Beyond fear centers – a distributed fMRI-based neuromarker for the subjective experience of fear

Feng Zhou[1,2], Weihua Zhao[1], Ziyu Qi[1], Yayuan Geng[1], Shuxia Yao[1], Keith M. Kendrick[1], Tor D. Wager[2*], Benjamin Becker[1*]

[1]Clinical Hospital of Chengdu Brain Science Institute, MOE Key Laboratory for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China

[2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

[*]Correspondence (Tor D. Wager and Benjamin Becker contributed equally to this work)

Tor D. Wager

Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03784, USA

Email: Tor.D.Wager@Dartmouth.edu; Telephone: (303) 895-8739

Benjamin Becker

Clinical Hospital of Chengdu Brain Science Institute, MOE Key Laboratory for Neuroinformation

University of Electronic Science and Technology of China, Chengdu 611731, China.

Email: ben_becker@gmx.de; Telephone: +86 2861 830 811

Abstract

The specific neural systems underlying the subjective feeling of fear remain vigorously debated in affective neuroscience. Here, we combined functional MRI with machine learning to identify and evaluate a sensitive and generalizable neural signature predictive of the momentary self-reported subjective fear experience across discovery (n=67), validation (n=20) and generalization (n=31) cohorts. We systematically demonstrate that accurate fear prediction crucially requires distributed brain systems, with important contributions from cortical (e.g., prefrontal, midcingulate and insular cortices) and subcortical (e.g., thalamus, periaqueductal gray, basal forebrain and amygdala) regions. We further demonstrate that the neural representation of subjective fear is distinguishable from the representation of conditioned threat and general negative affect. Overall, our findings suggest that subjective fear, which exhibits distinct neural representation with some other aversive states, is encoded in distributed systems rather than isolated 'fear centers'. This signature provides a neuromarker for monitoring fear-related neuropathology and evaluating novel treatments targeting pathological fear.

Introduction

Fear is probably the most studied emotion during the last decades, yet despite considerable advances in animal models and human neuroimaging research, vigorous debates on how to define and investigate fear and its facets continue [1-4]. When we talk about fear in everyday life, we primarily refer to the subjective feeling of being afraid. However, in psychological and neuroscientific conceptualizations, fear also describes defensive behaviors, such as freezing, and peripheral physiological changes that accompany such behaviors [5,6].

The neural basis of 'fear', or threat behaviors, has been extensively mapped in animal models using Pavlovian conditioning and predator exposure protocols [3,7]. These models provide compelling evidence for a pivotal role of subcortical systems, particularly the central extended amygdala, as well as the hypothalamus and periaqueductal gray (PAG), in mediating threat detection and defensive responses [7-13]. However, the subjective emotional experience of fear remains ultimately inaccessible in animal models, and recent conceptual frameworks argue that the evolutionarily conserved defensive survival circuits that account for the behavioral and physiological responses to threats might be distinct from those underlying the subjective experience of fear [3,6,14-16]. The differentiation between the defensive response and the subjective experience of fear has critical implications for translational research on pathological fear [6], given that animal models primarily evaluate novel treatments by means of effects on physiological and behavioral defensive threat reactivity [17], whereas feelings of exaggerated fear or anxiety represent the primary clinical outcome and reason for patients to seek treatment [18].

In humans, lesion and functional magnetic resonance imaging (fMRI) approaches have been employed to determine the specific brain systems that underlie the subjective feeling of fear. Early studies in a patient with focal amygdala lesions demonstrated impairments in fear-related processes, including recognition and experience of fear [19,20], which contributed to an amygdala-centric fear perspective. However, subsequent studies reported variable fear-related functional consequences in patients with focal amygdala lesions [21]. For example, an examination of homozygotic twins with similar amygdala lesions revealed that one sister demonstrated the hypothesized fear-specific deficits, while the other sister performed entirely within the normal range [22]. fMRI studies in healthy subjects suggest that it is time to move beyond an amygdala-centric fear perspective and demonstrate that stimuli that evoke subjective feelings of fear elicit activation not only in the amygdala but also PAG, hypothalamic and frontal regions [23-26]. However, the conventional fMRI approach applied in these studies has been limited. In particular, it is designed to permit the inference of whether a single brain region (e.g., a voxel) is activated conditionally on a stimulus, but does not allow reverse inferences about 'fear' states given brain activity [27]. Furthermore, conventional analyses do not model joint activity across brain regions working together to underpin fear experience [27], and stimulus-induced activation changes in single brain regions typically have only modest effect sizes [28-30]. These issues raise the question of whether isolated regions can provide sufficient brain-level descriptions of complex mental processes such as the subjective feeling of fear.

The perspective of an isolated fear center in the brain has additionally been challenged by recent appraisal [31] and constructionist [2] theories of emotion which suggest that emotional

experiences result from interactions between multiple systems including core affect, sensory, memory, motor, and cognitive systems [32], and by the two-system model suggesting that interactions of subcortical defensive systems with prefrontal regions engaged in consciousness are critical to establish a neural representation of the subjective fear experience [6].

To provide sufficient and process-specific brain-level descriptions of mental processes recent studies have combined fMRI with machine learning-based multivariate pattern analysis (MVPA). This approach can capture information at much finer spatial scales [33] and provide considerably larger effect sizes in brain-outcome associations [29] thus allowing the development of sensitive and specific brain signatures of mental processes [30,34-36], including acquired defensive responses [37] and subjective emotional states [38,39]. Moreover, an initial MVPA study has revealed promising findings suggesting that offline categorical fear ratings collected before fMRI are associated with a neural signature that is independent of online autonomic arousal indices acquired during fMRI [40] (henceforward referred to as V.T. decoder in this study for convenience). The MVPA approach additionally allows functional separation of mental processes based on population coding [29], despite overlapping univariate activation [35,41,42] and thus offers an opportunity to determine process-specific neural representations of (often) concurrent fear-related processes, such as the experience of fear and defensive responses.

Here, in the context of ongoing debates about the neural representations of fear, we capitalize on recent advances in MVPA-based neural decoding techniques to determine whether (1) it is possible to determine a sensitive and generalizable neural representation of the subjective momentary fear experience on the population level, (2) this neural representation can predict

momentary (trial-wise) fear experience on the individual level, (3) the neural representation in isolated systems such as the amygdala or 'cortical consciousness network' is sufficient to capture the subjective experience of fear, and (4) the neural representation of the momentary fear experience is distinct from the representations of the conditioned defensive threat response and unspecific aversive states. More specifically, we employed a support vector regression (SVR) algorithm in healthy participants (n = 67) to identify the brain signature that predicted the intensity of trial-by-trial rated subjective experiences of fear elicited by fear-evoking pictures ranging from low to high fear induction (Fig. 1A). The performance of the established visually induced fear signature (VIFS) was evaluated in (a) an independent validation cohort (n = 20), who underwent a similar but not identical fear induction paradigm (Fig. 1B) as well as a generalization cohort (n = 31) from a previous study that employed a different fear induction paradigm and MRI system [40] and (b) a comparison with the V.T. decoder (Fig. 1C). To extend the perspective from a population to an individual level we tested whether the VIFS can predict trial-wise fear experience for each subject in discovery and validation cohorts separately. We further systematically identified brain regions that were associated with (forward model) and predictive of (backward model) subjective fear experience and examined to what extent single brain systems or networks can capture subjective fear experience. Moreover, to determine the functional specificity of the neural fear experience signature we compared the spatial and functional similarities between the VIFS with the signature of conditioned defensive threat responses [37] and unspecific negative emotional experience [30], respectively (Fig. 1C). Together this systematic evaluation can advance ongoing debates on how the brain constructs subjective fear, whether the neural mechanisms of

the conscious experience of fear are distinct from defensive responses elicited by conditioning [3,6,14-16] or unspecific aversive emotional experience, and ultimately promote valid estimates of effect size with high clinical significance for the evaluation of novel treatments that specifically target subjective fear experience.

Results

A brain signature sensitive to predict visually induced subjective experience of fear

We applied SVR to identify a whole-brain signature of fMRI activation that predicted the intensity of self-reported fear ratings during observation of fear-evoking pictures in the discovery cohort (Fig. 2A). To evaluate the performance of the visually induced fear signature (VIFS), we applied the VIFS to data from test subjects in both discovery (n = 67; 10×10-fold cross-validated, see Methods for details) and validation (n = 20) cohorts to calculate the VIFS pattern expressions for individual participants' activation maps for each of 5 levels of reported fear. The developed VIFS accurately predicted ratings of reported fear in both discovery and independent validation cohorts. Specifically, for individual participants in the discovery cohort the average within-subject correlation between predicted and actual fear ratings was r = 0.89±0.01 (standard error (SE)), the mean explained variance was 72.5±2.1%, the average root mean squared error (RMSE) was 1.38±0.08 and the overall (between- and within-subjects) prediction-outcome correlation coefficient was 0.57 (bootstrapped 95% confidence interval (CI; averaged across 10 repetitions) = [0.49, 0.63]) (Fig. 2B). Testing the VIFS model identified in the discovery cohort, with no further model fitting, in the validation cohort (Fig. 1B) revealed comparably high prediction-outcome

correlations (within-subject r = 0.87±0.02; mean explained variance score = 68.3±5.6%; average RMSE = 1.40±0.14; overall prediction-outcome r = 0.59, 95% CI = [0.48, 0.69], permutation test one-tailed P < 0.001; Fig. 2C), indicating a sensitive and generalizable subjective fear signature on the neural level (see also generalization and benchmarking of the VIFS below). To further determine the sensitivity of the VIFS to predict levels of subjective fear experience a two-alternative forced-choice test was applied, comparing pairs of activation maps within each subject and choosing the one with higher VIFS response as more fearful. The VIFS response accurately classified high (average of rating 4 and 5) versus moderate (rating 3) and moderate versus low (average of rating 1 and 2) fear in both cohorts with 88-93% accuracy (Cohen's d: 1.18-1.40), and high versus low with 100% accuracy in both cohorts (Cohen's d: 2.20-2.58) (Fig. 1B, C; see also Table 1 for a detailed summary of classification performance). Moreover, the VIFS response could distinguish each successive pair of fear rating levels (e.g., rating 4 versus 5) with ≥ 80% accuracy, which were significantly better than chance level (50%; P < 0.001; except ratings of '1' versus '2' in the validation cohort) (Fig. 2B, C).

Retraining the decoder excluding the visual cortex revealed high prediction accuracies, suggesting that although the fear-predictive signals might be partly embedded in the visual cortex the contribution of visual cortical patterns is small (see Supplementary Results and Supplementary Fig. 1 for details). In addition, we applied the VIFS to time series data in the discovery (10×10-fold cross-validated) and validation datasets to determine the specificity of the visually induced fear pattern with respect to confrontation with imminent threat (rather than anticipation or cognitive evaluation). Examining each time point following stimulus onset revealed that VIFS

response began approximately 4s following picture onset and grew with increasing levels of reported fear during approximately 6-12s (Fig. 2D, E). These findings validate the adequacy of the hemodynamic response model and confirmed that the VIFS was specific to brain activity during threat exposure, as opposed to threat anticipation (pre-stimulus) or cognitive evaluation (response reporting).

Generalization and benchmarking of VIFS performance

An important feature of population-level neural signatures is that their performance can be evaluated in new datasets and paradigms, although prediction across cohorts, paradigms and different MRI systems has been challenging. Taschereau-Dumouchel, et al. [40] developed a neural decoder which predicted the general subjective fear of different animal categories (assessed before fMRI) and the authors shared the dataset used for training their model – which we term the 'generalization dataset' here – allowing us to compare the performance of the VIFS with the V.T. decoder on the discovery, validation and generalization cohorts. We found that the VIFS predicted all three datasets well while the V.T. decoder only performed well on its training dataset (Fig. 2F, G; Table 1; see also Supplementary results for details), indicating a robust generalization and high sensitivity of the VIFS to predict fear experience across populations, fear induction paradigms and MRI systems.

Within-subject trial-wise prediction

The feeling of fear is a momentary, highly subjective and individually constructed state [2,4] and thus a key question is to what extent the population-level model (i.e., the VIFS), which is a statistical summary of a highly variable set of instances, can predict momentary (trial-wise) fear on the individual level. To this end we performed single trial analyses to obtain a beta map for each trial for each subject in both discovery and validation cohorts (see Methods for details). The VIFS was next applied to these beta maps to calculate the pattern expressions which were further correlated with the true ratings for each subject separately. We found that the VIFS could significantly predict trial-by-trial ratings for 61 out of 67 subjects in the discovery cohort (cross-validated) and for 16 out of 20 subjects in the validation cohort. The mean prediction-outcome correlations were 0.38±0.01 and 0.40±0.03 for the discovery and validation cohorts, respectively. Our findings suggest that although fear experience differs between individuals [2,4] the VIFS could predict the level of momentary fear experience on the individual level in a large population.

Subjective fear is associated with and predicted by distributed neural systems

We systematically determined individual brain regions that were associated with subjective fear ratings and that provided consistent and reliable contributions to the whole-brain fear decoding model using different analytic strategies. We first examined the conventional univariate linear parametric effect of fear ratings, i.e., voxels that increased (yellow in Fig. 3A) or decreased (blue in Fig. 3A) linearly with within-subject fear ratings across trials. Subjective fear was associated with activation in a broad set of cortical and subcortical regions, including increased activation in the amygdala and surrounding sublenticular extended amygdala, anterior insula, anterior

midcingulate cortex (aMCC), thalamus, PAG and surrounding midbrain, ventrolateral prefrontal and lateral orbitofrontal cortices, and fusiform/ventral occipital-temporal regions. Conversely, we found negative correlations with fear ratings in the ventromedial prefrontal cortex (vmPFC), medial orbitofrontal cortex (OFC), posterior insula/operculum, and dorsolateral prefrontal cortex (dlPFC), posterior cingulate cortex (PCC), inferior parietal lobule (IPL) and supplementary motor area (SMA) ($q < 0.05$, FDR corrected; Fig. 3A).

We then compared these univariate (single-voxel) findings to multivariate models in several ways. First, we performed a one-sample t-test analysis (treating participant as a random effect) on the weights from within-subject (ideographic) multivariate predictive models (details see Methods). Like the univariate maps, within-subject predictive models (backward models) included consistent weights in brain regions spanning multiple large-scale cortical and subcortical systems, which exhibited a large overlap with the fear regions as determined by the univariate approach (Fig. 3B; $q < 0.05$, FDR corrected).

Some brain features that make large contributions to the multivariate models might capture and control for sources of noise in the data, rather than being directly related to mental events [43]. To provide a more transparent comparison between univariate and multivariate results, we thus calculated within-subject reconstructed 'activation patterns' (forward models; see Methods for details), which assess the relationships between each voxel and the response (fitted values) in the multivariate model. These are also referred to as 'structure coefficients' in the statistical literature [44]. Fig. 3C shows results of a group analysis of 'activation patterns' across individuals ($q < 0.05$, FDR corrected). As Haufe, et al. [43] suggest, voxels that exhibit significant predictive weights and

structure coefficients are important regions that are both directly correlated with the outcomes (i.e., fear ratings) and are predictive after accounting for other brain regions in the multivariate model. As shown in Fig. 3C, the thresholded 'model activation pattern' was remarkably similar to the univariate parametric effects of fear ratings (Fig. 3A). This suggests that the multivariate model is encoding activity across distributed regions and confirms that subjective fear is associated with activity in a large number of cortical and subcortical regions. Indeed, a formal assessment of overlap (Fig. 3D) showed that virtually all regions with consistent, significant model weights in the multivariate models also encoded model information (i.e., showed significant 'model activation patterns'). The broad conclusion is that the neural representation of human fear is not limited to a single or a set of focal regions (e.g., the amygdala), but rather includes a broad set of regions spanning multiple systems.

We next determined regions that reliably contributed to the fear prediction within the VIFS itself by applying a bootstrap test to identify regions with significant, consistent model weights $q < 0.05$, FDR corrected; [34]. In line with within-subject models, a set of distributed brain regions exhibited significant model weights (Fig. 3E) and structure coefficients (Fig. 3F), including amygdala, MCC, insula, inferior frontal cortex (IFG), PAG, occipital and sensorimotor areas.

Overall, regions that were most consistently associated with subjective fear across the analyses included key regions engaged in conditioned threat (amygdala, aMCC and PAG), and general avoidance motivation (anterior insula, posterior ventral striatum) as determined across species while other regions such as the right posterior lateral prefrontal cortex/inferior frontal junction and ventral occipito-temporal stream have been associated with cognitive emotion

construction in humans and dysregulated emotional experience in mental disorders [30,45]. Negative associations with fear were most consistently identified in medial prefrontal and sensorimotor regions. In conclusion, across both univariate and multivariate analyses, our results indicate that fear experience is represented in distributed neural systems involved in defensive responses, avoidance behavior, negative affect, emotional awareness as well as pathological fear and anxiety.

Alternative models to determine the contribution of isolated fear predictive systems: local searchlights, pre-defined regions and networks perform considerably worse than VIFS

Given the continuous debate on the contribution of specific brain regions, such as amygdala [8,13,20,46] and more recently the cortical consciousness network [3,6], to the subjective fear experience, we (1) located local brain regions that were predictive of subjective fear experience using both searchlight- and parcellation-based analyses, and (2) examined to what extent models trained on single brain region and network could predict subjective fear ratings as compared with the whole-brain model (i.e., the VIFS). As shown in Fig. 4A and 4B, subjective fear experience could be significantly predicted by activations in widely distributed regions (averaged across 10×10-fold cross-validation procedure; $q < 0.05$, FDR corrected). However, none of these local regions predicted subjective fear to the extent the VIFS did.

We next re-trained predictive SVR models (with the identical cross-validation and prediction procedure as used for the VIFS) restricted to activations in (a) the bilateral amygdala; (b) a pre-defined cortical network associated with consciousness [6] (see Methods for details); (c) a subcortical region group (including striatum, thalamus, hippocampus and amygdala); and (d) each

of 7 large-scale resting-state functional networks [47,48]. We found that the amygdala as well as the other brain networks could, to some extent ($0.21 < r < 0.37$, 10×10-fold cross-validated; Fig. 4C, D and Supplementary Table 1, predict subjective fear ratings. However, although statistically significant ($Ps < 0.001$, one-tailed permutation t-tests) the effect sizes in terms of prediction-outcome correlations (including searchlight- and parcellation-based predictions) were substantially smaller than those obtained from the VIFS, which used features that span multiple systems.

This comparison is fair even though the number of features differ, as models were always tested on hold-out participants, eliminating the problem of overfitting as more predictors are used (a substantial problem in models trained and tested without cross-validation). However, to assess the potential effect of the numbers of features in the prediction analyses (i.e., whole-brain model uses much more features/voxels), as shown in Fig. 4D, we randomly selected samples from a uniform distribution spanning the entire brain (black), consciousness network (red), subcortical regions (light purple) or a single resting-state network (averaged over 1,000 iterations) [29]. The asymptotic prediction when sampling from all brain systems as we did with the VIFS (black line in Fig. 4D) was substantially higher than the asymptotic prediction within individual networks (colored lines in Fig. 4D; see also Supplementary Table 1 for details). This analysis thus demonstrated that whole-brain models have much larger effect sizes than those using features from a single network. Furthermore, model performance reached asymptote when approximately 10,000 voxels were randomly sampled across the whole-brain, as long as voxels were drawn from multiple systems, further emphasizing that subjective fear experience is encoded in distributed neural patterns that span multiple systems. Importantly, we found similar results when applying

the models trained on the discovery cohort to the validation and generalization cohorts, indicating that models trained on ~10,000 randomly sampled voxels were robust and generalizable (Supplementary Fig. 2). Together the results from the systematic analyses provide the first evidence that the subjective experience of fear is represented in distributed neural systems which argues against fear experience being reducible to activations in any single brain region or network.

Subjective fear and conditioned defensive threat responses engage distinct neural representations in humans

Fear has often been assumed to be synonymous with conditioned threat (mostly referred to as 'conditioned fear' in the literatures). However, recent fear conceptualizations emphasize potential mechanistic and neural distinctions between acquired defensive responses and the subjective experience of fear [3,6,14-16]. Against this background we examined whether the neural representation of subjective fear and conditioned threat responses were dissociable by applying the VIFS to two datasets acquired during Pavlovian threat conditioning in which an auditory cue [37] or visual cue [49] (see Methods for details), respectively was paired with a shock (CS+) while a control cue was unpaired (CS-). We specifically tested whether the VIFS generalized to discriminate CS+ versus CS-. Second, Reddan, et al. [37] developed a threat-predictive signature (TPS) that accurately classified CS+ versus CS- in new individuals based on brain activity patterns. We applied the TPS to the fear paradigm data and assessed its performance in predicting subjective fear ratings. We propose that if subjective fear and conditioned threat share similar neural mechanisms, the VIFS and TPS should perform well in cross-prediction: i.e., VIFS responses could predict CS+ versus

CS-, and TPS responses should correlate with subjective fear ratings. Conversely, low cross-prediction indicates independence of the neural representations for 'fear' and 'conditioned threat' constructs (for similar approaches see e.g., refs [35,36,42,50]). As shown in Fig. 5 the VIFS did not classify CS+ from CS- above chance during auditory (accuracy = 57±6.0%, Cohen's d = 0.09, permutation test one-tailed P = 0.234, see also Fig. 5A) or visual threat conditioning (accuracy = 62±6.4%, Cohen's d = 0.35, permutation test one-tailed P = 0.265). Whereas the TPS predicted visual CS+ versus CS- cues with high accuracy (accuracy = 93±3.3%, Cohen's d = 1.30, permutation test one-tailed P = 0.003) in the visual threat conditioning data, it did not predict fear ratings in our discovery, validation, or generalization cohorts (discovery: r = 0.16, permutation test one-tailed P = 0.085, Fig. 5B; validation: r = 0.18, permutation test one-tailed P = 0.111; generalization: r = 0.24, permutation test one-tailed P = 0.183). In support of separable brain representations underlying subjective fear experience and defensive responses towards acquired threat signals we additionally found that the VIFS and TPS pattern weights were spatially uncorrelated on the whole-brain level (r = 0.02, permutation test one-tailed P = 0.125). Moreover, we explored the joint distribution of normalized (z-scored) voxel weights of these two patterns by plotting VIFS on the y-axis and the TPS on the x-axis (for a similar approach see ref [51]). As visualized in Fig. 5C, stronger weights across the whole-brain (sum of squared distances to the origin [SSDO]) were actually observed in the nonshared Octants (1, 3, 5, 7). Overall these results suggest distinct neural representations for subjective fear experience and conditioned threat responses. These findings provide the first evidence for separable whole-brain fMRI multivariate

patterns for subjective experience of fear and conditioned threat, indicating functionally independent neural representations for subjective fear and conditioned threat.

Moreover, we tested local VIFS and TPS patterns within integrative regions traditionally related to 'fear' but independent of sensory modality, including the amygdala, insula and aMCC. Voxels within each of these regions were reliable predictors in both signatures, confirming previous human neuroimaging studies indicating a general contribution of these regions to both, fear experience and threat conditioning, while animal studies suggest that specific neural ensembles in these regions may code separate aspects of threat-related processes (see e.g., ref [52]). Local VIFS and TPS patterns within the amygdala and aMCC were negatively correlated ($r = -0.25$ and $r = -0.21$, respectively), and the local insula patterns were uncorrelated ($r = -0.04$), further emphasizing that subjective fear- and conditioned threat-related representations within core 'fear' regions are coded by separable neural representations in humans.

VIFS responses mediate subjective fear induced by negative emotion

'Fear' is a highly aversive subjective state and represents a core construct within the negative valence systems domain in the Research Domain Criteria (RDoC) matrix. To separate fear from general negative affect we next investigated the spatial and functional similarities between the VIFS and PINES (picture-induced negative emotion signature) which was developed to track general negative emotion experience [30]. We found that these two signatures exhibited a weak positive spatial correlation ($r = 0.08$, permutation test one-tailed $P < 0.001$) and the VIFS was more specific to predict subjective fear rather than general negative emotions while the PINES

was more sensitive to predict general negative emotion as compared to fear (Fig. 6A; see Supplementary results for more details).

Given that the PINES significantly predicted fear ratings in both discovery ($r = 0.38$, permutation test one-tailed $P < 0.001$) and validation ($r = 0.37$, permutation test one-tailed $P < 0.001$) cohorts and PINES and VIFS exhibited similar patterns, we next tested whether the VIFS response could explain the association between the PINES response and the fear rating. Multilevel mediation analysis (details see Methods) revealed that the VIFS response partially mediated the effect of PINES response on fear ratings in both discovery (Fig. 6A) and validation cohorts (Fig. 6B), indicating that the VIFS expression could (partially) explain the association between the general negative emotion expression and subjective fear experience. In contrast the PINES could not (discovery data), or only with comparably lower effect size (Cohen's $d = 0.06$, versus $d = 0.21$ for the above model), mediate the VIFS response – fear rating association (see Supplementary Results for details).

Discussion

In the current study we developed and validated a sensitive and generalizable brain signature for the subjective experience of fear, which predicted momentary fear on a population and individual level and thus could have a high potential for translational applications aiming at yielding information about individual fear experience. Furthermore, we challenge the notion that subjective fear is a product of a single brain region or network and propose that subjective fear is encoded in brain regions that span multiple neural systems. Across a series of analyses subjective fear was

both associated with and predicted by distributed brain systems and fear prediction by isolated brain systems was substantial lower compared to the whole-brain approach. Driven by recent debates on whether subjective fear and the defensive response elicited by conditioned threat involve different brain circuits [3,6,14-16] we, moreover, employed a fine-grained analysis technique (MVPA) to show distinct neural representations underlying these two mental processes on the whole-brain level and in traditional 'fear' modules such as the amygdala. Finally, neural representations of subjective fear and general negative emotion exhibited shared yet separable representations, with the VIFS response mediating the association between the general negative emotion response and subjective fear. Together our findings shed light both on how subjective experience of fear is represented in the human brain and how this neural representation is separable from conditioned defensive responses and general negative emotion, respectively.

Machine learning techniques have been increasingly used to develop integrated predictive models of activation across multiple brain systems to predict mental processes with large effect sizes (or explained variance) [27,29]. Applying support vector regression, we developed and validated a sensitive and robust whole-brain signature (VIFS) that accurately predicted the intensity of subjective fear experience across different fear induction paradigms and MRI systems. The identification of this intermediate neural signature of subjective fear is pivotal, as it may (1) provide objective neurobiological measures that can supplement self-report which can be biased by self-reflection or communicative intentions [53] and, (2) promote the development and evaluation of process-specific interventions that target subjective fear experience.

Our findings have theoretical implications for ongoing debates about the neural circuits of fear, specifically the neural representation of subjective fear experience (for an overview see ref [1]). For instance, the subcortical fear system theory suggests that feelings of fear arise from highly conserved amygdala-centered subcortical circuits [8,10,13], while high-order perspectives emphasize the contribution of the fronto-parietal 'consciousness network' to fear experience and propose that subcortical circuits are not directly responsible for fear experience [3,6,16,40]. Although limitations of the structure-centric view are widely acknowledged [32] and appraisal [31] and constructionist [2] theories have suggested that fear experience results from interactions between multiple processes and brain systems, a systematic empirical comparison of structure-centric versus distributed representation models of subjective fear was previously lacking.

The present findings challenge the structure-centric and network-centric models of subjective fear by demonstrating that subjective fear is represented in distributed brain regions, including but not limited to amygdala, prefrontal, subcortical and sensory cortices. Whereas previous predictive models focused on identifying brain regions that reliably contributed to the model for interpretation purpose [30,34] we updated and extended the characterization of the predictive model. We divided voxels into different classes based on the combination of predictive weights and reconstructed 'activation patterns' [43] and revealed that distributed brain regions, which exhibited significant predictive weights and reconstructed 'activations', critically contribute to both, predictions of and associations with the outcome. Second, we demonstrated that isolated regions (e.g., amygdala) and networks (e.g., 'consciousness network') predicted subjective fear to a substantially lower extent than the whole-brain signature (VIFS). Finally, around 10,000 voxels

that were randomly sampled across the whole-brain could achieve high performance of predicting subjective fear, which could also be generalized to new data collected with different paradigms. Together our findings suggest that the fear circuits identified in previous studies may only represent aspects of the subjective fear experience, as reflected by comparably low effect sizes, yet that the subjective feeling of fear requires engagement of distributed brain systems. Our findings are consistent with recent MVPA studies demonstrating that affective processes including general negative emotion [30], vicarious [35] and self-experienced pain [29] are distributed across regions, and meta-analytic evidence suggesting that emotional experience is constructed in a set of highly interacting brain regions [32].

The RDoC matrix suggests several paradigms to study acute threat (or 'fear'), including fear conditioning and exposure to emotional evocative stimuli. Indeed, subjective feelings of fear and conditioned threat exhibit a pattern of similar brain activation particularly in subcortical and prefrontal cortices [23-25,37,49]. However, recent conceptualizations propose that due to the fact that the conditioned automatic defensive response represents an innate, fixed action pattern which does not necessarily require consciousness (as opposed to subjective fear which is a conscious experience), the underlying neural mechanisms might be distinct [3,6,14-16]. Traditional univariate activation analyses lack anatomical specificity and thus cannot determine whether the neural representations of overlapping activations are similar or distinct [36,54], while the MVPA approach can extract information at finer spatial scales [27,33] and permits support for or rejection of claims about neural mechanisms that are shared between mental processes [55]. Our findings indicate separable neural representations of subjective fear and the conditioned defensive response not only

on the whole-brain level but also in 'core fear regions' such as the amygdala which are commonly co-activated during both mental processes. Our study supports and extends current conceptualizations of neurobiologically orthogonal processes and implies that conditioned threat and subjective fear are distinct constructs within the negative valence system.

In line with the RDoC matrix suggesting that fear is a core construct of the negative valance systems, the VIFS shared similar yet different characteristic functions with the PINES which tracks general negative emotional responses including sadness, anger, disgust and fear [30]. The VIFS was more sensitive to predict subjective fear, together with the observation that VIFS responses mediates the association between PINES responses and fear ratings, suggesting that the VIFS is a more robust and specific brain marker for subjective feelings of fear.

In conclusion, we identified a whole-brain neural representation for the subjective experience of fear. This visually induced fear signature was validated and generalized across participants, paradigms and fMRI scanners. Our findings demonstrate the neural basis of subjective fear is not represented by isolated brain regions or networks but instead best captured by activations in distributed regions spanning multiple brain systems. The specificity of the fear signature was further tested with conditioned defensive responses and general negative emotion experience. Our work may provide objective neurobiological measures that can supplement self-report fear and be used as intermediate markers for treatment discovery that target pathological fear.

Methods

Participants in the discovery cohort. Seventy healthy, right-handed participants were recruited from the University of Electronic Science and Technology of China in this study. Exclusion criteria included color blindness; current or regular substance or medication use; current or history of medical or psychiatric disorders; any contraindications for MRI. Due to the excessive head motion (> 1 voxel) during fMRI scanning data from 3 participants were excluded, leading to a final sample of n = 67 participants (34 females; mean ± SD age = 21.5 ± 2.1 years). All participants provided written informed consent, and the study was approved by the local ethics committee at the University of Electronic Science and Technology of China and was in accordance with the most recent revision of the Declaration of Helsinki.

Stimuli and paradigm used in the discovery cohort. The fear rating task consisted of 4 runs with each run encompassing 20 photographs (including humans, animals and scenes) from the IAPS (International Affective Picture System), NAPS (Nencki Affective Picture System) and internet. Stimuli were presented using the E-Prime software (Version 2.0; Psychology Software Tools, Sharpsburg, PA). Participants were instructed to pay attention to the pictures when they came on the screen. Each trial consisted of a 6s presentation of the picture followed by a 2s fixation-cross separating the stimuli from the rating period. Participants then had 4s to report the fearful state they experienced for the stimuli using a 5-point Likert scale where 1 indicated neutral/slightest fear and 5 indicated most strongly fear. Finally, there was a 6s rest period (fixation-cross) before the presentation of the next picture (Fig. 1A). All of the subjects reported '1-4' in their responses while 2 out of 67 subjects did not use rating '5'.

Discovery cohort MRI data acquisition and preprocessing. MRI data were collected on a 3.0-T GE Discovery MR750 system (General Electric Medical System, Milwaukee, WI, USA) (see Supplementary Methods for details). Functional MRI data was preprocessed using Statistical Parametric Mapping (SPM12 v7487, https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The first 5 volumes of each run were discarded to allow MRI T1 equilibration. Prior to preprocessing of functional data, image intensity outliers resulting from gradient and motion-related artefacts were identified using CanlabCore tools (https://github.com/canlab/CanlabCore) based on meeting any of the following criteria: (a) signal intensity > 3 standard deviations from the global mean or (b) signal intensity and Mahalanobis distances > 10 mean absolute deviations based on moving averages with a full width at half maximum (FWHM) of 20 image kernels. Each time-point identified as outliers was included as a separate nuisance covariate in the first-level model. Then, functional images were corrected for differences in the acquisition timing of each slice and spatially realigned to the first volume and unwarped to correct for nonlinear distortions related to head motion or magnetic field inhomogeneity. The anatomical image was segmented into grey matter, white matter, cerebrospinal fluid, bone, fat and air by registering tissue types to tissue probability maps. Next, the skull-stripped and bias corrected structural image was generated and the functional images were co-registered to this image. The functional images were subsequently normalized the Montreal Neurological Institute (MNI) space (interpolated to 2 × 2 × 2mm voxel size) by applying the forward deformation parameters that were obtained from the segmentation procedure, and spatially smoothed using an 8-mm full-width at half maximum (FWHM) Gaussian kernel.

First-level fMRI analysis used in the discovery cohort. We conducted two separate subject-level GLM (general linear model) analyses. The first GLM model was used to obtain beta images for the prediction analysis. In this model we included five separate boxcar regressors logged to the presentations of pictures in each rating (i.e., 1-5), which allowed us to model brain activity in response to each fear level separately. To model any effects related to motor activity the model also included one boxcar regressor indicating the rating period. The fixation cross epoch served as implicit baseline. The second GLM model included two regressors of interest, with one modeling the picture viewing period and the other modeling the fear rating period. Additionally, the design matrix included fear ratings (1-5) reported for each picture as a parametric modulator for the picture viewing period.

All task regressors were convolved with the canonical HRF function and a high-pass filter of 128 seconds was applied to remove low frequency drifts. Time series from multiple runs were concatenated using SPM's spm_fmri_concatenate.m function, which included an intercept for each run and corrected the high-pass filter and temporal non-sphericity calculations. Regressors of non-interest (nuisance variables) included (1) six head movement parameters and their squares, their derivatives and squared derivatives (leading to 24 motion-related nuisance regressors in total); and (2) indicator vectors for outlier time points (see above for details).

Participants in the validation cohort. Twenty-two healthy, right-handed participants were recruited from the University of Electronic Science and Technology of China in this study. Due to

excessive head motion (> 1 voxel) during fMRI scanning data from 2 participants were excluded leading to a final sample of n = 20 participants (6 females; mean ± SD age = 21.75 ± 2.61 years).

Stimuli and paradigm used in the validation cohort. The fear rating task consisted of 2 runs with each run encompassing 30 photographs (60 in total) from the IAPS, NAPS and internet. Fifty-eight out of sixty stimuli were overlapped with the stimuli used in the discovery cohort. Stimuli were presented using the E-Prime software. Participants were instructed to pay attention to the pictures when they came on the screen. Each trial consisted of a 6s presentation of the picture followed by a jittered fixation-cross (1s, 2s or 3s). Participants then had 4s to report the emotional state they experienced for the stimuli using a 5-point Likert scale where 1 indicated minimal fear/neutral and 5 indicated very strong fear. Finally, there was a jittered fixation-cross epoch (4s, 5s, or 6s) before the presentation of the next picture (Fig. 1B). All of the subjects reported rating '1-5' in their responses.

Validation cohort MRI data acquisition, preprocessing and first-level fMRI analysis. Imaging data acquisition, preprocessing and subject-level GLM analysis were identical to the discovery cohort.

Generalization cohort. The details of the generalization cohort were reported in previous studies (Taschereau-Dumouchel et al., 2018; 2019). Briefly, 31 participants (15 females; mean ± SD age = 23.29 ± 4.21 years) underwent a 1h fMRI session where they were presented with 3600 images consisting of 30 animal categories and 10 object categories (90 different images per category). The stimuli were grouped in blocks of 2, 3, 4 or 6 images of the same category with each stimulus

presented for 1s (no interblock or interstimulus interval). Subjective fear ratings for each category were established before the fMRI procedure without presenting any fearful stimuli. The least-square separate single trial analysis approach was employed to iteratively fit a GLM to estimate the brain response to the first image of each block and then the within-subject beta images with the same fear ratings were averaged, which resulted in one beta map per rating for each subject (for paradigm, MRI acquisition and analysis details see ref [40]).

Multivariate pattern analysis. We applied whole-brain (restricted to a grey matter mask) multivariate machine learning pattern analysis to obtain a pattern of brain activity that best predicted participants' self-reported fear ratings. We employed the support vector regression (SVR) algorithm implemented in the Spider toolbox (http://people.kyb.tuebingen.mpg.de/spider) with individual beta maps (one per rating for each subject) as features to predict participants' fear ratings of the grouped pictures they viewed while undergoing fMRI. Of note, we only used data from the discovery cohort to develop the VIFS. To evaluate the performance of our algorithm, we used a 10×10-fold cross-validation procedure on the discovery cohort during which all participants were randomly assigned to 10 subsamples of 6 or 7 participants using MATLAB's cvpartition function. The optimal hyperplane was computed based on the multivariate pattern of 60 or 61 participants (training set) and evaluated by the excluded 7 or 6 participants (test set). This procedure was repeated 10 times with each subsample being the testing set once. To avoid a potential bias of training-test splits, the cross-validation procedures throughout the study were repeated 10 times by producing different splits in each repetition and the resultant prediction

performance and P values were averaged to produce a convergent estimation [36,56]. In addition to the cross-validation procedure we applied the fear-predictive pattern (trained on the whole discovery cohort) to the validation and generalization cohorts (that is, the dot product of the VIFS weight map and the test image plus the intercept) to assess the prediction performance of the VIFS using a permutation test with 10,000 random shuffles. Given that the cross-validated permutation test is very time consuming the inferences on model performance were only performed using permutation testing on the validation and generalization cohorts.

Within-subject trial-wise prediction. Here we tested whether the VIFS could predict individual trial-by-trial subjective fear. To this end we performed a single-trial analysis, which was achieved by specifying a GLM design matrix with separate regressors for each stimulus. Each task regressor was convolved with the canonical hemodynamic response function. Nuisance regressors and high-pass filter were identical to the above GLM analyses. One important consideration for single-trial analysis is that the beta estimates for a given trial could be strongly affected by acquisition artifacts (e.g., sudden motion and scanner pulse artifacts) that cooccur during a trial. For each subject we therefore excluded trials with variance inflation factors > 3 (indicating potential multicollinearity) from subsequent analyses (overall ~6% trials were excluded). Next, we calculated the VIFS pattern expressions of these single-trial beta maps (i.e., the dot-product of vectorized activation images with the VIFS weights), which were finally correlated with the true ratings for each participant separately.

Comparing the performance of VIFS with the V.T. decoder. A previous study has developed a whole-brain fear decoder [40]. To compare the performance of VIFS with the V.T. decoder we

applied both patterns to the discovery, validation and generalization cohorts and assessed the overall prediction-outcome correlation as well as two-alternative forced-choice classification accuracies between low, moderate and high fear based on the pattern expressions.

Determining brain regions associated with and predictive of subjective fear. To identify neural circuits underlying subjective experience of fear we employed a series of analyses. Firstly, we performed one-sample t-tests on the first-level univariate parametric modulation beta maps to see which brain regions' activation was associated with fear ratings. Next, we used multivariate analyses to locate brain regions that predictive of and associated with fear ratings separately as well as brain regions showing an overlapping effect. Specifically, we evaluated the consistency of each weight for every voxel in the brain across within-subject multivariate classifiers (developed with single-trial data) using a one-sample t-test. The thresholded map ($q < 0.05$, FDR corrected) showed the consistent fear-predictive brain regions across subjects. For this analysis we only included participants in the discovery cohort whose fear ratings could be significantly predicted by their brain data (10×10-fold cross-validated; $n = 60$). Of note, similar results were found when including the entire sample. Given that the predictive brain regions could be related to (in this case) fear processing as well as suppressing the noise in the data [43] we transformed the within-subject patterns to 'activation patterns' using the following formula: $A = cov(X)*W*cov(S)^{-1}$, where A is the reconstructed activation pattern, cov(X) is the covariance matrix of training data, W is the pattern weight vector, and cov(S) is the covariance matrix of the latent factors, which is defined as W^T*X. The significant brain regions (one-sample t-test thresholded at $q < 0.05$, FDR corrected) exhibited the consistent fear-associative effect. In parallel with with-subject models we

conducted bootstrap tests, where we took 10,000 samples with replacement from the discovery cohort, repeated the prediction process with each bootstrap sample, and calculated Z scores and two-tailed uncorrected P-values with the mean and standard deviation of the sampling distribution, on the population-level fear-predictive pattern (i.e., the VIFS) as well as the transformed 'activation pattern' from the VIFS to identify the reliable fear predictive and associative brain regions of the VIFS (thresholded at $q < 0.05$, FDR corrected).

Furthermore, we asked whether fear processing could be reducible to activations in a single brain region (e.g., amygdala) or network (e.g., subcortical regions). To examine this hypothesis, we employed whole-brain searchlight (three-voxel radius spheres) [57] – and parcellation (274 cortical and subcortical regions) [58] – based analyses to identify local regions predictive of fear and compared model performances of local regions with the whole-brain model (i.e., the VIFS). In addition, we compared prediction performances of amygdala (based on Anatomy Toolbox version 2.2c; available online in the Cognitive and Affective Neuroscience Laboratory Github repository at https://github.com/canlab/Neuroimaging_Pattern_Masks) and large-scale networks to the whole-brain approach. The networks of interest included 7 resting-state functional networks [47,48], a subcortical network (including the striatum, thalamus, hippocampus and amygdala) and a 'consciousness network' proposed by LeDoux and Pine [6], which composed of anterior cingulate cortex, inferior frontal gyrus, middle frontal gyrus, superior frontal gyrus, orbitofrontal gyrus, rectus, olfactory and insula from the AAL atlas and the posterior parietal cortex from Shirer, et al. [59].

Comparing the similarities of the VIFS and a threat-predictive signature. To examine the functional and spatial similarities between the VIFS and the TPS threat-predictive signature; [37] which predicts the defense responses elicited by threat conditioning, we (1) applied VIFS to distinguish unreinforced CS+ versus CS- and predicted subjective fear ratings using the TPS, and (2) examined the voxel-level spatial similarity between these two signatures. Inference on model performance was performed using permutation testing with 10,000 random shuffles. Given that the fear and threat conditioning studies employed visual and auditory cues, respectively, we next tested whether the dissociable effects based simply on differences in sensory processing by applying both signatures to an independent visual threat conditioning dataset [49]. If the signatures were sensory-dependent the TPS would not distinguish visual (unreinforced) CS+ versus CS- whereas the VIFS might predict unreinforced CS+ from CS- if fear and conditioned threat shared similar neural representations. For this analysis we included two datasets that employed an auditory and visual threat conditioning paradigm during which a previously neutral stimulus was paired with an unpleasant shock (CS+) while another matched stimulus was not paired (see Supplementary Methods for details)

Multilevel two-path mediation analysis. In order to test the relationship between VIFS response, fear rating and the PINES response we conducted two multilevel mediation analyses using the Mediation Toolbox (https://github.com/canlab/MediationToolbox). A mediation analysis tests whether the observed covariance between an independent variable (X) and a dependent variable (Y) could be explained by a third variable (M). Significant mediation effect is obtained when inclusion of M in a path model of the effect of X on Y significantly alters the slope of the X–Y

relationship. That is, the difference between total (path c) and direct (non-mediated, path c′) effects of X on Y (i.e., c - c′), which could be performed by testing the significance of the product of the path coefficients of path a × b, is statistically significant. The multilevel mediation analysis is designed to explain both within- and between-subject variations in the same model by treating the subject as a random effect [54]. The first-level accounts for the mediation effects within each individual participant and the second-level tests for consistency across participants and allows for population inference. In the current study we tested whether (1) VIFS response mediated the association between PINES response and fear rating and (2) PINES response mediated the relationship between VIFS response and fear rating. To this end the VIFS and PINES responses were calculated by dot-product of the single-trial beta maps with the VIFS (through cross-validation procedure for the discovery cohort) and PINES patterns respectively for each subject. We used bias corrected accelerated bootstrapping (10,000 replacements) for significance testing.

Acknowledgements

We thank Vincent Taschereau-Dumouchel and colleagues for sharing their data. This work was supported by the National Key Research and Development Program of China (Grant No. 2018YFA0701400) to B.B.; National Institute of Mental Health (R01MH076136; R01MH116026) to T.D.W; National Natural Science Foundation of China (Grant No. 31530032) to K.M.K; China Postdoctoral Science Foundation (Grant No. 2018M643432) to W.Z..

Author Contributions

F.Z., T.D.W. and B.B. designed the experiment, analyzed the data and drafted the manuscript. F.Z. W.Z., Q.Y., and Y.G. conducted the experiment. W.Z., S.Y. and K.M.K provided feedback and revised the manuscript.

Competing interests

The authors declare no competing financial interests.

Data and materials availability

The visually induced fear signature is available at https://figshare.com/articles/dataset/Subjective_fear_dataset/13271102. The analysis code is available at https:// github.com/canlab. Additional data related to this paper may be requested from the authors.

References

1 Mobbs, D. *et al.* Viewpoints: Approaches to defining and investigating fear. *Nature neuroscience* **22**, 1205-1216 (2019).

2 Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience* **12**, 1-23 (2017).

3 LeDoux, J. E. Coming to terms with fear. *Proceedings of the National Academy of Sciences* **111**, 2871-2878 (2014).

4 Adolphs, R. The biology of fear. *Current Biology* **23**, R79-R93 (2013).

5 Gross, J. J. & Feldman Barrett, L. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion review* **3**, 8-16 (2011).

6 LeDoux, J. E. & Pine, D. S. Using neuroscience to help understand fear and anxiety: a two-system framework. *American journal of psychiatry* (2016).

7 Gross, C. T. & Canteras, N. S. The many paths to fear. *Nature Reviews Neuroscience* **13**, 651-658 (2012).

8 Davis, M. The role of the amygdala in fear and anxiety. *Annual review of neuroscience* **15**, 353-375 (1992).

9 Janak, P. H. & Tye, K. M. From circuits to behaviour in the amygdala. *Nature* **517**, 284-292 (2015).

10 LeDoux, J. The emotional brain, fear, and the amygdala. *Cellular and molecular neurobiology* **23**, 727-738 (2003).

11 Shackman, A. J. & Fox, A. S. Contributions of the central extended amygdala to fear and anxiety. *Journal of Neuroscience* **36**, 8050-8063 (2016).

12 Takahashi, L. K. Olfactory systems and neural circuits that modulate predator odor fear. *Frontiers in behavioral neuroscience* **8**, 72 (2014).

13 Panksepp, J. *Affective neuroscience: The foundations of human and animal emotions*. (Oxford university press, 2004).

14 Barrett, L. F. *et al.* Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspectives on Psychological Science* **2**, 297-312 (2007).

15 Mobbs, D., Hagan, C. C., Dalgleish, T., Silston, B. & Prévost, C. The ecology of human fear: survival optimization and the nervous system. *Frontiers in neuroscience* **9**, 55 (2015).

16 LeDoux, J. E. & Brown, R. A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences* **114**, E2016-E2025 (2017).

17 Milad, M. R. & Quirk, G. J. Fear extinction as a model for translational neuroscience: ten years of progress. *Annual review of psychology* **63**, 129-151 (2012).

18 Shin, L. M. & Liberzon, I. The neurocircuitry of fear, stress, and anxiety disorders. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **35**, 169-191 (2010).

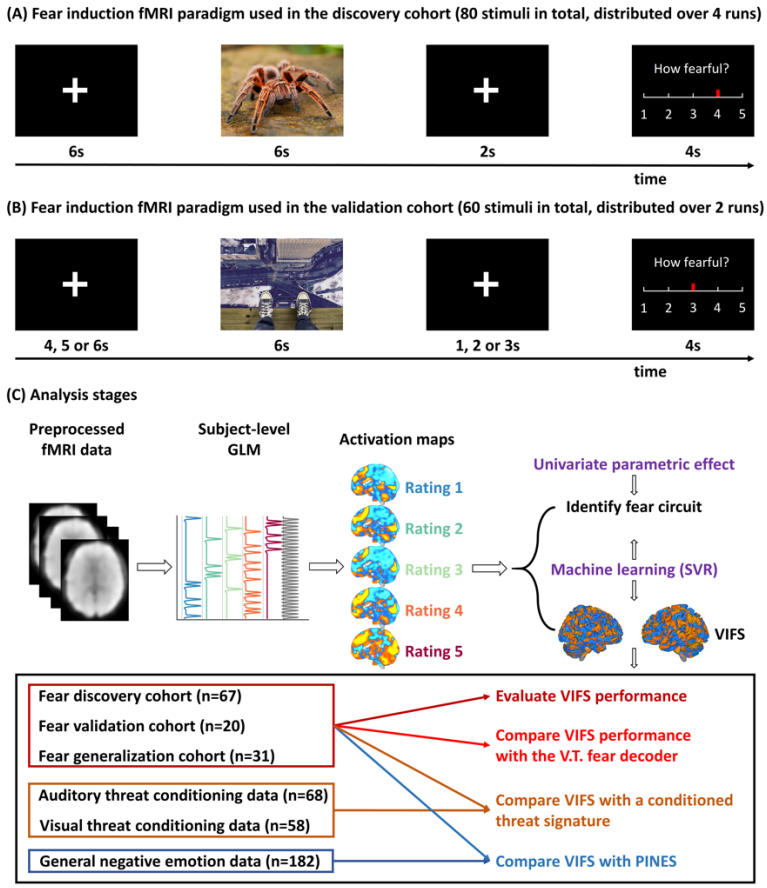
19 Adolphs, R. *et al.* A mechanism for impaired fear recognition after amygdala damage. *Nature* **433**, 68-72 (2005).
20 Feinstein, J. S., Adolphs, R., Damasio, A. & Tranel, D. The human amygdala and the induction and experience of fear. *Current biology* **21**, 34-38 (2011).
21 Hurlemann, R. *et al.* Amygdala control of emotion-induced forgetting and remembering: evidence from Urbach-Wiethe disease. *Neuropsychologia* **45**, 877-884 (2007).
22 Becker, B. *et al.* Fear processing and social networking in the absence of a functional amygdala. *Biological psychiatry* **72**, 70-77 (2012).
23 Alpers, G. W. *et al.* Attention and amygdala activity: an fMRI study with spider pictures in spider phobia. *Journal of Neural Transmission* **116**, 747-757 (2009).
24 Stark, R. *et al.* Hemodynamic brain correlates of disgust and fear ratings. *NeuroImage* **37**, 663-673 (2007).
25 Hudson, M. *et al.* Dissociable neural systems for unconditioned acute and sustained fear. *NeuroImage*, 116522 (2020).
26 Mobbs, D. *et al.* When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. *Science* **317**, 1079-1083 (2007).
27 Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nature neuroscience* **20**, 365 (2017).
28 Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience* **18**, 115 (2017).
29 Kragel, P. A., Koban, L., Barrett, L. F. & Wager, T. D. Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron* **99**, 257-273, doi:https://doi.org/10.1016/j.neuron.2018.06.009 (2018).
30 Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A sensitive and specific neural signature for picture-induced negative affect. *PLoS biology* **13**, e1002180 (2015).
31 Clore, G. L. & Ortony, A. Appraisal theories: How cognition shapes affect into emotion. (2008).
32 Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Barrett, L. F. The brain basis of emotion: a meta-analytic review. *The Behavioral and brain sciences* **35**, 121 (2012).
33 Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nature neuroscience* **8**, 679 (2005).
34 Wager, T. D. *et al.* An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine* **368**, 1388-1397 (2013).
35 Krishnan, A. *et al.* Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *Elife* **5**, e15166 (2016).

36 Zhou, F. *et al.* Empathic pain evoked by sensory and emotional-communicative cues share common and process-specific neural representation. *eLife* **9**, e56929 (2020).
37 Reddan, M. C., Wager, T. D. & Schiller, D. Attenuating neural threat expression with imagination. *Neuron* **100**, 994-1005. e1004 (2018).
38 Kragel, P. A. & LaBar, K. S. Multivariate neural biomarkers of emotional states are categorically distinct. *Social cognitive and affective neuroscience* **10**, 1437-1448 (2015).
39 Saarimäki, H. *et al.* Discrete neural signatures of basic emotions. *Cerebral cortex* **26**, 2563-2573 (2016).
40 Taschereau-Dumouchel, V., Kawato, M. & Lau, H. Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. *Molecular Psychiatry*, 1-13 (2019).
41 Peelen, M. V., Wiggett, A. J. & Downing, P. E. Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron* **49**, 815-822 (2006).
42 Woo, C.-W. *et al.* Separate neural representations for physical pain and social rejection. *Nature communications* **5**, 5380 (2014).
43 Haufe, S. *et al.* On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* **87**, 96-110 (2014).
44 Courville, T. & Thompson, B. Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement* **61**, 229-248 (2001).
45 McTeague, L. M. *et al.* Identification of common neural circuit disruptions in emotional processing across psychiatric disorders. *American Journal of Psychiatry* **177**, 411-421 (2020).
46 LeDoux, J. *The emotional brain: The mysterious underpinnings of emotional life*. (Simon and Schuster, 1998).
47 Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C. & Yeo, B. T. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of neurophysiology* (2011).
48 Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* (2011).
49 Zhou, F. *et al.* Human extinction learning is accelerated by an angiotensin antagonist via ventromedial prefrontal cortex and its connections with basolateral amygdala. *Biological psychiatry* **86**, 910-920 (2019).
50 Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P. & Singer, T. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nature communications* **7**, 10904 (2016).
51 Koban, L., Jepma, M., López-Solà, M. & Wager, T. D. Different brain networks mediate the effects of social and conditioned expectations on pain. *Nature communications* **10**, 1-13 (2019).

52 Gothard, K. M. Multidimensional processing in the amygdala. *Nature Reviews Neuroscience*, 1-11 (2020).
53 Barrett, L. F. Are emotions natural kinds? *Perspectives on psychological science* **1**, 28-58 (2006).
54 Woo, C.-W., Roy, M., Buhle, J. T. & Wager, T. D. Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biol* **13**, e1002036 (2015).
55 Peelen, M. V. & Downing, P. E. Using multi-voxel pattern analysis of fMRI data to interpret overlapping functional activations. *Trends in cognitive sciences* **11**, 4-4 (2007).
56 Zhou, F. *et al.* Shifted balance of dorsal versus ventral striatal communication with frontal reward and regulatory regions in cannabis-dependent males. *Human brain mapping* **39**, 5062-5073 (2018).
57 Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences* **103**, 3863-3868 (2006).
58 Fan, L. *et al.* The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex* **26**, 3508-3526 (2016).
59 Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V. & Greicius, M. D. Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral cortex* **22**, 158-165 (2012).

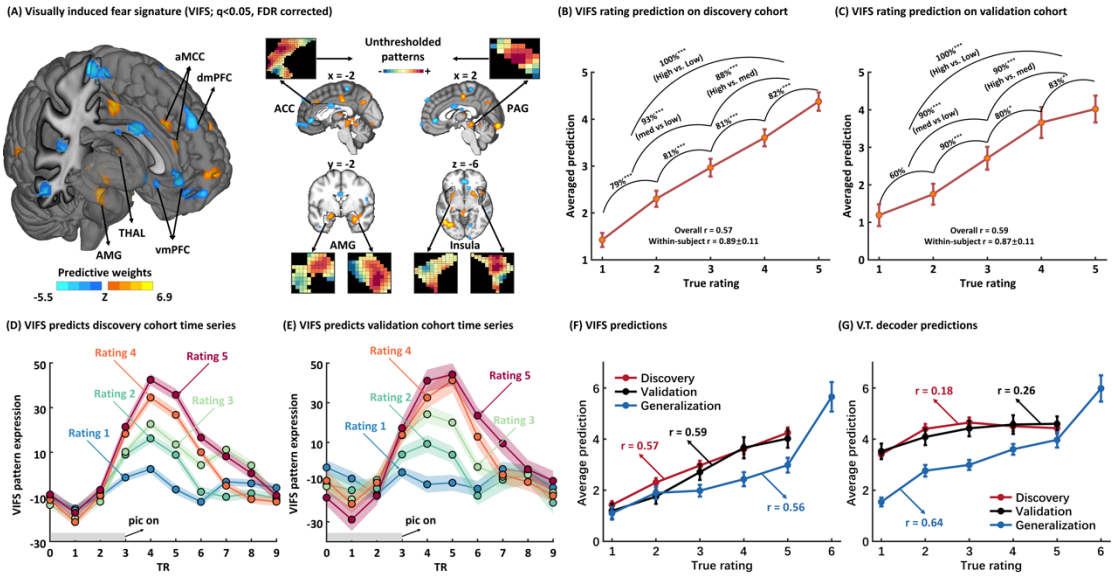
Figures and legends

Fig. 1 Experimental paradigms and analysis stages.



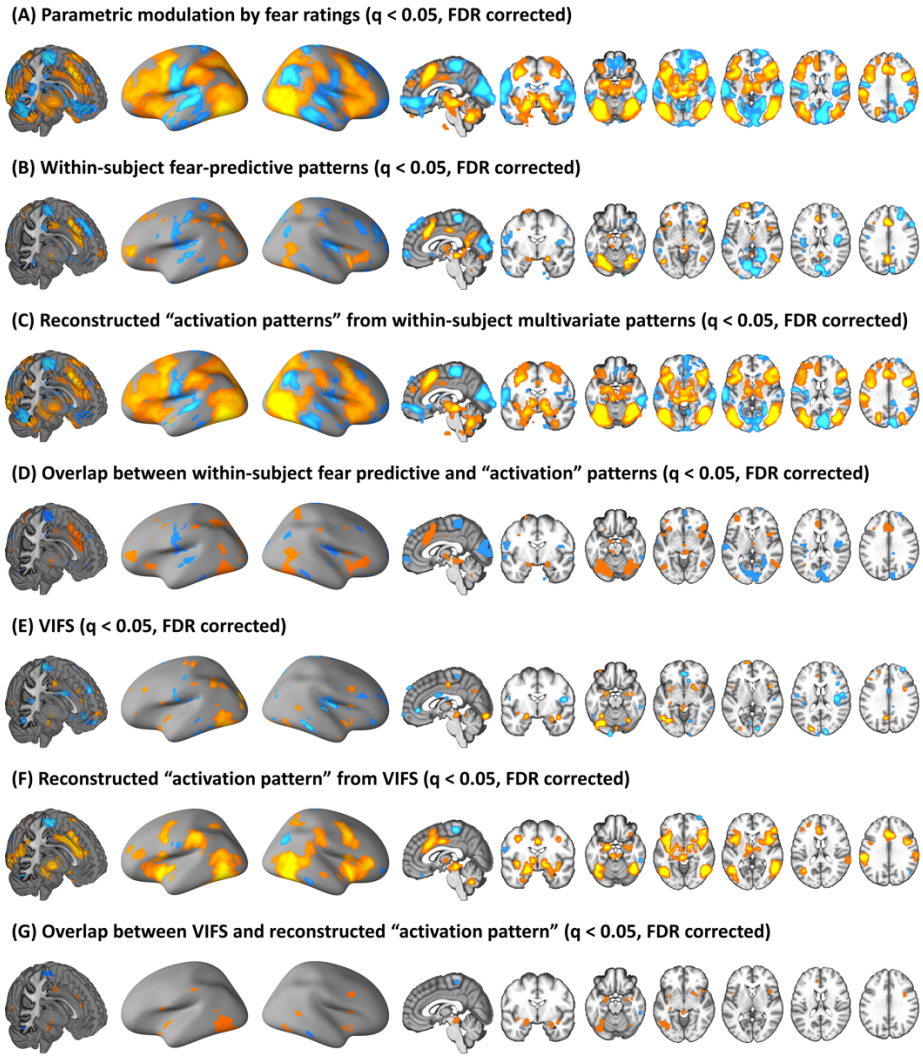
Discovery (panel A) and validation (panel B) cohorts underwent two similar but not identical fear induction and rating paradigms during fMRI acquisition. Of note, examples of the fear-evoking photos are pictures (free for commercial use) only for display purposes and not included in the original stimulus set. Panel C depicts the analytic stages and datasets used in the present study. Specifically, a whole-brain multivariate pattern predictive of the level of subjective experience of fear was trained on the discovery sample (n=67) using support vector regression and further evaluated in discovery (cross-validated), validation (n=20) and generalization (n=31) cohorts. We next systematically applied univariate and multivariate analyses to determine the spatial scale and local contributions of specific regions to the momentary subjective fear representation. Finally, we tested whether subjective fear was encoded with a neural signature that was distinct from the representation of conditioned threat (CS+ versus CS- cues) and general negative affect. SVR, support vector regression; VIFS, visually induced fear signature developed in the current study; PINES, picture-induced negative emotion signature developed by Luke Chang and colleagues; V.T. fear decoder, a fear-predictive pattern developed by Vincent Taschereau-Dumouchel and colleagues. See Methods and Results for the details of the datasets and brain signatures used in this study.

Fig. 2 Visually induced fear signature (VIFS).



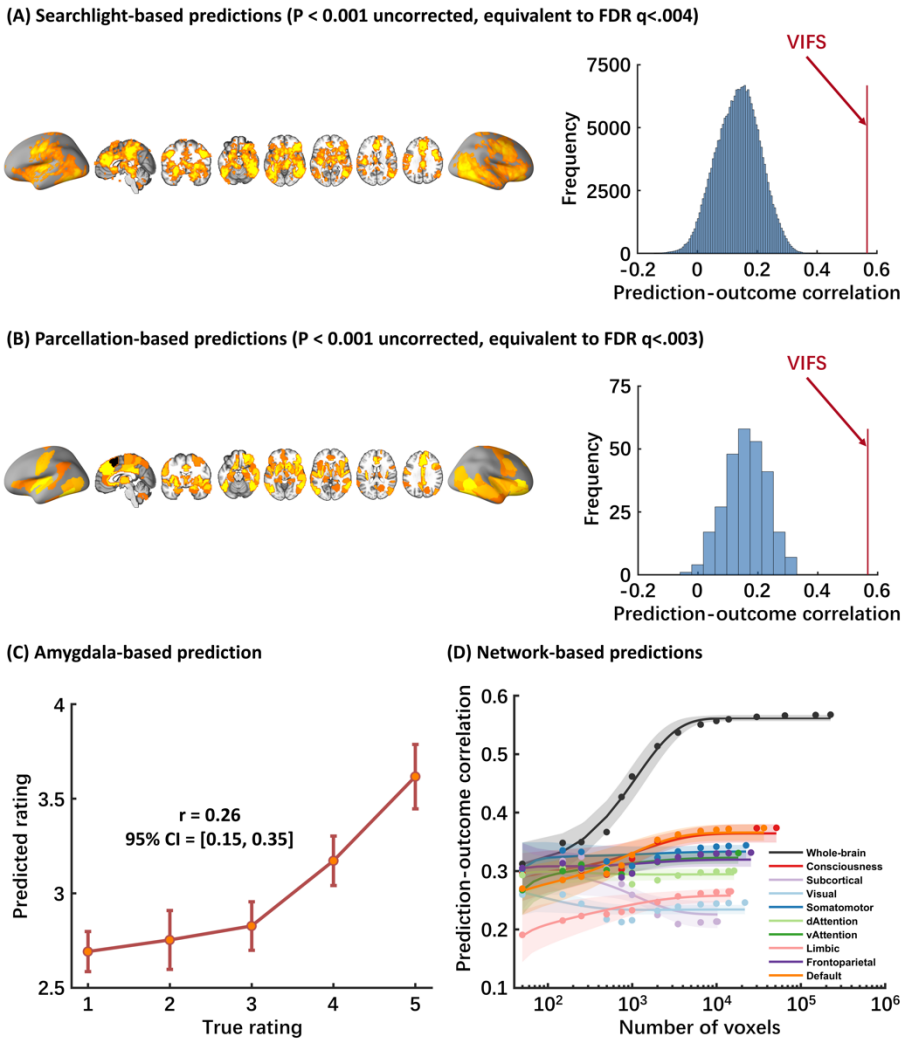
Panel A depicts the VIFS pattern thresholded using a 10,000-sample bootstrap procedure at $q < 0.05$, FDR corrected. Inserts show the spatial topography of the unthresholded patterns in the left ACC, right PAG, bilateral AMG and bilateral insula. AMG denotes amygdala, THAL thalamus, vmPFC ventromedial prefrontal cortex, dmPFC dorsal medial prefrontal cortex, dACC dorsal anterior cingulate cortex, MCC middle cingulate cortex, ACC anterior cingulate cortex, and PAG periaqueductal gray. Panels B and C depict the predicted fear experience (subjective ratings) compared to the actual level of fear for the cross-validated discovery cohort and the independent validation cohort, respectively. Accuracies reflect forced-choice comparisons. Panels D and E depict an average peristimulus plot of the VIFS response to the cross-validated discovery cohort and the independent validation cohort. This reflects the average VIFS response at every repetition time (TR; 2s) in the timeseries separated by the fear ratings. Panel F and G compare the fear prediction of V.T. decoder with the VIFS on discover, validation and generalization cohorts, respectively. *indicates $P < 0.05$, **$P < 0.01$ and ***$P < 0.001$. Error bars and shaded regions indicate standard errors.

Fig. 3 Subjective experience of fear is associated with and predicted by distributed brain regions.



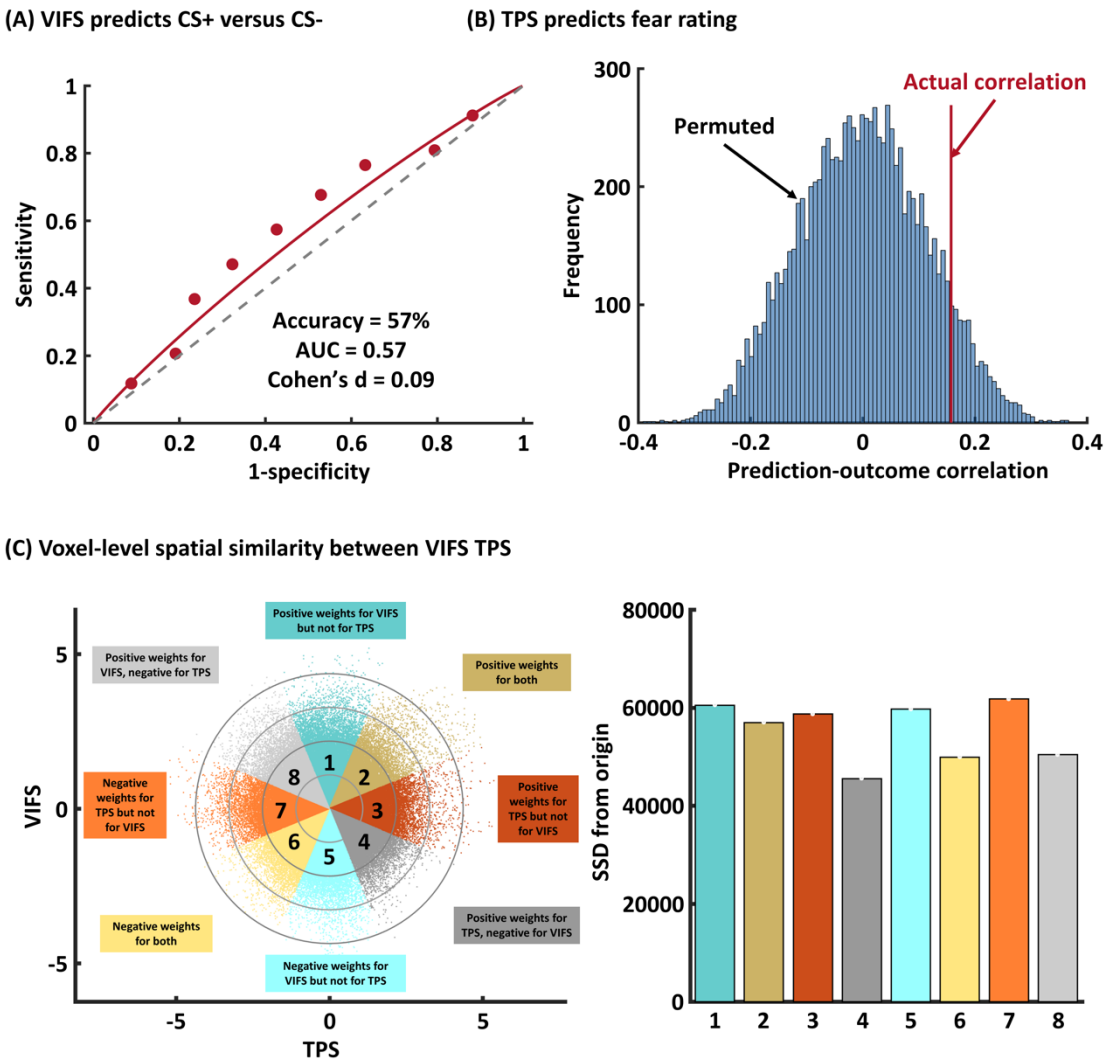
Panel A shows univariate parametric effects of fear ratings. Panel B depicts brain regions consistently predictive of subjective fear across participants. Panel C shows thresholded transformed 'activation patterns' from within-subject fear-predictive patterns. Panel D demonstrates overlapping brain regions between panel B and C. Panel E shows the thresholded VIFS. Panel F depicts threshholded transformed 'activation pattern' from the VIFS. Panel G shows the overlap between VIFS and transformed 'activation pattern'. Red color indicates positive association and weight, and blue color indicates negative association and weight. All images are thresholded at $q < 0.05$, FDR corrected. Hot color indicates positive associations (panels A, C and F) or weights (panels B and E) where as cold color indicates negative associations (panels A, C and F) or weights (panels B and E). VIFS, visually induced fear signature.

Fig. 4 Local brain region and network predictions.



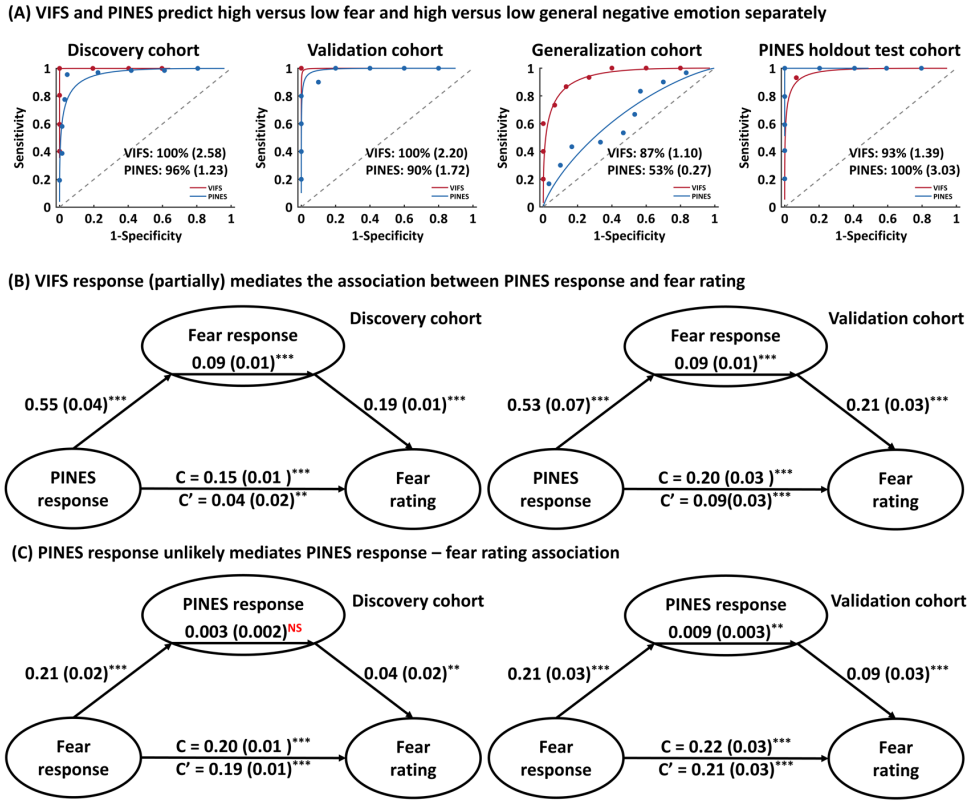
Panel A shows brain regions that can significantly predict subjective fear ratings revealed by searchlight analysis. Histograms: cross-validated predictions (correlations) from local searchlight analysis. Red line indicates the prediction-outcome correlation from VIFS. Panel B depicts brain regions which can significantly predict subjective fear revealed by parcellation-based analysis. Histograms: cross-validated predictions (correlations) from parcellations. Red line indicates the prediction-outcome correlation from VIFS. Panel C shows cross-validated predictions from amygdala-based prediction analysis. Error bar indicates standard error. Panel D demonstrates the information about subjective experience of fear is distributed across multiple systems. Model performance was evaluated as increasing numbers of voxels/features (x axis) were used to predict subjective fear in different regions of interest including the entire brain (black), consciousness network (red), subcortical regions (light purple) or large-scale resting-state networks. The y axis denotes the cross-validated prediction-outcome correlation. Colored dots indicate the correlation coefficients, solid lines indicate the mean parametric fit and shaded regions indicate standard deviation. Model performance is optimized when approximately 10,000 voxels are randomly sampled across the whole-brain. VIFS, visually induced fear signature.

Fig. 5 Comparing fear- and threat-predictive signatures.



Panel A depicts that visually induced fear signature (VIFS) does not distinguish unreinforced CS+ versus CS-. Panel B shows the histograms of prediction of threat-predictive signature (TPS) on fear data from nonparametric permutation test. Histograms show the distribution of null-hypothesis prediction-outcome correlations, and the red line shows the actual correlation coefficient. Panel C demonstrates the scatter plot displaying normalized voxel weights for VIFS (y-axis) and TPS (x-axis). Bars on the right represent the sum of squared distances from the origin (0,0) for each Octant. Different colors are assigned to the eight Octants that reflect voxels of shared positive or shared negative weights (Octants 2 and 6, respectively), selectively positive weights for the VIFS (Octant 1) or for TPS (Octant 3), selectively negative weights for the VIFS (Octant 5) or TPS (Octant 7), and voxels with opposite weights for the two neural signatures (Octants 4 and 8).

Fig. 6 Comparing VIFS and PINES responses.



Panel A depicts that VIFS more accurately (shown as forced-choice classification accuracy and Cohen's d) predict high versus low subjective fear while PINES is more sensitive to distinguish high versus low general negative emotion. Panel B shows the multilevel mediation analysis results showing that VIFS response mediates the PINES response – fear rating association in both discovery and validation cohorts. Panel C shows that the PINES response does not mediate the VIFS response – fear rating association in the discovery cohort. Although the mediation effect is significant in the validation cohort, the effect size (Cohen's d = 0.06) is very small. **indicates $P < 0.01$, $^{***}P < 0.001$, ^{NS} $P > 0.5$.

Table 1 Comparing performance of VIFS with V.T. decoder

classifications		VIFS	V.T. decoder
Discovery dataset	high versus low	100±0%*** (2.58)	79±5.0%*** (0.77)
	high versus moderate	88±3.9%*** (1.18)	42±6.0%^{NS} (-0.22)
	moderate versus low	93±3.1%*** (1.40)	84±4.5%*** (1.02)
Validation dataset	high versus low	100±0%*** (2.20)	95±4.9%*** (1.88)
	high versus moderate	90±6.7%*** (1.21)	55±11.1%^{NS} (0.19)
	moderate versus low	90±6.7%*** (1.27)	65±10.7%^{NS} (0.78)
Generalization dataset	high versus low	87±6.2%*** (1.10)	90±5.5%***(1.56)
	high versus moderate	83±6.8%*** (0.97)	87±6.2%*** (0.79)
	moderate versus low	83±7.0%*** (0.86)	93±4.7%*** (1.52)

For each dataset we used VIFS and V.T. decoders to classify high, moderate and low subjective fear using two alternative forced-choice tests. Performance was shown as accuracy ± SE (Cohen's d). ***denotes $P < 0.001$, and ^{NS} denotes non-significant based binomial tests.

Table 2 Comparing prediction (correlation) of VIFS and PINES

Datasets	VIFS (correlation, 95%CI)	PINES (correlation, 95%CI)
Discovery	0.57 [0.49, 0.63] (cv)	0.38 [0.28 0.47]
Validation	0.59 [0.48, 0.69]	0.37 [0.21, 0.51]
Generalization	0.56 [0.45, 0.64]	0.20 [0.02, 0.36]
PINES holdout	0.29 [0.17, 0.38]	0.72 [0.65, 0.77]

We applied the VIFS and PINES to subjective fear and general negative emotion holdout datasets and calculated the correlation (bootstrapped 95% CI) between the pattern expressions and the true ratings.