. Behav. Res. Ther 12, 319–326. [PubMed: 4155622]

Jovanovic T, Kazama A, Bachevalier J, Davis M, 2012. Impaired safety signal learning may be a biomarker of PTSD. Neuropharmacology 62, 695–704. [PubMed: 21377482]

Kalin NH, Shelton SE, Davidson RJ, 2004. The role of the central nucleus of the amygdala in mediating fear and anxiety in the primate. J. Neurosci 24, 5506–5515. [PubMed: 15201323]

Kasper L, Bollmann S, Diaconescu AO, Hutton C, Heinzle J, Iglesias S, Hauser TU, Sebold M, Manjaly ZM, Pruessmann KP, Stephan KE, 2017. The PhysIO Toolbox for modeling physiological noise in fMRI data. J. Neurosci. Methods 276, 56–72. [PubMed: 27832957]

Keifer OP, Hurt RC, Ressler KJ, Marvar PJ, 2015. The physiology of fear: reconceptualizing the role of the central amygdala in fear learning. Physiology 30, 389–401. [PubMed: 26328883]

Kim SH, Hamann S, 2007. Neural correlates of positive and negative emotion regulation. J. Cogn. Neurosci 19, 776–798. [PubMed: 17488204]

Knight DC, Nguyen HT, Bandettini PA, 2005. The role of the human amygdala in the production of conditioned fear responses. Neuroimage 26, 1193–1200. [PubMed: 15961053]

Koban L, Jepma M, López-Solà M, Wager TD, 2019. Different brain networks mediate the effects of social and conditioned expectations on pain. Nat. Commun 10, 4096–4109. [PubMed: 31506426]

Labar KS, Gatenby JC , Gore JC, LeDoux JE, Phelps EA, 1998. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron 20, 937–945. [PubMed: 9620698]

LeDoux JE, Brown R, 2017. A higher-order theory of emotional consciousness. Proc. Natl. Acad. Sci E2016–E2025. [PubMed: 28202735]

LeDoux JE, Pine DS, 2016. Using neuroscience to help understand fear and anxiety: a two-system framework. Am. J. Psychiatry 173, 1083–1093. [PubMed: 27609244]

Lewis PA, Critchley H, Rotshtein P, Dolan RJ, 2007. Neural correlates of processing valence and arousal in affective words. Cereb. Cortex 17, 742–748. [PubMed: 16699082]

Linnman C , Zeidan MA, Pitman RK, Milad MR, 2012. Resting cerebral metabolism correlates with skin conductance and functional brain activation during fear conditioning. Biol. Psychol 89, 450–459. [PubMed: 22207247]

Lonsdorf TB , Menz MM , Andreatta M, Fullana MA, Golkar A, Haaker J, Heitland I, Hermann A, Kuhn M, Kruse O, 2017. Don't fear 'fear conditioning': methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. Neurosci. Biobehav. Rev 77, 247–285. [PubMed: 28263758]

MacKinnon DP, Krull JL, Lockwood CM, 2000. Equivalence of the mediation, confounding and suppression effect. Prev. Sci 1, 173–181. [PubMed: 11523746]

MacNamara A, Rabinak CA, Fitzgerald DA, Zhou XJ, Shankman SA, Milad MR, Phan KL, 2015. Neural correlates of individual differences in fear learning. Behav. Brain Res 287, 34–41. [PubMed: 25819422]

Marin MF, Hammoud MZ, Klumpp H, Simon NM, Milad MR, 2020. Multimodal categorical and dimensional approaches to understanding threat conditioning and its extinction in individuals with anxiety disorders. JAMA Psychiatry 77, 618–627. [PubMed: 32022832]

Marques JP, Kober T, Krueger G, van der Zwaag W , Van de Moortele P-F, Gruetter R, 2010. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. Neuroimage 49, 1271–1281. [PubMed: 19819338]

Marstaller L, Burianová H, Reutens DC, 2017. Adaptive contextualization: a new role for the default mode network in affective learning. Hum. Brain Mapp 38, 1082–1091. [PubMed: 27767246]

Milad MR, Quirk GJ, Pitman RK, Orr SP, Fischl B, Rauch SL, 2007. A role for the human dorsal anterior cingulate cortex in fear expression. Biol. Psychiatry 62, 1191–1194. [PubMed: 17707349]

Mobbs D, Petrovic P, Marchant JL, Hassabis D, Weiskopf N, Seymour B, Dolan RJ, Frith CD, 2007. When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. Science 317, 1079–1083. [PubMed: 17717184]

Morris LS, Kundu P, Costi S, Collins A , Schneider M, Verma G, Balchandani P, Murrough JW, 2019. Ultra-high field MRI reveals mood-related circuit disturbances in depression: a comparison between 3-Tesla and 7-Tesla. Transl Psychiatry 9, 94–105. [PubMed: 30770788]

Nagai Y, Critchley HD, Featherstone E, Trimble MR, Dolan RJ, 2004. Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: a physiological account of a “default mode” of brain function. Neuroimage 22, 243–251. [PubMed: 15110014]

Patterson JC, Ungerleider LG, Bandettini PA, 2002. Task-independent functional brain activity correlation with skin conductance changes: an fMRI study. Neuroimage 17, 1797–1806. [PubMed: 12498753]

Paulus MP, Stein MB, 2006. An insular view of anxiety. Biol. Psychiatry 60, 383–387. [PubMed: 16780813]

Phan KL, Taylor SF, Welsh RC, Decker LR, Noll DC, Nichols TE, Britton JC, Liberzon I, 2003. Activation of the medial prefrontal cortex and extended amygdala by individual ratings of emotional arousal: a fMRI study. Biol. Psychiatry 53, 211–215. [PubMed: 12559653]

Phelps EA, Delgado MR, Nearing KI, LeDoux JE , 2004. Extinction learning in humans: role of the amygdala and vmPFC. Neuron 43, 897–905. [PubMed: 15363399]

Phelps EA, O’Connor KJ, Gatenby JC, Gore JC, Grillon C, Davis M, 2001. Activation of the left amygdala to a cognitive representation of fear. Nat. Neurosci 4, 437–441. [PubMed: 11276236]

Ryan KM, Neumann DL, Waters AM, 2021. Does the assessment of different combinations of within-phase subjective measures influence electrodermal responding and between-phase subjective ratings during fear conditioning and extinction experiments? Biol. Psychol 162, 108085–108098. [PubMed: 33775735]

Satpute AB, Wager TD, Cohen-Adad J, Bianciardi M, Choi J-K, Buhle JT, Wald LL, Barrett LF, 2013. Identification of discrete functional subregions of the human periaqueductal gray. Proc. Natl. Acad. Sci 110, 17101–17106. [PubMed: 24082116]

Savage HS, Davey CG, Fullana MA, Harrison BJ, 2020a. Clarifying the neural substrates of threat and safety reversal learning in humans. Neuroimage 207, 116427. [PubMed: 31801684]

Savage HS, Davey CG, Fullana MA, Harrison BJ, 2020b. Threat and safety reversal learning in social anxiety disorder-an fMRI study. J. Anxiety Disord 76, 102321. [PubMed: 33099070]

Schiller D, Delgado MR, 2010. Overlapping neural systems mediating extinction, reversal and regulation of fear. Trends Cogn. Sci 14, 268–276. [PubMed: 20493762]

Schiller D, Levy I, Niv Y, LeDoux JE , Phelps EA, 2008. From fear to safety and back: reversal of fear in the human brain. J. Neurosci 28, 11517–11525. [PubMed: 18987188]

Setsompop K, Gagoski BA, Polimeni JR, Witzel T, Wedeen VJ, Wald LL, 2012. Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. Magn. Reson. Med 67, 1210–1224. [PubMed: 21858868]

Tak S, Noh J, Cheong C, Zeidman P, Razi A, Penny WD, Friston KJ, 2018. A validation of dynamic causal modelling for 7T fMRI. J. Neurosci. Methods 305, 36–45. [PubMed: 29758234]

Taschereau-Dumouchel V, Kawato M , Lau H, 2019. Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. Mol. Psychiatry 25, 2342–2354. [PubMed: 31659269]

Tashjian SM, Zbozinek TD, Mobbs D, 2021. A decision architecture for safety computations. Trends Cogn. Sci 25 (5), 342–354. [PubMed: 33674206]

Terracciano A, McCrae RR, Hagemann D, Costa PT, 2003. Individual difference variables, affective differentiation, and the structures of affect. J. Personal 71, 669–703.

VanderWeele TJ, 2016. Mediation analysis: a practitioner's guide. Annu. Rev. Public Health 37, 17–32. [PubMed: 26653405]

Via E, Fullana MA, Goldberg X, Tinoco-Gonzalez D, Martinez-Zalacain I, Soriano–Mas C, Davey CG, Menchon JM, Straube B , Kircher T, Pujol J, Cardoner N, Harrison BJ, 2018. Ventromedial prefrontal cortex activity and pathological worry in generalised anxiety disorder. Br. J. Psychiatry 213, 437–443. [PubMed: 29739481]

Wager T, Waugh C, Lindquist M, Noll D, Fredrickson B, Taylor S, 2009. Brain mediators of cardiovascular responses to social threat. Neuroimage 47, 821–835. [PubMed: 19465137]

Wang YC, Bianciardi M, Chanes L, Satpute AB , 2020. Ultra high field fMRI of human superior colliculi activity during affective visual processing. Sci. Rep 10, 1–7. [PubMed: 31913322]

Wilke M, 2012. An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. Neuroimage 59, 2062–2072. [PubMed: 22036679]

Zhang S, Hu S, Chao HH, Luo X, Farr OM, Li CSR, 2012. Cerebral correlates of skin conductance responses in a cognitive task. Neuroimage 62, 1489–1498. [PubMed: 22634217]

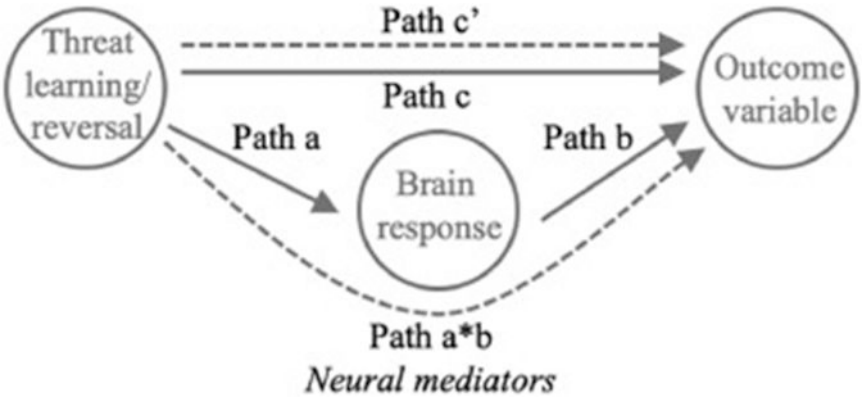

Fig. 1.
Schematic of a standard three variable path-modelling framework.

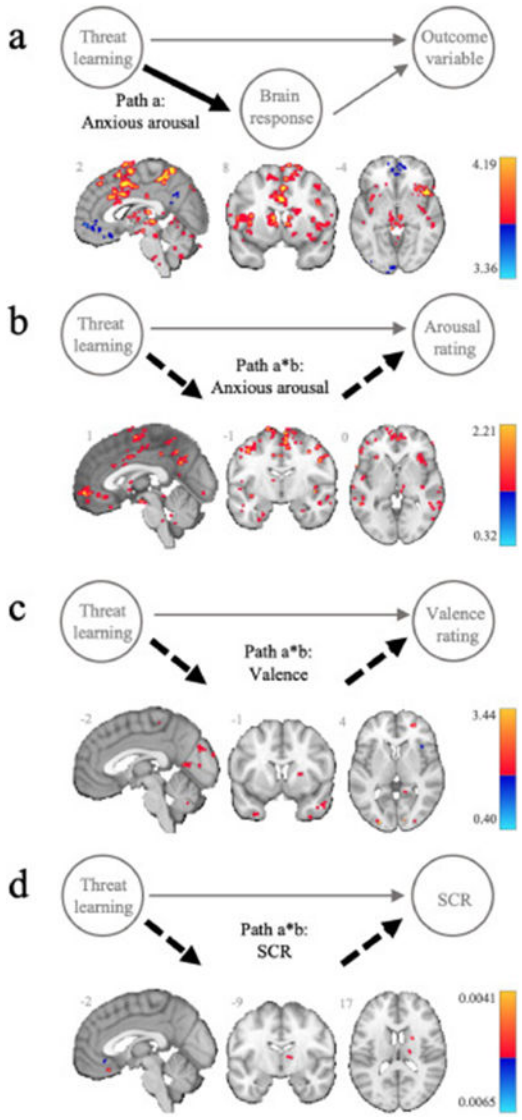

Fig. 2.

Threat learning and neural mediators of subjective and autonomic responses. (a) Representative task modulation of brain activity (*Path a*) during threat learning (CS+ > CS−), taken from the anxious arousal mediation model. Threat learning mediation effects (*Path a*b*) for subjective ratings of (b) anxious arousal and (c) valence, and (d) skin conductance responses (SCR). Positive (yellow/red) and negative (blue) effects are scaled by effect size and presented at an FDR corrected threshold ($q < 0.05$, cluster extent = 5).

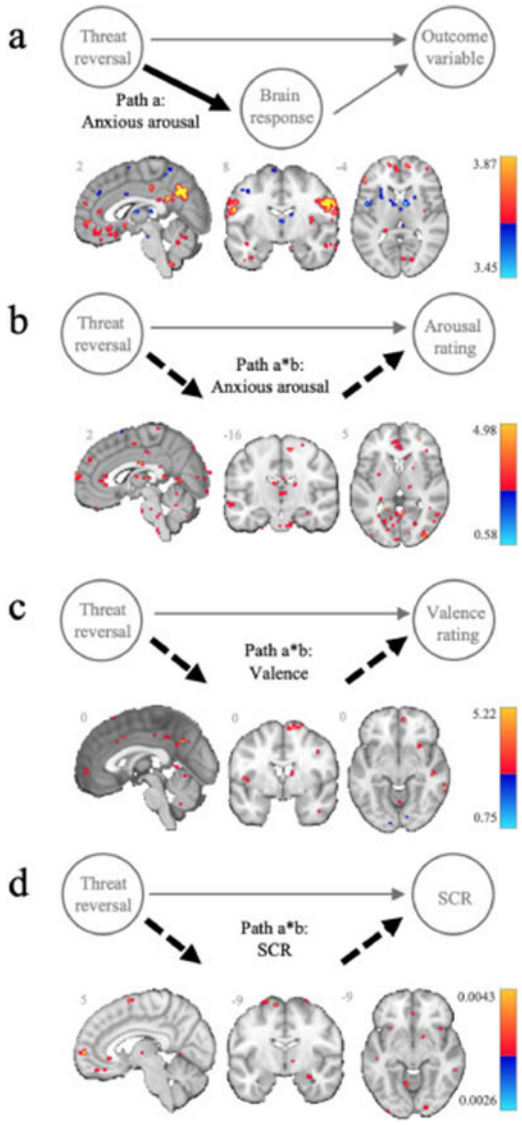

Fig. 3.

Threat reversal and neural mediators of subjective and autonomic responses. (a) Representative task modulation of brain activity (Path a) during threat reversal (new CS− > CS+), taken from the anxious arousal mediation model. Threat reversal mediation effects (Path a*b) for subjective ratings of (b) anxious arousal and (c) valence, and (d) skin conductance responses (SCR). Positive (yellow/red) and negative (blue) effects are scaled by effect size and presented at an FDR corrected threshold ($q < 0.05$, cluster extent = 5).

Table 1

Threat learning and neural mediators of subjective and autonomic responses.

A note on interpretation: Interpreting the mediation effects, depends on which of the 4 relationship types we observe: (i) positive to threat, positive to outcome measure, (ii) negative to threat, negative to outcome measure, (iii) positive to threat, negative to outcome measure, (iv) negative to threat, positive to outcome measure. Relationship (i) is the common positive mediator effect. Relationship (ii) represents a negative mediator, whereby greater deactivation to the threat cue mediates higher anxious arousal ratings, greater SCR amplitude, or more negative valence ratings. Relationships (iii) and (iv) can be called suppressor effects (MacKinnon et al., 2000). In relationship (iv), for example, greater deactivation to threat, decreases ratings of anxious arousal, SCR amplitude, or increases ratings of positive valence.			***Task modulation of brain activity (Path a, X-M)***			
			Positive (CS+> CS−) • midline cingulo-frontal cortex (incl. dACC, pre-SMA) • dorsomedial prefrontal cortex (PFC) • bilateral AIC • central amygdala (medial sector) • anterior caudate nucleus • posterior putamen • ventrolateral and lateral posterior thalamus • hippocampus (CA1, 2) • superior colliculus • red nucleus • PAG • the cerebellum (sixth and eighth lobes and vermis).		**Negative** (CS-> CS+) • distinct clusters across the vmPFC • bilateral lateral OFC (~BA8/9) • posterior cingulate cortex extending to precuneus • temporoparietal junction • bilateral angular gyrus • lateral visual association areas	
Brain mediators of subjective responses and SCR during threat learning (Path a*b; X-M-Y)						
			Task positive		**Task negative**	
Outcome variables			**Positive Mediators (i)**	**Negative Mediators (iii)**	**Positive Mediators (iv)**	**Negative Mediators (ii)**
Anxious arousal	CS+ Mean rating ± SD: 2.81±1.2 Range: 1–5 CS− Mean rating ± SD: 1.46±0.9 Range: 1–5	$t_{25} = 4.59$ $p < 0.001$ Cohen's $d = 1.1$	cingulo-frontal cortex (incl. dACC, pre- SMA, dorsomedial PFC); bilateral AIC; anterior and posterior caudate; posterior putamen; ventral striatum; hippocampus (CA1, CA2); mediodorsal, intralaminar and pulvinar thalamus; red nucleus; dorsomedial and dorsolateral PAG; superior and inferior colliculus; cerebellum (sixth, seventh and eighth lobes).		anterior vmPFC (~BA10/11) extending to the right lateral OFC (~BA8/9); temporoparietal junction; fusiform gyrus; angular gyrus.	left visual association area.
Valence	CS+ Mean rating ± SD: 1.65±0.9 Range: 1–5	$t_{25} = -8.94$ $p < 0.001$ Cohen's $d = 2.4$	posterior caudate; anterior and posterior putamen; cerebellum (sixth, seventh and eighth lobes).	right AIC.	primary visual cortex; visual association areas.	left lateral OFC (~BA8); right superior temporal gyrus;

Brain mediators of subjective responses and SCR during threat learning (Path a*b; X-M-Y)

			Task positive		Task negative	
Outcome variables			**Positive Mediators (i)**	**Negative Mediators (iii)**	**Positive Mediators (iv)**	**Negative Mediators (ii)**
	CS− Mean rating ± SD: 4.04±1.1 Range: 2–5					angular gyrus; superior precuneus.
SCR	CS+ Mean response ± SD: 0.14μS ± 0.1 Range: 0.02–0.48 μS CS− Mean response ± SD: 0.09μS ± 0.06 Range: 0.02–0.25 μS	t_{24} = 3.44 p = 0.002 Cohen's d = 0.6	small cluster of the pre-SMA; anterior and posterior caudate; posterior putamen; ventrolateral thalamus; cerebellum (sixth, seventh and eighth lobes).		vmPFC sub-region with a distinct ventral focus (~BA11); the right lateral OFC (~BA11); right primary visual cortex.	sub-genual ACC (~BA25); tip of the dorsal temporal pole; bilateral temporoparietal junction; supramarginal gyrus; precuneus; right visual association area.

AIC, anterior insular cortex; BA, Brodmann's area; CA, cornu ammonis; CS+, conditioned threat cue; CS−, conditioned safety cue; d, Cohen's d for paired *t*-test; dACC, dorsal anterior cingulate cortex; OFC, orbitofrontal cortex; PAG, periaqueductal grey; PFC, prefrontal cortex; pre-SMA, pre-Supplementary Motor Area; SCR, skin conductance response; S.D, standard deviation; vmPFC, ventromedial prefrontal cortex; μS, micro Siemens.

Table 2

Threat reversal and neural mediators of subjective and autonomic responses.

A note on interpretation: Interpreting the mediation effects, depends on which of the 4 relationship types we observe:
(i) positive to new safety, positive to outcome measure,
(ii) negative to new safety, negative to outcome measure,
(iii) positive to new safety, negative to outcome measure,
(iv) negative to new safety, positive to outcome measure.
Relationship (i) is the common positive mediator effect.
Relationship (ii) represents a negative mediator, whereby greater deactivation to the new safety cue mediates lower anxious arousal ratings, smaller SCR amplitude, or more positive valence ratings.
Relationships (iii) and (iv) can be called suppressor effects (MacKinnon et al., 2000). In relationship (iv), for example, greater deactivation to new safety, higher anxious arousal ratings, greater SCR amplitude, or increases ratings of negative valence.

Task modulation of brain activity (Path a, X-M)

Positive (new CS−> CS+)

- dorsal vmPFC (BA10)
- lateral OFC (~BA8/9)
- posterior cingulate cortex
- medial precuneus
- angular gyrus
- ventral striatum
- hippocampal-den
- tate cerebellum (seventh and eight lobes)
- primary visual cortex visual association areas

Negative (CS+ > new CS−)

- mid-cingulate cortex
- pre-SMA
- bilateral AIC
- superior lateral precuneus
- anterior and posterior caudate
- anterior putamen
- mediodorsal, lateral posterior and intralaminar thalamus
- cerebellum (sixth lobe)

Brain mediators of subjective responses and SCR during threat learning (Path a*b; X-M-Y)

Outcome variables			Task positive		Task negative	
			Positive Mediators (i)	**Negative Mediators (iii)**	**Positive Mediators (iv)**	**Negative Mediators (ii)**
Anxious arousal	new CS− Mean rating ± SD: 1.50±0.7 Range: 1–3 CS+ Mean rating ± SD: 2.81±1.2 Range: 1–5	$t_{25} = -5.93$ $p < 0.001$ Cohen's $d = 1.3$	dorsal vmPFC (~BA10); a sub-region of the vmPFC on the genu of the dACC; supramarginal gyrus; medial temporal lobe; temporal pole; fusiform gyrus; right nucleus accumbens; ventral striatum; cerebellum (seventh and eighth lobes); primary visual cortex; visual association areas.	superior dmPFC (~BA6).	mid cingulate cortex; pre-SMA; anterior and posterior caudate; posterior putamen; mediodorsal, intralaminar and pulvinar thalamus; red nucleus; rostral and caudal pontine regions; cerebellum (sixth lobe).	
Valence	new CS− Mean rating ± SD: 3.5 ± 1.0 Range: 1–5 CS+ Mean rating ± SD: 1.65±0.9 Range: 1–5	$t_{25} = 8.14$ $p < 0.001$ Cohen's $d = 1.9$	dorsal vmPFC (~BA10); lateral PFC (~BA8); superior medial PFC (~BA6); angular gyrus; left temporoparietal junction; right fusiform gyrus; cerebellum (seventh and eighth lobes).	inferior parietal lobule; primary visual cortex; visual association areas.	mid- and posterior cingulate cortex; AIC; right anterior caudate; mediodorsal extending to anterior ventral thalamus; cerebellum (sixth lobe).	
SCR	new CS− Mean response ± SD: 0.08μS ± 0.06 Range: 0.02–0.21 μS CS+	$t_{24} = -3.62$ $p = 0.001$ Cohen's $d = 0.7$	a lateral region of the vmPFC (~BA10); ventral vmPFC (~BA11); the pregenual ACC (~BA25); dmPFC (~BA8); fusiform gyrus; hippocampus (CA1); cerebellum (dentate nucleus).		mid cingulate cortex; SMA; AIC; intralaminar thalamus; anterior pallidum; cerebellum (sixth lobe).	

Brain mediators of subjective responses and SCR during threat learning (Path a*b; X-M-Y)				
	Task positive		Task negative	
Outcome variables	**Positive Mediators (i)**	**Negative Mediators (iii)**	**Positive Mediators (iv)**	**Negative Mediators (ii)**
Mean response ± SD: 0.14μS ± 0.1 Range: 0.02–0.48 μS				

AIC, anterior insular cortex; BA, Brodmann's area; CA, cornu ammonis; CS+, conditioned threat cue; CS−, conditioned safety cue; d, Cohen's d for paired *t*-test; dACC, dorsal anterior cingulate cortex; OFC, orbitofrontal cortex; PAG, periaqueductal grey; PFC, prefrontal cortex; pre-SMA, pre-Supplementary Motor Area; SCR, skin conductance response; SMA, Supplementary Motor Area; S.D, standard deviation; vmPFC, ventromedial prefrontal cortex; μS, micro Siemens.