Neural predictors of subjective fear depend on the situation

Yiyu Wang[1], Philip A. Kragel[2], and Ajay B. Satpute[1,3]

[1]Department of Psychology, Northeastern University

[2]Department of Psychology, Emory University

[3]Department of Radiology, Massachusetts General Hospital

The extent to which neural representations of subjective fear depend on or generalize across the situational context has remained unclear. We systematically manipulated variation within and across three distinct fear-evocative situations including fear of heights, spiders, and social threats. Searchlight multivoxel pattern analysis (MVPA) was used to identify whether and which brain regions carry information that predicts subjective fear, and the degree to which the fear-predictive neural codes in these areas depend upon or generalize across the situations. The overwhelming majority of brain regions carrying information about fear did so in a situation dependent manner. These findings suggest that neural representations of fear are unlikely to involve a singular pattern or neural signature, but rather a collection of multiple heterogeneous brain states.

Keywords: Fear, subjective experience, multivariate pattern analysis

Correspondence: *wang.yiyu@northeastern.edu*

"'Fear' of getting wet is not the same fear as fear of a bear." ~ William James

For over a century, philosophers, psychologists, and neuroscientists have debated the nature of emotions [1,2]. Much of this debate concerns the uniformity or heterogeneity of representation for particular emotion categories. For example, is there a particular brain state or set of core regions that underlie fearful experiences, or does fear involve a collection of heterogeneous brain states? Addressing this question has broad implications for developing translational models of mood and anxiety disorders and brain-based interventions such as biofeedback or neurostimulation therapies [3,4].

Of potential relevance to this question are recent functional magnetic resonance imaging (fMRI) studies that have searched for the "neural signatures" of emotion categories using multivariate analytical approaches [5–9]. A neural signature of emotion refers to a statistical model that can be used to classify, when given some input brain data, whether and which emotion category is most likely to be present in independent subjects (Kragel et al., 2018). Several studies have now shown that neural signatures of emotion draw on information that is widely distributed throughout much of the brain. However, the degree to which this information is organized into a single prototypical pattern or reflects an amalgam of heterogeneous functional states remains unclear in part because neural signatures do not inherently provide direct insight about the neural representations of emotion [10,11]. Significant classification is possible even when few of the brain regions comprising the whole brain pattern carry category-level representations of an emotion, since different areas may be used to predict different instances of the emotion (Azari et al., 2020; Clark-Polner et al., 2017; Kragel et al., 2018; Lindquist et al., 2022). Determining whether neural signatures reflect a single, categorical representation of fear or multiple representations for different subtypes of fear requires direct comparisons across different instances that vary in their perceptual and semantic content, yet are commonly evaluated as producing fear.

To effect, a handful of univariate fMRI studies have shown that distinct sets of brain regions are engaged depending on the fear-evocative content (e.g. pictures of spiders, blood/injections/injuries, social encounters) [12–14]. These findings are suggestive of heterogeneity in the neural representation of fear. However, there remain two main limitations to this work that preclude a clear conclusion. First, to model fear, these studies either compared stimuli with distinct semantic content (e.g. stimuli with spider content vs. non-spider content), or participant groups (phobic v. non-phobic). Accordingly, the findings may be due to idiosyncratic individual differences, or differences in the semantic content of the stimuli, rather than fearful subjective experiences, per se. Second, these studies did not use an information-based, multivariate pattern analysis approach (MVPA). Patterns of activation may carry information about psychological states irrespective of the presence or absence of differences in overall activation magnitude.

Here, we tested which, if any, brain regions carry category level representations of fear. Focusing on subjective experiences of fear enabled us to systematically manipulate variation within and across three distinct and externally valid situations including fear of heights, spiders, and social threats. Circumventing limitations of prior work, evocative video stimuli were curated to evoke a wide range of subjective fear within each situation, which enables a parametric examination of subjective fear and provides a greater control over differences in semantic content than in previous studies while still maintaining a naturalistic context.

We used searchlight MVPA[15] to identify which brain regions

carry situation general or situation dependent information that predicts subjective experiences of fear. Critically, theoretical models that assume uniformity v. heterogeneity in fear-predictive activation patterns suggest that different methods should be used to optimally test the hypotheses. If a brain region carries situation general information, it would be preferable to train the model using instances across situations to enable the model to learn which signals are indeed situation general. Alternatively, if a brain region carries situation dependent information, it would be preferable to train the model using instances from a specific situation to enable the model to learn which signals are indeed situation general. We trained our models both ways and compared their performance to assess how fear is represented in brain activity.

Methods

Participants. Neurotypical participants (N = 24) who reported no clinical psychiatric diagnosis were recruited from the greater Los Angeles area. Exclusion criteria consisted of claustrophobia, psychiatric medication, left-handedness, metal in the body, and age (under 18 years or over 55 years). After excluding individuals with excessive motion (N = 3; criteria described below), the effective sample included 21 participants (11 male, 10 female, ages 22-40, mean age = 30.4).

Stimuli. To test for the generalization or dependence of fear across situations, evocative video stimuli were obtained and normed (see Supplemental Materials) to elicit parametric variation in self-reported fear across three distinct situations (Figure S2). For ecological validity, all videos depicted naturalistic footage and were shot from an immersive 1st person perspective. To introduce parametric variation within each situation, videos (12/situation) were selected that ranged across low to high in normative fear ratings. For example, in a high fear heights video, 1st person footage depicts walking along the edge of a cliff, whereas in a control video, 1st person footage depicts walking down a set of stairs. Thus, low normative fear videos retain much of the semantic and perceptual similarity of high normative fear videos, albeit without evoking fear.

Experimental Task. Stimuli were presented using MATLAB (MathWorks) and the Psychophysics Toolbox, and behavioral responses were recorded using a scanner-compatible trackball. Participants were instructed to watch each video stimulus and immerse themselves in the situation depicted. After each video, participants made self-report ratings on a sliding scale for experienced fear, arousal, and valence. Fear and arousal scales were tethered using "low" and "high", and the valence scale was tethered from "negative" to "positive". Participants used the trackball to move a cursor along the scale and then clicked a button under their right thumb to log their rating. Four seconds were allotted to make each rating. A jittered fixation interval of 3-5 seconds was presented before the next trial. Prior to each video stimulus, a word ("Heights", "Social", or "Spider") corresponding to the category of the upcoming video was presented for 3 seconds, after which a sliding scale for expected fear tethered from "low" to "high", however, this period was not analyzed to address the present hypotheses. Trials were presented across three runs, each approximately 9 minutes long. Each run consisted of twelve videos in which two normatively high fear and two normatively low fear videos from each category were presented (based on median splits of normative fear ratings). The order of video stimulus presentations was randomized with two exceptions to ensure uniformity of stimulus types over time. First, videos of a given category were preceded by videos of the same and different categories equally often. Second, videos with a given normative fear rating were preceded by videos of higher and lower normative fear ratings equally often.

Participants failed to provide fear ratings within the allotted time on a small proportion of trials. Missing fear ratings were interpolated (see Supplementary Materials). The following analysis used the interpolated ratings.

fMRI data acquisition and preprocessing. MRI data were collected using a 3T Siemens Trio MRI scanner. Functional images were acquired in interleaved order using a T2*-weighted multiband echo planar imaging (EPI) pulse sequence (transverse slices, TR=1000ms, TE = 3000ms, flip angle=60°, FOV = 200mm, 2.5 mm thickness slices, voxel dimension= 2.5 x 2.5 x 2.5 mm, Phase Encoding Direction anterior to posterior (AP), multi band acceleration factor = 4). Anatomical images were acquired at the start of the session with a T1-weighted pulse sequence (TR = 2400ms, TE = 2600ms, flip angle=8°, FOV = 256mm, 1 mm thickness slices, voxel dimension= 1 x 1 x 1 mm). Image volumes were preprocessed using fMRIprep (Esteban et al., 2019). Preprocessing included motion correction, slice-timing correction, removal of high frequency drifts using a temporal high-pass filter (discrete cosine transform, 100s cutoff), and spatial smoothing (6mm FWHM). For all analyses, functional volumes were downsampled to a 3 mm space to speed searchlight analyses and registered to participants' anatomical image and then to a standard template (MNI152) using FSL FLIRT [16]. Participants with at least two runs without excessive head motion (defined as > 2 mm maximum framewise displacement), were included in the analysis yielding 18 participants with three runs of data and 3 participants with two runs of data.

General Linear Model. A general linear model (GLM) was used to model the neural data. The GLM included a separate boxcar regressor for each video stimulus, convolved with a canonical hemodynamic response function from SPM12. Nuisance regressors included six regressors corresponding with motion parameters, four regressors for physiological noise artifacts (CSF, white matter, Framewise Displacement, and non-steady states), and three regressors to model activity related to incidental visual properties of luminance, contrast, and complexity of the stimuli. The GLM was conducted using custom scripts in Python. Beta maps for each video and participant were used for training the searchlight LASSO-PCR (see below).Analysis scripts are available in https://github.com/yiyuwang/AffVids_mvpa.

Searchlight LASSO-PCR. We analyzed the data using a searchlight multivariate pattern analysis approach. Betas

from the estimated GLM were extracted from each voxel within the voxel's searchlight neighborhood using a 15mm (5 voxel) radius. Because voxel data is non-independent, and because the outcome variable (self-reported fear) is continuous, the data was submitted to a Least Absolute Shrinkage and Selection Operator and Principal Components Regression (LASSO-PCR) [17–19]. LASSOPCR performs a principal component analysis to represent non-independent activity across voxels as a set of orthogonal components. Fear ratings are then regressed on component scores using a LASSO regularized regression. A relatively lenient penalty term of 0.05 was used since the searchlight analysis already constrains the dimensionality of the fMRI data. The analysis was performed using the modified scikit-learn and nilearn [64] Python module.

Cross Validation. We implemented a k-fold cross validation procedure to increase robustness, as recommended by [20], and since our factorial design necessarily constrains the number of samples/situation/subject. Participants were randomly divided into three groups (folds) of seven participants each (we selected three since it evenly divides the 21 participants in the sample). Models (i.e. searchlight with LASSO-PCR) were iteratively trained on two groups and tested on the left-out group. The dot product of the weights from LASSO-PCR (fit to the training data) and activation data from the testing sample yields predicted fear ratings. Pearson correlations between the predicted fear ratings and the actual fear ratings were calculated and assigned to the center voxel of the spherical searchlight. The whole brain map of the correlation values between the predicted ratings and the observed ratings of the testing sample were calculated and averaged across three folds.

Permutation Testing and Statistical Correction. Permutation testing (n=1,200) was used to identify voxels with nonzero predictions. Models were trained and tested using shuffled data to generate a null distribution of correlation values. Comparing the original correlation values against the null distribution resulted in a statistical map that could be thresholded, using the p-values obtained for each voxel. A Familywise Error (FWE) rate of 0.05 was used to threshold the permutation [21,22].

Results

Different theoretical models motivate different analytical approaches for evaluating brain representations of fear. According to the situation dependent hypothesis, neural codes are expected to predict subjective fear in some situations, but not in others. In this case, it would be preferable to train the model situation by situation. According to the situation general hypothesis, neural codes are expected to predict subjective fear across situations. In this case, it would be preferable to train the model using data across situations. We performed analyses from both theoretical perspectives and examined which voxelwise neighborhoods contained situation dependent or situation general neural codes for predicting subjective fear. We defined a region as containing situation general information if it predicted fear across situations, and situation dependent if it predicted only one or two situations out of the three. Of note, we conceptually distinguish situation dependency from situation specificity; while our study can compare situation general v. situation dependent representations, it would require testing a much larger variety of situations to address whether a brain region contains situation specific information.

Situation Dependent Model Training. For situation dependent model training, we first trained and tested the searchlight with LASSO-PCR models situation by situation. This approach resulted in three Pearson correlation maps in which data from a given voxel's searchlight neighborhood was first trained to predict fear and then tested on left out data for every situation separately. As illustrated in Figure 1, brain regions containing information that predicts fear for each situation were widely distributed, consistent with prior work. To see if the same or different brain regions predicted fear depending on the situation, we performed a conjunction analysis (Figure 1).

A breakdown of the conjunction map showed that only 4.6% of the total voxels predicted fear in all three situations (Figure 1). Of the remaining voxels, 64.8% predicted fear in only one situation, and 30.5% predicted fear in two of the three situations. We present our findings using FWE corrected significance tests, however, we performed analyses across a range of lenient and stringent statistical thresholds to ensure that conclusions were robust across thresholding. The locations and proportions of voxels classified as situation general or situation dependent did not meaningfully change when using a more lenient threshold (see Supplementary Materials Figure S1A).

Even though this analytical approach is optimized from the situation dependent perspective, it is still possible that training in all three cases would identify some brain regions in common that predict fear. The conjunction analysis identified the posterior insula, superior temporal cortex, and somatosensory cortex as areas with information that might predict fear across situations (Figure 1B, red). However, regional overlap does not indicate whether the neural codes within these regions are also generalizable. We address this latter point in the next subsection.

Situation General Model Training. The situation general model assumes there are brain regions that contain situation invariant representations that predict fear. Thus, we trained the searchlight MVPA using data across all three situations. We then examined how well the model predicted fear for every situation using left out data. Importantly, this approach may lead to diverse outcomes. If an area relies on different weights for different situations, it is possible that even fewer areas will be characterized as situation general here than in the situation dependent analysis strategy, which shows spatial overlap (Figure 1) but does not test for generalization across shared neural codes. Alternatively, if a brain region contains situation general information in predicting fear, then training across situations may do a better job of identifying these codes rather than overfitting to idiosyncrasies of a particular situation. Thus, in this analysis, voxels that predict fear across situations also rely on the same codes and hence may contain situation general information. Alternatively, we may

find that even upon training with instances across situations, many of these areas nonetheless contain situation dependent codes.

Brain regions predicting fear were, again, widely distributed throughout cortical and subcortical areas (Figure 2B). A breakdown of the situation general map showed that only 0.9% of voxels met criteria for being situation-general. Of the remaining voxels, 37.1% predicted fear in one situation, whereas 62% predicted fear in two of three situations. While reported findings use FWE corrected significance tests, the proportions and locations of voxels classified as situation general and situation dependent did not meaningfully change when using a more lenient threshold (see Supplementary Materials, Figure S1B). Areas with information that predicted fear across situations included the right posterior insula and the right superior temporal cortex (Figure 2B, red). The remaining areas predicted fear in one or two situations.

Strikingly, in comparison to situation dependent model training, the nominal proportion of voxels that predicted fear in two situations more than doubled when using situation general training (i.e. from 30.5% to 62%). However, if anything, there was a nominal decrease in the proportion of voxels that contained situation general codes (i.e. from 4.6% to 0.9%). Thus, even when training our models using a situation general approach, the vast majority of brain regions nonetheless appeared to contain situation dependent information in predicting fear.

Discussion

A central question in affective neuroscience concerns how the brain represents subjective experiences of fear or emotions more generally. Here, we investigated the extent to which region-wise neural codes for fear are situation dependent or situation general. Answering this question is critical for determining whether the specific brain regions implicated in fear in a given study are likely to generalize to other contexts, and thus, may inform issues of external validity or generalization that have been at the crux of theoretical debates on emotion representation. Extending prior work, we combined a searchlight MVPA approach with a parametric experimental design to systematically examine the relationship between neural activity and subjective fear within and across three contexts. The findings showed that a small portion of voxel-wise neighborhoods (~1%) may contain situation general neural codes, whereas the overwhelming majority (~95-99%) of regional activation patterns contained situation dependent neural codes.

These results may help inform findings on human neuroimaging studies of fear that at first glance may appear to be paradoxical. Univariate analyses in fMRI studies have shown that different brain regions are associated with fear depending on the stimulus content [12–14], yet multivariate fMRI studies have identified neural signatures for states of fear or other specific emotion categories that may also generalize across stimulus types [5,6,8,9,23]. There are theoretical and methodological reasons that may address this issue. Theoretically, the goal of characterizing the representational architecture of emotion categories is distinct from a search for neural signatures of emotion categories. Indeed, it is possible to successfully classify or predict an emotion category from brain data irrespective of its internal structure [10,24]. Methodologically, this issue has generally been confounded with univariate and multivariate analytical approaches. Differences in activation magnitude across content conditions may have little bearing on the pattern of neural codes within an area. Our study addresses this issue by examining representations of fear using more recent, multivariate methods. In this regard, our study may help draw connections between these sets of studies. Consistent with studies searching for neural signatures of fear and emotion, we found that widely distributed brain regions were implicated in fear. But we also found that the overwhelming majority of these brain regions carried situation dependent neural codes.

Understanding how the brain represents subjective fear may provide guidance on how to develop more successful neural signatures in the long run. In particular, our findings inform future work on neural signatures of fear. Most studies search for category-level neural signatures of emotion. Here, our findings underscore the importance of training and testing models across diverse and heterogeneous situational contexts to promote generalization. However, our findings do raise the possibility that the functional heterogeneity of within-category representations may limit the success of this approach. Models that incorporate both category-level and more fine-grained representations of fear may ultimately provide the best prediction of fear experience, as is the case for neural signatures that predict the aversiveness of affective experience [25]. This issue may further be addressed by using machine learning approaches that accommodate non-linearities in the brain-behavior relationship [26,27].

Moving in another direction, some studies have begun to develop contextualized neural signatures of emotion. For instance, neural signatures have been developed based on stimulus modality for emotion (e.g. a neural signature for visually evoked fear [28] and also for valence [17,29–33] and pain [34,35]. Intriguingly, we only used the visual modality for stimulus delivery but found that the neural representations of fear nevertheless differentiated by situational content. Correspondingly, our findings suggest that contextualized neural signatures of emotion may benefit by considering situational content, too.

Questions regarding how the brain represents emotions lie at the crux of emotion theory [36]. Indeed, the present study was inspired by certain constructionist models of emotion which posit substantial within-category heterogeneity in neural representations of emotion [37–41]. According to this view, fear refers to a population category constituted from instances with diverse and heterogeneous features [37,42,43]. We reasoned that instances from similar situations are more likely to share features in common with each other, and thus may provide a useful heuristic to approximate subcategory representations of fear [40], which in turn serve as priors for conceptualizing future sensory inputs as yet another instance of fear. Technically, the theoretical model does not "hard-code" situations. It is possible for some instances of fear involving spiders to

be more similar to instances of fear involving heights than other instances of fear involving spiders. Consistent with this notion, a substantial portion of brain regions contained fear-predictive codes that generalized across two situations. These findings suggest that instances of fear may be better represented in a latent space [40,44,45]. While our study design was inspired by a constructionist approach, importantly our findings are also consistent with appraisal models that propose that diverse and heterogeneous patterns of appraisals may underlie a singular emotion category [46,47], albeit they are inconsistent with appraisal models that hypothesize a singular appraisal configuration for fear.

We also searched for brain regions that carried situation general neural codes. Few areas carried information that predicted fear across all three situations. These included the right posterior lateral temporal cortex and right posterior insula. These areas have been inconsistently implicated in prior MVPA studies on emotion experience (see Table S1). Expanding a bit to include MVPA studies on emotion perception, areas adjacent [44] or contralateral to [48] the posterior superior temporal sulcus have been implicated in classification in emotion perception tasks. It has been proposed that this region may encode appraisal dimensions [44] or semantic representations of emotion which may help support generalization across diverse instances. The posterior insula receives sensory inputs from the body and may play a more general role in arousal or interoception that is shared across emotions [49–53].

Our findings may also speak to theory on the relationship between fear and defensive behavior. It is well-known that freezing, fleeing, and fighting behaviors depend on predatory imminence [36,54,55]. These behaviors rely on a common macrolevel architecture involving the amygdala, hypothalamus, and periaqueductal gray among certain other areas with specific behaviors depending on more fine-grained circuitry within these regions. It remains hotly contested as to whether this macrolevel circuit should be functionally characterized as underlying "core fear circuitry" including subjective experiences of fear [54,56–58], or survival behaviors that are to a degree independent of fear [4,59].

Our results lead to a more nuanced conclusion. Although standard spatial resolution of fMRI cannot distinguish between these more fine-grained circuits, it may capture macro level functional properties of these regions. We found little evidence that this circuitry contains situation general neural codes in predicting fear, but that they do carry situation dependent fear-predictive codes (Figure 2). These findings suggest that circuitry underlying defensive behavior cannot be assumed to generalize to all cases of subjective fear [4,24,36], but this circuitry may also carry information pertaining to subjective fear in certain situations. Establishing boundary conditions for generalization would be a critical avenue for future work. While our study examined certain potent fears (e.g. fear of heights), one possibility is macrolevel architecture for defensive behavior may be implicated in fearful experiences when they share features of a predator-prey interaction [60] which we did not investigate.

Importantly, our study was optimized to investigate questions concerning heterogeneity in neural representations of fear, but not the specificity of those representations. Indeed, even brain regions we observed as carrying situation general neural codes may or may not carry fear-specific neural codes. Estimating specificity requires comparing neural representations of fear against other emotion categories (and perhaps also non-emotion categories). However, this has already been the primary focus of most previous MVPA studies and thus was not the main focus here. We also limited our experimental protocol to three situations in order to systematically examine variation in subjective fear within each situation. Doing so was crucial for extending beyond aforementioned limitations of prior studies while simultaneously ensuring enough samples to conduct a parametric analysis.

A better understanding of the neural representations of fear may be highly informative for advancing translational research [61]. While there has been an extensive focus on translating non-human animal models of defensive behavior toward understanding fear and anxiety disorders in humans [4,61,62], far less is known about how the brain constructs subjective experiences of fear and how these may vary across situations and individuals. Understanding this variation may be key to developing targeted and personalized treatments. While many translational studies examine the effects of interventions in a particular situation (e.g. on behavioral or subjective responses in anticipation of cutaneous pain) [63], our findings suggest it may be beneficial to test for external validity of findings in relation to the specific situations of clinical relevance for specific individuals.

References

1. Barrett, L. F. & Satpute, A. B. Historical pitfalls and new directions in the neuroscience of emotion. *Neurosci. Lett.* **693**, 9–18 (2019).

2. Lindquist, K. A., Siegel, E. H., Quigley, K. S. & Barrett, L. F. The hundred-year emotion war: are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). (2013).

3. Fanselow, M. S. & Pennington, Z. T. The danger of LeDoux and Pine's two-system framework for fear. *Am. J. Psychiatry* **174**, 1120–1121 (2017).

4. LeDoux, J. E. & Pine, D. S. Using neuroscience to help understand fear and anxiety: a two-system framework. *Am. J. Psychiatry* (2016).

5. Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G. & Just, M. A. Identifying emotions on the basis of neural activation. *PloS One* **8**, e66032 (2013).

6. Kragel, P. A. & LaBar, K. S. Multivariate neural biomarkers of emotional states are categorically distinct. *Soc. Cogn. Affect. Neurosci.* **10**, 1437–1448 (2015).

7. Kragel, P. A. & LaBar, K. S. Decoding the nature of emotion in the brain. *Trends Cogn. Sci.* **20**, 444–455 (2016).

8. Saarimäki, H. *et al.* Discrete neural signatures of basic emotions. *Cereb. Cortex* **26**, 2563–2573 (2016).

9. Saarimäki, H. *et al.* Distributed affective space represents multiple emotion categories across the human brain. *Soc.*

Cogn. Affect. Neurosci. **13**, 471–482 (2018).
10. 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 (2018).
11. Lindquist, K. A., Jackson, J. C., Leshin, J., Satpute, A. B. & Gendron, M. The cultural evolution of emotion. *Nat. Rev. Psychol.* 1–13 (2022).
12. Caseras, X. *et al.* Dynamics of brain responses to phobic-related stimulation in specific phobia subtypes. *Eur. J. Neurosci.* **32**, 1414–1422 (2010).
13. Lueken, U. *et al.* How specific is specific phobia? Different neural response patterns in two subtypes of specific phobia. *NeuroImage* **56**, 363–372 (2011).
14. Michalowski, J. M. *et al.* Neural response patterns in spider, blood-injection-injury and social fearful individuals: New insights from a simultaneous EEG/ECG–fMRI study. *Brain Imaging Behav.* **11**, 829–845 (2017).
15. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci.* **103**, 3863–3868 (2006).
16. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
17. 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 Biol.* **13**, e1002180 (2015).
18. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
19. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
20. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
21. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
22. Nichols, T. & Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* **12**, 419–446 (2003).
23. Wager, T. D. *et al.* A Bayesian model of category-specific emotional brain responses. *PLoS Comput. Biol.* **11**, e1004066 (2015).
24. Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Barrett, L. F. The brain basis of emotion: a meta-analytic review. *Behav. Brain Sci.* **35**, 121 (2012).
25. Čeko, M., Kragel, P. A., Woo, C.-W., López-Solà, M. & Wager, T. D. Common and stimulus-type-specific brain representations of negative affect. *Nat. Neurosci.* 1–11 (2022).
26. Azari, B. *et al.* Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Sci. Rep.* **10**, 1–17 (2020).
27. Khan, Z. *et al.* A computational neural model for mapping degenerate neural architectures. *Neuroinformatics* 1–15 (2022).
28. Zhou, F. *et al.* A distributed fMRI-based signature for the subjective experience of fear. *Nat. Commun.* **12**, 1–16 (2021).
29. Chikazoe, J., Lee, D. H., Kriegeskorte, N. & Anderson, A. K. Population coding of affect across stimuli, modalities and individuals. *Nat. Neurosci.* **17**, 1114–1122 (2014).
30. Kim, H.-C., Bandettini, P. A. & Lee, J.-H. Deep neural network predicts emotional responses of the human brain from functional magnetic resonance imaging. *NeuroImage* **186**, 607–627 (2019).
31. Kim, J., Shinkareva, S. V. & Wedell, D. H. Representations of modality-general valence for videos and music derived from fMRI data. *NeuroImage* **148**, 42–54 (2017).
32. Lee, K. M., Ferreira-Santos, F. & Satpute, A. B. Predictive Processing Models and Affective Neuroscience. *Neurosci. Biobehav. Rev.* (2021).
33. Miskovic, V. & Anderson, A. Modality general and modality specific coding of hedonic valence. *Curr. Opin. Behav. Sci.* **19**, 91–97 (2018).
34. Krishnan, A. *et al.* Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *elife* **5**, e15166 (2016).
35. Wager, T. D. *et al.* An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
36. Mobbs, D. *et al.* Viewpoints: Approaches to defining and investigating fear. *Nat. Neurosci.* **22**, 1205–1216 (2019).
37. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23 (2017).
38. Ekman, P. Are there basic emotions? (1992).
39. Riskind, J. H. Looming vulnerability to threat: A cognitive paradigm for anxiety. *Behav. Res. Ther.* **35**, 685–702 (1997).
40. Satpute, A. B. & Lindquist, K. A. The default mode network's role in discrete emotion. *Trends Cogn. Sci.* **23**, 851–864 (2019).
41. Wilson-Mendenhall, C. D., Barrett, L. F. & Barsalou, L. W. Variety in emotional life: within-category typicality of emotional experiences is associated with neural activity in large-scale brain networks. *Soc. Cogn. Affect. Neurosci.* **10**, 62–71 (2015).
42. Barrett, L. F. *How emotions are made: The secret life of the brain*. (Pan Macmillan, 2017).
43. Siegel, E. H. *et al.* Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychol. Bull.* **144**, 343 (2018).
44. Skerry, A. E. & Saxe, R. Neural representations of emotion are organized around abstract event features. *Curr. Biol.* **25**, 1945–1954 (2015).
45. Tamir, D. I., Thornton, M. A., Contreras, J. M. & Mitchell, J. P. Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proc. Natl. Acad. Sci.* **113**, 194–199 (2016).
46. Meuleman, B. & Scherer, K. R. Nonlinear appraisal modeling: An application of machine learning to the study of emotion production. *IEEE Trans. Affect. Comput.* **4**, 398–411 (2013).

47. Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. Appraisal theories of emotion: State of the art and future development. *Emot. Rev.* **5**, 119–124 (2013).
48. Peelen, M. V., Atkinson, A. P. & Vuilleumier, P. Supramodal representations of perceived emotions in the human brain. *J. Neurosci.* **30**, 10127–10134 (2010).
49. Carvalho, G. B. & Damasio, A. Interoception and the origin of feelings: A new synthesis. *BioEssays* **43**, 2000261 (2021).
50. Craig, A. D. How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* **3**, 655–666 (2002).
51. Craig, A. D. How do you feel—now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10**, 59–70 (2009).
52. Damasio, A. R. *The feeling of what happens: Body and emotion in the making of consciousness*. (Houghton Mifflin Harcourt, 1999).
53. Kleckner, I. R. *et al.* Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nat. Hum. Behav.* **1**, 1–14 (2017).
54. Fanselow, M. S. Neural organization of the defensive behavior system responsible for fear. *Psychon. Bull. Rev.* **1**, 429–438 (1994).
55. Fendt, M. & Fanselow, M. S. The neuroanatomical and neurochemical basis of conditioned fear. *Neurosci. Biobehav. Rev.* **23**, 743–760 (1999).
56. Mobbs, D. *et al.* From threat to fear: the neural organization of defensive fear systems in humans. *J. Neurosci.* **29**, 12236–12243 (2009).
57. Panksepp, J. The basic emotional circuits of mammalian brains: do animals have affective lives? *Neurosci. Biobehav. Rev.* **35**, 1791–1804 (2011).
58. Panksepp, J., Fuchs, T. & Iacobucci, P. The basic neuroscience of emotional experiences in mammals: The case of subcortical FEAR circuitry and implications for clinical anxiety. *Appl. Anim. Behav. Sci.* **129**, 1–17 (2011).
59. LeDoux, J. E. & Brown, R. A higher-order theory of emotional consciousness. *Proc. Natl. Acad. Sci.* **114**, E2016–E2025 (2017).
60. Mobbs, D. *et al.* When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. *Science* **317**, 1079–1083 (2007).
61. Griebel, G. & Holmes, A. 50 years of hurdles and hope in anxiolytic drug discovery. *Nat. Rev. Drug Discov.* **12**, 667–687 (2013).
62. Bourin, M. Animal models for screening anxiolytic-like drugs: a perspective. *Dialogues Clin. Neurosci.* (2022).
63. Choy, Y., Fyer, A. J. & Lipsitz, J. D. Treatment of specific phobia in adults. *Clin. Psychol. Rev.* **27**, 266–286 (2007).
64. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Figures

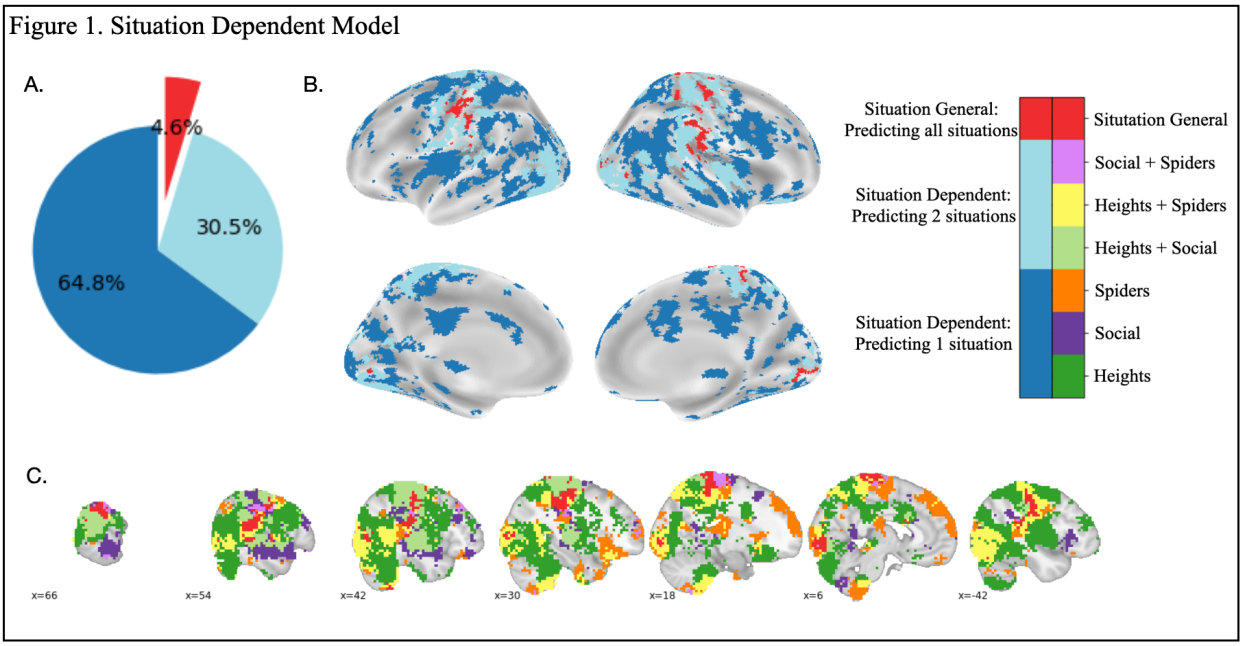

Figure 1 The situation dependent model: the proportion of situation general and situation dependent voxels. The situation dependent model was trained separately on each situation and tested to predict the same situation, resulting in three maps indicating voxels predictive of fear in heights, social interactions, and spiders. We ran a conjunction analysis on the three maps to compare whether each voxel is indicated in significantly predicting one, two or all three situations. The pie chart showed the percentage of the voxels in the total significant voxels. 64.8% of the voxels predicted fear in one situation (dark blue), 30.5% of voxels predicted fear in two situations (light blue). 4.6% of voxels predicted all three situations. The situation general area included that posterior insula, somatosensory areas showed to contain situation general (red). Bottom: Breakdown of the neural pattern by each situation and their combinations.

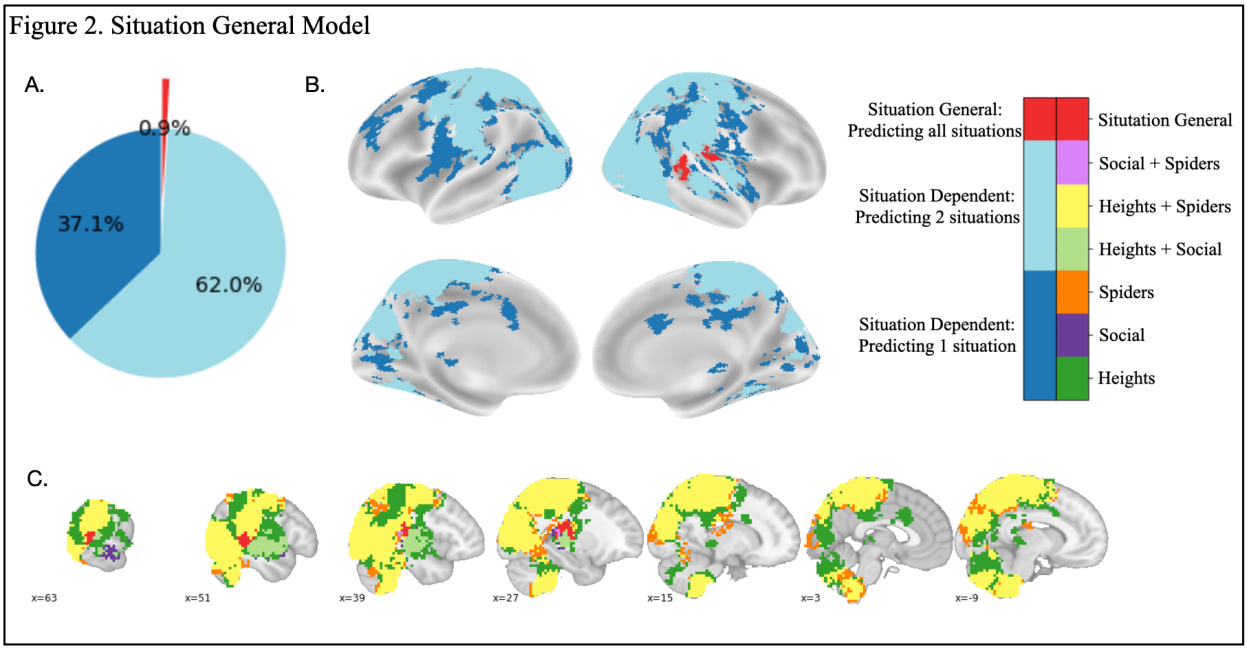

Figure 2 The situation general model: the proportion of situation general and situation dependent voxels. The situation-general model was trained in a sample of all three situations, and was applied to predict on each of the situations. The conjunction analysis was run on the three resulting maps to compare whether each voxel is indicated in significantly predicting one, two or all three situations. The pie chart showed the percentage of the voxels in the total significant voxels. 62.0% of the voxels predicted fear in one situation (dark blue), 37.1% of voxels predicted fear in two situations (light blue). Only 0.9% of voxels predicted all three situations. The situation general area included that superior temporal cortex and posterior insula areas showed to contain situation general (red). Bottom: Breakdown of the neural pattern by each situation and their combinations.