UNIVERSITY OF CALIFORNIA

Los Angeles

Understanding fear and threat responding in the human brain

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Psychology

by

Cody Alexander Cushing

2022

ABSTRACT OF THE DISSERTATION

Understanding fear and threat responding in the human brain

by

Cody Alexander Cushing

Doctor of Philosophy in Psychology, University of California, Los Angeles, 2022

Professor David Clewett, Chair

Goal: The goal of this dissertation is to investigate threat and fear responses in humans in order to understand both the multifaceted nature of these responses and how they can go awry in fear and anxiety related disorders. By understanding the mechanisms behind fear and anxiety disorders, effective treatments can be designed that minimize distressing or panic-inducing experiences that too often lead to attrition from the clinic. These issues are investigated across three studies utilizing behavioral tasks and neuroimaging via functional magnetic resonance imaging (fMRI) with the following aims:

Aim 1. Establishing efficacy of decoded neurofeedback as an intervention for specific phobia in a clinical cohort. There is an unmet need for non-distressing treatments for fear-related disorders like specific phobia where exposure therapy is the gold standard. While the effectiveness of exposure therapy should not be diminished, it is an inherently distressing experience leading to high rates of attrition and a need for alternative options for those who can not tolerate the experience. In the first study of this dissertation, I tested multi-voxel neuro-reinforcement as an intervention for specific phobia in a randomized double-blind placebo-controlled clinical trial. I found evidence for reduced amygdala activation to phobia as well as less attentional capture by the target phobia in an affective stroop task post-treatment.

Aim 2. Classifying the subjective awareness of threat from multi-voxel patterns. Most studies of threat processing have relied on direct contrasts between conditioned stimulus types as a proxy for threat detection. However, such contrasts fail to take individual differences in threat learning success and generalization into consideration. In the second study of this dissertation, I use machine-learning techniques to classify the subjective awareness of threat from whole-brain multi-voxel patterns. Additionally, I classify threat awareness iteratively within brain regions to identify which brain regions are most critical for threat awareness.   These analyses reveal a distributed brain response to threat that is organized hierarchically along the visual stream. Results are also characterized in terms of self-reported participant symptomatology.

Aim 3. Investigating brain networks involved in acquisition, extinction, and recall of learned threat. While many brain regions have been implicated in the formation of threat memories, the whole-brain dynamics of learned threat in humans are still poorly understood. In the third study of this dissertation, I applied group independent component analysis to whole-brain fMRI data. I identified brain networks involved in the acquisition, extinction, and recall of learned threat in a Pavlovian threat conditioning paradigm. This revealed stable networks across task phases that responded in opposing fashions across the same timescale. Results from this study highlight the multitude of parallel processes and networks that are engaged in human threat detection and learning.

The dissertation of Cody Cushing is approved.

Michelle Craske

Hongjing Lu

Megan Peters

David Clewett, Committee Chair

University of California, Los Angeles

2022

I dedicate this dissertation to
my parents, brother, and all my family and friends
that supported and inspired me to reach this point.

To the loving memories of:
Colman Nakano
Tristan "TJ" Townsend
Taylor Townsend

"Let them all play again, in some other way, and let them be happy." - Philip K. Dick

Table of Contents

List of Figures

List of Tables

Acknowledgments

Parts of this research were supported by the following:

UCLA Graduate Research Mentorship Award
UCLA Edwin W. Pauley Fellowship

I would like to thank all of my mentors, collaborators, colleagues, and friends who have helped shape this work through their generous support, feedback, and guidance. Thanks to:

My graduate advisor and committee chair Prof. David Clewett, dissertation committee, Prof. Michelle Craske, Prof. Hongjing Lu, Prof. Megan Peters; Prof. Hakwan Lau for advising me in my first 4 years of grad school and continued support; Prof. Mitsuo Kawato, Prof. Vincent Taschereau-Dumouchel, Prof. Brian Odegaard, Prof. Aurelio Cortese, Prof. Yujia Peng, Prof. Jesse Rissman, Zachary Anderson, Prof. Robin Nusslock, Prof. Joseph E. LeDoux, Prof. Stefan G. Hofmann, Dr. Jeffrey D. Knotts, Dr. Matthias Michel, Dr. Toshinori Chiba, Dr. Jesse Taylor, Dr. Ben Smith, Dr. Ben Rosenberg, Dr. Taylor Webb, Dr. Cathie Tsz Yan So, Dr. Seong Hah Cho, Dr. Brian Maniscalco, Dr. Piercesare Grimaldi, Prof. Jorge Morales, Prof. Kayuet Liu, Prof. Alan Lee, Prof. Kiyofumi Miyoshi, Prof. Kestas Kveraga, Prof. Reginald B. Adams Jr., Prof. Hee Yeon Im, Dr. David De Vito, Prof. Nouchine Hadjikhani, Prof. Avniel Ghuman, Prof. Troy Steiner, Dr. Jasmine Boshyan, Dr. Daniel Albohn, Prof. Gideon P. Caplovitz, Dr. J. Daniel McCarthy, Prof. Christopher Blair, Dr. Kyle Killibrew, Dr. Gennadiy Gurariy, Dr. Gennady Erlikhman, Lisa Lee, Kaori Nakamura, Mieko Hirata, Freya Vaughn, Mary Susselman, Trent Thixton, Mouslim Cherkaoui, Sivananda Rajananda, Raihyung Lee, Andrew Eastwick, Angela Clague, Ana Costello, Annelise Murillo, Shawn Wang, Shiyun Wang, and Marlene Berke.

Special thanks to my partner Amanda Baker for supporting and encouraging me through this work.

Vita

Education

University of California, Los Angeles, Los Angeles, CA **2018**

M.A. Psychology

University of Nevada-Reno, Reno, Nevada **2015**

B.S. Neuroscience, *Cum Laude*

B.A. Philosophy, *Cum Laude*

Minor: Biophysical-Organic Chemistry

Publications and Papers

Cushing, C. A., Im, H. Y., Adams, R. B., Jr., Ward, N., & Kveraga, K. (2019). Magnocellular and parvocellular pathway contributions to facial threat cue processing. *Social Cognitive and Affective Neuroscience.* https://doi.org/10.1093/scan/nsz003

Cushing, C. A., Im, H .Y., Adams, R. B. , Jr., Ward, N., Albohn, D. N., Steiner, T. G. & Kveraga, K. (2018). Neurodynamics and connectivity during facial fear perception: The role of threat exposure and signal congruity.. *Scientific Reports.* https://doi.org/10.1038/s41598-018-20509-8

Cushing, C. A.*, Cho, S.H.*, Patel, K., Kothari, A., Lan, R., Michel, M., Cherkaoui, M., Lau, H. (2018). Blockchain and human episodic memory. *Pre-print on arXiv:* https://arxiv.org/abs/1811.02881 *co-first author

Taschereau-Dumouchel, V., **Cushing, C. A.**, Lau, H. (2022). Real-time fMRI in the treatment of mental health disorders. *Annual Review of Clinical Psychology.* https://doi.org/10.1146/annurev-clinpsy-072220-014550

Yeung, A. W. K., **Cushing, C. A.**, Lee, A. L. F. (2022). A bibliometric evaluation of the impact of theories of consciousness in academia and on social media. *Consciousness and Cognition.* https://doi.org/10.1016/j.concog.2022.103296

Im, H.Y., **Cushing, C.A.**, Ward, N., & Kveraga, K. (2021). Differential neurodynamics and connectivity in the dorsal and ventral visual pathways during perception of emotional crowds and individuals: a MEG study. *Cognitive, Affective, and Behavioral Neuroscience*.

Adams, Jr., R., Im, H. Y., **Cushing, C. A.**, Boshyan, J., Ward, N., Albohn, D. N., & Kveraga, K. (2019). Differential magnocellular versus parvocellular pathway contributions to the combinatorial processing of facial threat. *Progress in Brain Research,* Elsevier B.V.. https://doi.org/10.1016/bs.pbr.2019.03.006

Kveraga, K., De Vito, D., **Cushing, C. A.,** Im, H. Y., Albohn, D. N., Adams, R. B., Jr., (2019). Spatial and feature-based attention to expressive faces. *Experimental Brain Research.* https://doi.org/10.1007/s00221-019-05472-8

Im, H.Y., Adams, R. B. Jr., **Cushing, C. A.,** Boshyan, J., Ward, N. & Kveraga, K. (2018). Sex-related differences in behavioral and amygdalar responses to compound facial threat cues. *Human Brain Mapping*. doi: 10.1002/hbm.24035.

Im, H.Y., Adams, R.B., Jr., Boshyan, J., Ward, N., **Cushing, C. A.**, Kveraga, K. (2017) Observer's anxiety facilitates magnocellular processing of clear facial threat cues, but impairs parvocellular processing of ambiguous facial threat cues. *Scientific Reports.* doi: 10.1038/s41598-017-15495-2

Im, H. Y., Albohn, D. N., Steiner, T. G., **Cushing, C. A.,** Adams, R. B. Jr., & Kveraga, K. (2017) Differential hemispheric and visual stream contributions to ensemble coding of crowd emotion. *Nature Human Behavior*. doi:10.1038/s41562-017-0225-z

Conference presentations

Talks

Cushing, C.A., Lau, H. (2019). Unconscious learning of value in a model-based paradigm with decoded neurofeedback. Talk given at 23nd annual meeting of the Association for the Scientific Study of Consciousness, London, Ontario, CA.

Cushing, C.A., Lau, H. (2018). Visual pathways differentially modulate subjective confidence. Talk given at 22nd annual meeting of the Association for the Scientific Study of Consciousness, Krakow, Poland.

Posters

Cushing, C. A., Cherkaoui, M., Rissman, J., Kawato, M., Lau, H. (2019). Visual representations outside of conscious awareness can support sensory preconditioning. Poster presented at Vision Sciences Society Annual Meeting, St. Pete Beach, FL.

Cushing, C. A., Adams, R. B., Jr., Im, H. Y., Ward, N., & Kveraga, K. (2017). Differential visual pathway contributions to compound facial threat cue processing. Poster presented at Vision Sciences Society Annual Meeting, St. Pete Beach, FL.

Cushing, C. A., Adams, R. B. Jr., Im, H. Y., Ward, N., & Kveraga, K. (2016). Neurodynamics and connectivity during compound threat cue perception. Poster presented at 20th International Conference on Biomagnetism, Biomag, Seoul, Korea.

Cushing, C. A., Adams, R. B. Jr., Im, H. Y., Ward, N., & Kveraga, K. (2016). Neurodynamics of facial threat cue perception modulated by anxiety: A MEG study. Poster presented at Vision Sciences Society Annual Meeting, St. Pete Beach, FL.

Cushing C. A., & Caplovitz, G.P. (2015). Neural Correlates of Illusory Contour Formation. Poster presented at Undergraduate Research Poster Conference, Reno, NV.

Cushing C. **A.**, McCarthy J. D., & Caplovitz G. P. (2014). The Neural Time Course of Spatiotemporal Form Integration. Poster presented at Annual Psi Chi Poster Session, Reno, NV.

Invited Talks

Using decoded neurofeedback to treat phobia and other mental health disorders. *Early-career seminar, University of Nevada, Reno, Feb. 2021*

Fellowships and Awards

Graduate Research Mentorship Award, UCLA	**2020-2021**
Graduate Summer Research Mentorship Award, UCLA	**2018**
Edwin W. Pauley Fellowship, UCLA	**2017-2021**
General Undergraduate Research Award, UNR	**2014-2015**

Chapter 1. Introduction

For decades, attempts have been made to pinpoint neural mechanisms in anxiety and fear-related disorders. Yet, anxiety, phobia, post-traumatic stress disorder (PTSD), and other related disorders still do not have a clearly understood neuropathophysiology. This difficulty may partly stem from the subjective experience of fear itself not having received adequate attention in the search for these neural mechanisms and effective treatments (LeDoux & Hofmann, 2018). As such, most effort towards reducing fear disorders to a manageable level has relied on behaviorally targeting the object or circumstances of fear rather than the experience of fear itself. However, exposure to the target of fear typically leads to a distressing and uncomfortable experience for patients. This leaves an unmet need for treatment options that are both effective and non-distressing to those seeking treatment. By understanding the different foundational aspects of fear and threat responding in the human brain, individually tailored treatments can be developed that potentially minimize distressing patient experiences.

The difficulty in treating fear disorders is not solely for lack of effective treatment options. One currently popular treatment is 'exposure therapy' which involves direct exposure to fear-causing or panic-inducing stimuli. Exposure therapy attempts to extinguish or counter-condition the existing fear to be associated with something more positive. Inherently, this is a disturbing and unpleasant experience. Despite the ultimate effectiveness for individuals completing treatment, there are high rates of attrition due to discomfort (Loerinc et al., 2015; Zayfert et al., 2005). However, a promising new treatment using a fMRI method called multi-voxel neuro-reinforcement has demonstrated the ability to lessen fear responses without direct exposure to fear-inducing stimuli (Koizumi et al., 2017; Taschereau-Dumouchel et al., 2018). Fear-response reduction is achieved through a kind of 'unconscious exposure' (Chiba et al., 2019). By using a machine-learning classifier to 'decode' online BOLD activity from patients in the scanner, neuro-reinforcement can be provided based on a specific stimulus category (e.g. spider). Analyzing patterns across voxels rather than average brain activity alone enables more specific neural targets. Importantly, this can be done outside of conscious awareness.

Anxiety-related disorders also involve these same feelings of fear but generalized over time and place. The definition of a clear treatment target for something like exposure can be more difficult in anxiety disorders. Though exposure therapy can not always be applied in the same way, a useful target for multi-voxel neuro-reinforcement may still be found. Under the fear responses seen in fear and anxiety-related disorders likely lies a unifying mechanism. This mechanism has been theorized to involve threat conditioning processes with fear and anxiety-related disorders arising from dysfunctional threat learning (Craske et al., 2017; Fenster et al., 2018). A wealth of evidence supports this notion with differences in traditional threat learning processes being observed in phobia (Lange et al., 2019), anxiety disorders (Marin et al., 2017; Pittig et al., 2018), and PTSD (Bremner et al., 2005; Hennings et al., 2020; Milad et al., 2009). However, the consequences of these differences for designing treatments have not been so immediately clear. The reasons for this are likely also stemming from non-clarity surrounding just what the fear response is (LeDoux & Hofmann, 2018).

As such, much of the focus has been placed on low-level threat response processes in regions such as the amygdala and hippocampus. But modulating these low-level regions alone may not be sufficient to impact a patient’s distressing subjective experience (Taschereau-Dumouchel et al., 2018, 2022). Indeed, at the human level it is not so clear exactly how necessary classic translational targets like the amygdala, which has been popularized as the “fear center”, are for threat learning as it is at times not detected during threat acquisition in human neuroimaging studies (Visser et al., 2021). Additionally, amygdala lesions do not seem to alter subjective experiences of either positive or negative affect (Anderson & Phelps, 2002) nor do they prevent the experience of fear or panic (Feinstein et al., 2013).

However, these low-level processes are not to be trivially dismissed either. They are surely a critical piece in understanding how threats are detected and how new threats are learned in the human brain. Additionally, they are undoubtedly foundational to how fear and anxiety-related disorders develop. The role of each process just needs to be understood as it contributes to things like behavior, physiological responses, or subjective experiences. The search for brain mechanisms behind these phenomena must also be expanded beyond a handful of focused regions as brain responses related to subjective experience tend to be distributed across the entire cortex (Taschereau-Dumouchel et al., 2019; Zhou et al., 2021).

As an early response to the inconsistencies surrounding the term "fear" in the scientific literature, Peter Lang proposed the "three-systems model" of fear (Lang et al., 1983). This divided the fear response into three systems: verbal (as a proxy for cognition), motor, and somatic. The theory purported fear and anxiety express themselves through these 3 systems and therapy should seek to alter a system specifically in order to impact the fear response. More modern approaches have sought to put subjective experience in the front seat in place of verbal report as methods to investigate subjective experience have come to fruition through technical and experimental development (Lau, 2022; Taschereau-Dumouchel et al., 2022). A modern understanding of the brain mechanisms behind anxiety and fear needs to understand the mechanisms behind both subjective experience and low-level threat processes alike as well as their influence on behavior and physiology.

Dissertation Overview

This dissertation is divided into 3 main chapters (Chapters 2-4). Chapter 2 describes a randomized double-blind placebo-controlled clinical trial of multi-voxel neuro-reinforcement as a clinical intervention for specific phobia. In it, real-time activations of nonconscious visual representations (e.g. spider) are paired with reward during an fMRI scan. Patients are assessed pre-treatment and post-treatment with a fear rating and affective stroop task while in the fMRI scanner. The primary hypothesis is that multi-voxel neuro-reinforcement leads to selective reduction in threat responses to the phobia targeted for intervention.

Chapter 3 investigates multi-voxel fMRI response patterns responsible for representing the subjective awareness of threat in a large cohort of participants in a Pavlovian threat conditioning task. The paradigm is a 2-day three-phase procedure with participants completing threat acquisition and extinction on day 1 and extinction recall on day 2 at least 48 hours later. Whole-brain classification performance is assessed in a cross-validation procedure. Within-region classification is performed across a parcellation of the cortex to identify brain regions that contain significant information regarding the subjective awareness of threat. Generalization of classification performance is assessed by applying the developed machine-learning classifier to a large independent dataset of participants that completed a similar Pavlovian threat conditioning paradigm. Participant symptomatology is considered as it relates to classifier evidence of threat awareness during extinction memory recall.

Chapter 4 utilizes the same Pavlovian threat conditioning paradigm and describes a group network analysis using group ICA analysis of fMRI data. Using group independent component analysis, brain networks involved in the acquisition, extinction, and retention of extinction memory are explored. Networks that are common across multiple phases of threat learning are given focus.

Chapter 2. Establishing efficacy of decoded neurofeedback as an intervention for specific phobia in a clinical cohort

Introduction

Fear disorders such as specific Phobia and Post-traumatic stress disorder (PTSD) are among the most difficult mental disorders to treat. The current best treatment is ‘exposure therapy’ which involves direct exposure and experience of fear-causing or panic-inducing stimuli. This is done in an attempt to extinguish existing fear or to counter-condition the fear-evoking stimuli to be associated with something more positive. Inherently, this is a disturbing and unpleasant experience for the patient undergoing treatment, leading to high rates of attrition (Loerinc et al., 2015; Zayfert et al., 2005).

However, a promising new treatment using a fMRI method called multi-voxel neuro-reinforcement has demonstrated the ability to lessen fear responses to both lab-conditioned fears and pre-existing fears through a kind of ‘unconscious exposure’ (Koizumi et al., 2017; Taschereau-Dumouchel et al., 2018). By using a machine-learning classifier (also referred to as a ‘decoder’) to ‘decode’ online BOLD activity from patients in the scanner, neuro-reinforcement can be provided based on a specific stimulus category (e.g. spider) rather than average brain activity alone.

Importantly, this can be accomplished at an implicit nonconscious level as participants undergoing neuro-reinforcement are simply trying to make a feedback disc on the screen grow in size with no specific instruction as to what makes the disc grow. As they are unaware of the relation of the feedback score to the feared stimulus category (e.g. spider), their brain is able to learn to activate a nonconscious representation of the feared stimulus outside of the patient’s awareness. Critically this results in no subjective discomfort for the patient.

The procedure is based on reinforcement learning. When a brain pattern is induced, it is paired with reward. Through this process, either exposure or counter-conditioning effects alter neural and behavioral responses to feared stimuli. The exact mechanism is not yet understood, but early results are consistent with an exposure effect over a counter-conditioning mechanism (Chiba et al., 2019). Regardless of the mechanism, multi-voxel neuro-reinforcement has shown early promise as a clinical intervention that can be applied outside of conscious awareness, eliminating stressful conscious exposures. Essentially any neural pattern that can be identified reliably with multivariate-pattern analysis (MVPA) can be used as a target for intervention.

Typically the construction of such a machine-learning classifier to decode visual representations in a patient's brain involves repeated visual presentation of the representation attempting to be decoded. This would seemingly nullify the entire appeal of the multi-voxel neuro-reinforcement procedure. However, recent advances in fMRI methodology have enabled the functional alignment of fMRI brain data allowing brain data to be moved from the native space of one person into another (Haxby et al., 2011). Functional alignment is thought to be superior to simple anatomical registration based on structural landmarks as cortical regions tend to be more functionally organized rather than structurally organized. Aligning fMRI data functionally results in superior between-subject decoding for functionally aligned data over structurally aligned data (Haxby et al., 2011). The functional organization of the cortex is why popular modern parcellations of the human brain have begun to be based on functional connectivity patterns rather than alignment of brain structures (Schaefer et al., 2018).

By leveraging functional alignment approaches, a decoder can be built for a phobic patient's brain using brain data from a non-phopic control for whom viewing repeated images of a target representation (e.g. spider) produces no stress. The phobic patient simply needs to undergo a similar task (minus the phobic images) while fMRI data are collected in order to calculate the necessary functional alignment. This provides the opportunity to produce a "nonconscious exposure" in phobic patients that have not had to interact with the feared stimulus in any capacity other than non-distressing activation of the target representation at a nonconscious level.

The specificity of the decoder also allows the opportunity for a within-subject placebo control provided the patient has more than one phobia. For example, if a patient has a snake and a spider phobia, a decoder can be built specifically for spiders while snakes remain a placebo control. Such within-subject placebo controls are not possible with other forms of neurofeedback as something like increasing univariate BOLD signal within a region of interest cannot be specifically related to one image category. Here, I describe a double-blind placebo-controlled clinical trial of this method as an intervention in a population with specific phobia.

Methods

Participants

For multi-voxel neuro-reinforcement, phobic patients (N=18) with at least 2 specific phobias were enrolled for treatment. Eligibility was confirmed and phobias diagnosed via an ADIS-V (T. Brown & Barlow, 2014) interview conducted by a certified researcher. Exclusion criteria were any MRI contraindications or meeting the diagnostic criteria for Post-traumatic Stress Disorder, Obsessive Compulsive Disorder, Substance Use Disorder, current Major Depressive Disorder, Bipolar Disorder, or Psychosis. Patients were randomly assigned to complete either 1, 3, or 5 days of multi-voxel neuro-reinforcement to determine the dose-response relationship between neuro-reinforcement and clinical outcomes.

Decoder Construction

Prior to neuro-reinforcement, a between-subject machine learning classifier was trained for the target phobic image category (Fig. 1). To eliminate the need for any exposure to phobic stimuli to receive treatment, the classifier was constructed using brain data from healthy controls using a process called hyperalignment (Haxby et al., 2011). During an initial fMRI session (Fig. 2), each healthy control (N=28) viewed the same image dataset of 3600 images consisting of 40 categories of animals and objects (e.g. birds, butterflies, snakes, spiders). Conversely, phobic patients (N=18) viewed the same image dataset but with their specific phobias removed to avoid unnecessary exposure. In place of phobic images, phobic patients viewed happy human faces using stimuli from the Chicago Face Database and NimStim Set of Facial Expressions (Ma et al., 2015; Tottenham et al., 2009). These stimuli have their emotional expression verified by independent raters and were used to provide a non-disturbing stimulus replacement that was sufficiently orthogonal to the task image set of animals and objects. The decoder construction task consisted of 6 runs of 600 trials each. Each trial consisted of a .98 second image presentation with no inter-trial interval. This rapid event-related design was used to maximize the number of images each participant viewed. To ensure attention, participants were given the task of pressing a button each time the image category changed (i.e. a 1-back task). Image categories were presented in chunks of 2, 3, 4, or 6 consecutive images.

Decoder construction fMRI data were processed using a combination of SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm) and custom python scripts using pyMVPA and sklearn packages (Hanke, Halchenko, Sederberg, Olivetti, et al., 2009; Pedregosa et al., 2011). All 6 runs of the task were concatenated and preprocessed in SPM using default parameters

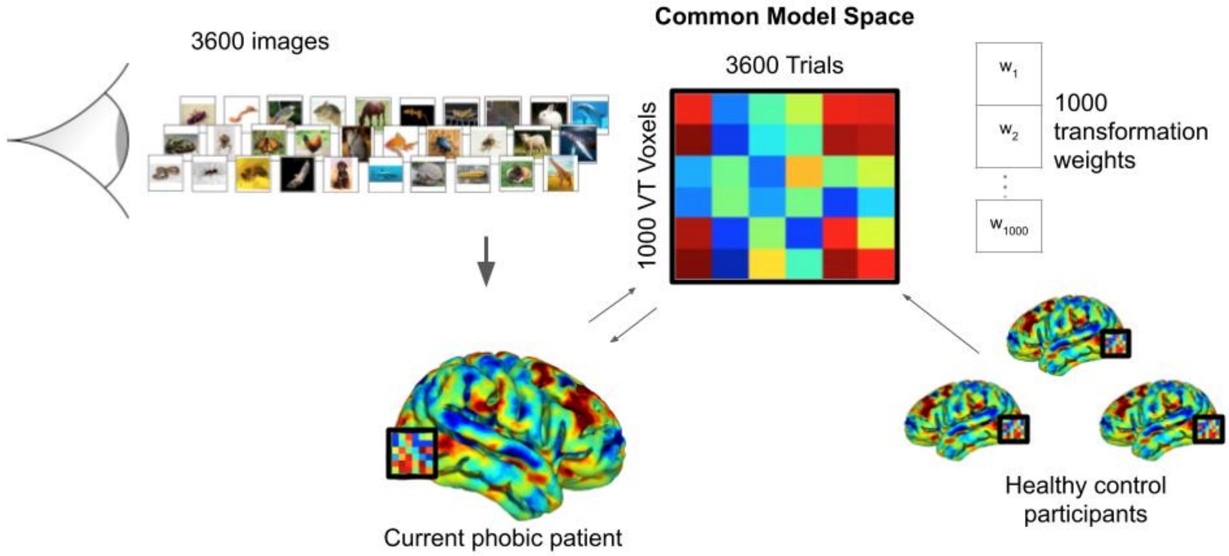

Figure 1. Functional alignment of brain data into phobic patient brain using hyperalignment. All participants complete a near-identical task in the fMRI scanner where 3600 images are rapidly viewed during 1 second presentations. Phobic patients view happy human faces instead of their own phobic categories. Healthy controls view images from all categories. Transformation parameters into the functionally aligned common model space are determined with phobic image trials withheld. Data from all participants for all categories (including phobic categories) are transformed into the common model space and then reverse transformed into the native space of the current phobic participant. A machine-learning classifier can then be trained on phobic images in the patient's native brain space despite the patient never having personally viewed the images.

unless otherwise explicitly specified. Data were realigned to the first image from the first run of the task and segmented into tissue classes. Anatomical and functional data were coregistered using the gray matter image from segmentation as a reference. Motion was then regressed out of the functional data using the parameters from realignment. Single-trial estimates were then generated with pyMVPA using the least-squares 2 (LS-2) method (Turner et al., 2012) in which a separate GLM is computed for each trial where the current trial is assigned to one regressor while the remaining trials are equally split between two "rest" regressors.

Using hyperalignment, single-trial estimates from healthy controls in the target brain region (ventral temporal cortex) were functionally transformed to the current phobic patient's brain and used to train a machine-learning pattern classifier (decoder) using the phobic images that the patient did not see (Fig. 1). To ensure double-blind treatment target selection, the target for treatment was automatically selected by a computer program that calculated which phobic category had the highest cross-validated area under the receiver operating characteristic curve (AUC) for binary classification.

To determine AUC metrics, a 6-fold cross-validation (CV) procedure was used. FMRI data for each participant were loaded and masked to the ventral temporal (VT) area in their own native space using an anatomical mask derived from Freesufer parcellations of the fusiform, lingual, parahippocampal, and inferior temporal areas (Fischl et al., 2004). Single-trial parameter estimates were standardized by feature within subject and within each of the 6 task runs. The data were split into 6 folds for training and testing based on the 6 runs completed by each participant. That is, for each CV split, the withheld testing set consisted of all the data from each participant for one of the six task runs. The remaining preprocessing was calculated using only the training data to avoid overfitting. As hyperalignment requires a stable number of features across participants, 1000 voxels were selected within the VT area via F-test to select which voxels accounted for the most variance elicited by all image categories across all training trials. For each phobic participant, a unique set of hyperalignment transformation parameters into the common model space was calculated for the current phobic participant and all healthy controls. The fitting of the hyperalignment parameters was done using trials for all image categories except the current patient’s phobia. For example, if a phobic patient had spider and snake phobias, all spider and snake trials were withheld from all participants when fitting the transformation parameters.

After hyperalignment transformation parameters were determined, the data from all healthy controls were moved into the native space of the current phobic patient by transforming the data into the common model space and then reverse transforming the data from the common model space into the native space of the current patient. The transformed data included the previously withheld phobic category images from the healthy controls as well as the testing dataset.

With all data in the current patient's native space, class sizes (target vs. non-target image categories) were balanced by random undersampling balanced between the 39 non-target image categories. Following previous work (Taschereau-Dumouchel et al., 2018), a Sparse Multinomial Logistic Regression (SMLR) classifier was trained to perform binary (one-vs-rest) classification between the potential target category and all remaining categories (Krishnapuram et al., 2005). AUC scores for each CV split were calculated based on classifier estimates.

Of the potential phobic categories to be selected for treatment for the current patient, the phobia with the highest AUC scores across all 6 CV splits was blindly selected via computer program as the target for treatment. The within-subject control was also blindly selected through automated random selection from the remaining phobic categories. For the final decoder to be used in neuro-reinforcement, the same procedure was performed but trained using all 6 runs of data.

Specific phobia treatment

Pre- and Post-treatment assessments

Each participant completed a pre-treatment and post-treatment fMRI session (Fig. 2). During the pre-treatment and post-treatment sessions, participants completed a fear test as well as an affective stroop task while their BOLD activity was recorded.

Fear test. To assess physiological, neural, and behavioral responses to phobic images, participants completed a task in which they rated how fearful they found images from select

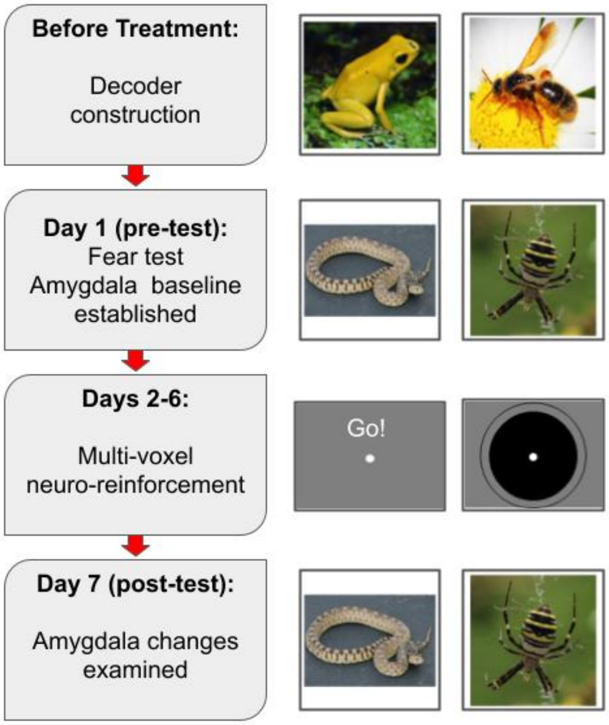

Figure 2. Study design. Timeline detailing patient activities during each day’s fMRI session with sample stimuli from each day. Before beginning the treatment program patients undergo a decoder construction session where they view non-phobic images to enable hyperalignment with healthy control subjects. On day 1 of treatment, patients complete a pre-test in which phobic (and non-phobic) images are rated for fearfulness. Over the next 5 days, patients complete their assigned number of multi-voxel neuro-reinforcement sessions (1, 3, or 5 days). On day 7, patients complete the same task as a post-test to assess changes in amygdala and SCR response to treated and untreated phobias.

categories following the previous proof-of-concept study (Taschereau-Dumouchel et al., 2018). During each trial, a fixation cross was presented for 3-7 seconds, followed by a static image for 6 seconds. After the static image, a blank screen was displayed for 4-12 seconds followed by a prompt to enter how fearful they found the image on a 7-point scale. Images displayed either belonged to the target phobia, control phobia, neutral animal, or neutral object categories. Neutral animals and objects were randomly selected based on categories for which the patient reported no levels of fear during their diagnosis interview. Participants completed two runs of 15 images each with a self-paced break in between runs. Within each run, patients viewed 5 target phobia images, 5 control phobia images, and 2-3 neutral animal/object images, counterbalanced across runs. The first image of each run was a neutral object, always immediately followed by either a target phobia or control phobia image, counterbalanced across runs. The remaining images within a run were randomly selected from the remaining images.

Affective Stroop. In order to assess patients' reflexive attentional responses to phobic stimuli, patients also completed an affective stroop task. In this task, patients started with a 1 second red fixation cross and then briefly (300 ms) saw an image from either a phobic or neutral control category. As soon as the image appeared, patients were instructed to as quickly and accurately as they could make a size judgment about whether the presented animal could fit in their hand (i.e. is it the size of your hand or smaller?). Patients pressed one of two buttons with their index and middle finger to indicate yes or no. Response-key mappings were counterbalanced across participants. There was a 1.2 second response period (indicated by a blue fixation cross) following stimulus offset for participants to enter their response followed by a fixed 1 second inter-trial interval. Stimuli were selected from 7 animal categories: target phobia, control phobia, and 5 neutral animal categories.   Similar to the fear test, neutral animal categories were selected from categories for which patients reported no fear during their diagnosis interview. The task consisted of 210 randomly distributed trials split over 2 fMRI runs with a self-paced break in between runs.

Multi-voxel neuro-reinforcement

In a number of additional fMRI sessions (based on dosage grouping), patients underwent multi-voxel neuro-reinforcement (Fig. 2). Using multi-voxel neuro-reinforcement, successful activation of the phobic image category was paired with reward. While participants laid in the fMRI scanner instructed to do “whatever they can” to get the best feedback, a neuro-reinforcement method (Taschereau-Dumouchel et al., 2018) was used to reward a nonconsciously represented phobic image category (e.g., spider). Feedback during these training sessions was based on real-time output of the decoder constructed for the individual corresponding to the specific phobia selected for treatment.

Each neuro-reinforcement run began with an extended rest period of 50 seconds while scanner image reconstruction processing caught up to real time. Then, an additional rest period of 10 seconds was collected to determine baseline BOLD activity levels followed by 16 trials of neuro-reinforcement. Each trial began with 6 seconds of rest, followed by 6 seconds of “induction” where patients modulated their brain activity in an attempt to receive high feedback. Following induction, real-time decoder output was calculated during a 4 second period and the decoder output was then displayed as a green disc for 2 seconds. The size of the disc directly corresponded to the likelihood estimate of the decoder such that a likelihood of 100% was associated with a maximum disc size (indicated by a visual boundary) and a likelihood of 0% associated with no disc displayed. The size of the disc also determined the amount of reward the patient received at the end of each run, with their average feedback score determining the percentage of that run’s total bonus received. For example, a run finished with an average score of 60% resulted in 60% of the potential $6.00 bonus being received (i.e. $3.60 bonus received). To further motivate patients, an additional bonus was also given when participants were able to generate a feedback score of 70% or more for 3 trials in a row. Patients were given an additional $2.00 per high-score streak bonus which was visually indicated by the feedback disc turning blue with a written message alerting them to their high-score streak.

Amygdala Response Analysis

Based on previous work (Taschereau-Dumouchel et al., 2018), my primary hypothesis was that multi-voxel neuro-reinforcement would result in selectively reduced amygdala response to the target phobia of treatment. To test this hypothesis, fMRI data from the fear test task were processed in the following manner.

FMRI task runs were distortion corrected using FSL’s topup (Andersson et al., 2003; Smith et al., 2004) according to spin echo field map sequences collected in opposite phase-encoding directions. Due to technical issues with spin echo field map collection, 5 participants were excluded from distortion correction. Anatomical T1 images were brain extracted using bet (Smith, 2002). Then, preprocessing and ICA-decomposition were performed using FSL’s melodic and FEAT (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). During preprocessing, fMRI data were motion corrected using mcflirt (Jenkinson et al., 2002), brain extracted using bet (Smith, 2002), spatially smoothed with a Gaussian kernel of FWHM 4.0mm, intensity normalized, and highpass filtered with a gaussian-weighted least-squares straight line fitting with sigma=50.0s. Images were then registered to the standard MNI space using FLIRT and then refined using nonlinear registration with FNIRT (Jenkinson et al., 2002; Jenkinson & Smith, 2001). Registration of multi-band images were improved by using a high-contrast single-band reference image collected at the start of each functional run as an initial reference image for registration.

ICA components were then manually investigated with components resulting from movement or other sources of noise removed. To further account for movement in this clinical cohort, data were processed with the Artifact Detection Tools (ART, https://www.nitrc.org/projects/artifact_detect) toolbox to generate motion regressors and identify outlier timepoints for censoring. First-level GLMs were then calculated in SPM12 with a temporal derivative to account for slice-timing differences. Regressors were fit for the onset of target phobia, control phobia, neutral animal, and neutral object images with a duration of 0 seconds to model the event-related response. Following previous work (Taschereau-Dumouchel et al., 2018), only the first 2 trials within each run were analyzed for target phobia and control phobia images.

Bilateral amygdala masks were generated from the automatic Freesurfer segmentation of the T1 image and transformed into the patient’s native functional space. Average parameter estimates were extracted from the Amygdala using marsbar (Brett et al., 2002). Average parameter estimates were then corrected to baseline by subtracting the average amygdala response to the neutral animal from the target phobia and control phobia, within runs. Baseline-corrected phobia responses were then averaged across runs for pre-treatment and post-treatment sessions. Amygdala responses were tested with a 2 (condition: target phobia/control phobia) x 2 (time: pre-treatment/post-treatment) repeated-measures ANOVA using JASP software (JASP Team 2022). As my specific hypothesis was that there would be a significant reduction in amygdala response for the target phobia category post-treatment compared to the control phobia, planned t-tests were performed on pre- and post-treatment reaction times for the target phobia and control phobia.

Behavioral Analyses

Fear test. For the fear test task, fear ratings were extracted for the following categories using custom scripts in python: target phobia, control phobia, neutral animal, and neutral object. To test if there was a specific reduction in self-reported fear following treatment, planned t-tests were performed on pre- and post-treatment ratings for target phobia and control phobia. One patient was excluded from this analysis due to not properly completing the task, resulting in 17 patients analyzed.

Affective stroop. For the affective stroop task, response times were extracted for target phobia, control phobia, and neutral animal stimuli using custom scripts in MATLAB (Mathworks Inc., Natick, MA). Response times were tested with a 2 (condition: target phobia/control phobia) x 2 (time: pre-treatment/post-treatment) repeated-measures ANOVA using JASP software (JASP Team 2022). As my specific hypothesis was that there would be a significant reduction in response time for the target phobia category post-treatment compared to the control phobia, planned t-tests were performed on pre- and post-treatment reaction times for the target phobia and control phobia. Also, to verify phobic images were modulating attention as intended, an additional t-test was performed on reaction times to phobic images (grouping target and control) and neutral animal images pre-treatment. Technical issues occurred for 2 patients during data collection, resulting in 16 patients analyzed for this task.

Results

Amygdala Response

Before neuro-reinforcement, there was a significant response in the amygdala for both the target phobia (*t*(17)=2.20, *p*=0.042) and control phobia (*t*(17)=2.27, *p*=0.037) compared to neutral animals as confirmed by one-sample t-tests performed on the baselined parameter estimates. This indicates successful capturing of threat responding in the amygdala for phobic images.

Following neuro-reinforcement, there was a main effect of time shown by a 2 (condition) x 2 (time) repeated-measures ANOVA (*F*(1,17)=4.56, *p*=0.048). This result indicates that there was a generalized reduction in amygdala response to phobic images (as compared to neutral animal images) following neuro-reinforcement. Amygdala responding to both phobias was reduced to levels comparable to neutral animals as confirmed by separate one-sample t-tests on the baselined parameter estimates for the target (*t*(17)=-0.53, *p*=0.60) and control (*t*(17)=-0.45, *p*=0.66) phobias. However, my primary hypothesis was that neuro-reinforcement would selectively decrease amygdala responding for the target phobia alone. Although the interaction between condition and time was not significant (*F*(1,17)=0.276, *p*=0.606), I performed planned t-tests for the target and control phobia conditions to examine target-specific engagement

After neuro-reinforcement, there was a trend of decreasing amygdala response to the target phobia (*t*(17)=1.80, *p*=0.09) but not for the control phobia (*t*(17)=1.58, *p*=0.13, Fig. 3A). These

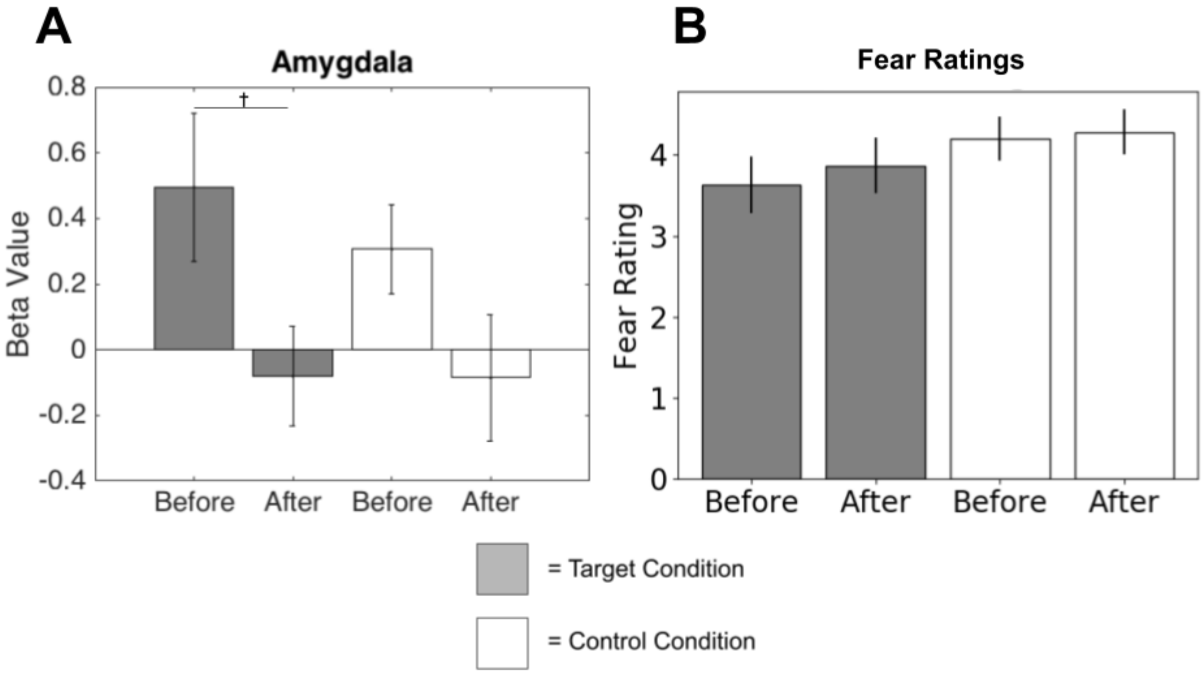

Figure 3. Amygdala response and self-reported fear following neuro-reinforcement. (A) Amygdala responses to phobic stimuli (baseline-corrected to response to neutral animal stimuli) before and after neuro-reinforcement. There was trending statistical significance for decreased amygdala response to the target phobia following neuro-reinforcement. (B) Self-reported fear ratings to phobic stimuli following neuro-reinforcement. There were no changes in self-reported fear for either the target or control phobia following neuro-reinforcement. Error bars represent standard error from mean. † $p<0.10$

findings broadly support my hypothesis that amygdala activation would be selectively reduced for the target phobia following neuro-reinforcement. Though only reaching trending significance, these results corroborate previous neuro-reinforcement findings of reduced amygdala activation to a feared animal category following neuro-reinforcement (Taschereau-Dumouchel et al., 2018). Reduction of amygdala activation following neuro-reinforcement indicates that physiological threat response to the target phobia is reduced by neuro-reinforcement.

Although the control phobia did not reach trending significance on its own, amygdala response to the control phobia was more reduced in the current study compared to the original proof-of-concept study (Taschereau-Dumouchel et al., 2018). In order to test whether the more generalized reduction in amygdala responding in the current study was related to decoder performance during decoder construction (mean AUC=0.63(0.032)), I performed an exploratory follow-up analysis. The same contrasts were performed on a subset of participants that demonstrated AUCs of 0.60 or higher during decoder construction cross validation (15 participants). This subset of participants with the highest decoding performance had greater specificity in amygdala response reduction following neuro-reinforcement. Following neuro-reinforcement, reduction of the amygdala response to the target phobia increased (*t*(14)=2.13, *p*=0.051) while it decreased for the control phobia (*t*(14)=1.09, *p*=0.30) in this subset of participants. This finding indicates decoding performance may need to reach a sufficient threshold to achieve target-specific reduction of threat responding following neuro-reinforcement.

Self-reported Fear

After neuro-reinforcement there was no significant change in self-reported fear levels in response to either the target phobia (*t*(16)=-1.52, *p*=0.15) or the control phobia (*t*(16)=-0.56, *p*=0.58, Fig. 3B). These findings match previous findings that self-reported fear levels are not modulated by neuro-reinforcement (Taschereau-Dumouchel et al., 2018). This could be due to implicit neuro-reinforcement being more effective for automatic physiological responses to threat compared to the subjective experience of fear itself.

Affective Stroop

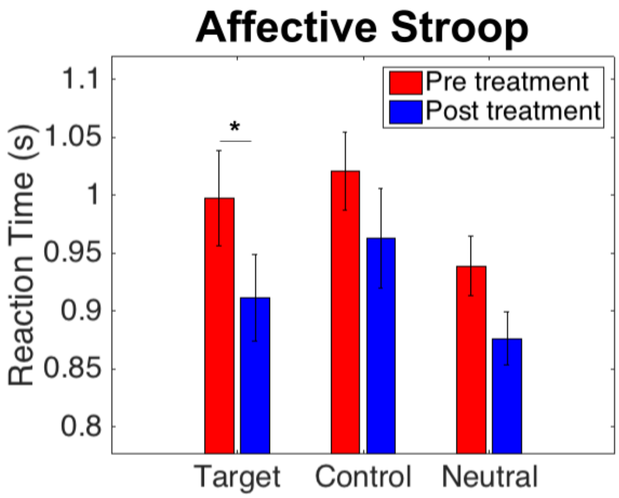

Figure 4. Reaction times in affective stroop task following neuro-reinforcement. Participants made size judgments about briefly viewed phobic and neutral animal stimuli. After neuro-reinforcement, patients were significantly faster at responding to the target phobic stimulus but not the control phobic stimulus. This indicates attention was less captured by the target phobia following neuro-reinforcement. Error bars represent standard error from mean. * p<0.05

Before treatment with neuro-reinforcement, response times for phobic stimuli were significantly slower compared to responses to neutral stimuli (*t*(15)=2.62, *p*=0.019). Slower response times for phobic stimuli indicate that attention is successfully captured by phobic stimuli in this task. Following neuro-reinforcement, there was a significant effect of time similar to that observed in the amygdala during the fear test (*F*(1,15)=5.644, *p*=0.031). This result indicates that attention was less captured by phobia following neuro-reinforcement. Although the interaction of condition and time was not significant (*F*(1,15)=1.49, *p*=0.24), I performed planned t-tests to explore the main hypothesis that reaction times would be selectively decreased for the target phobia. There was a significant decrease in reaction time to the phobic target compared to pre-treatment (*t*(15)=2.50, *p*=0.025) but not for the control phobia (*t*(15)=1.92, *p*=0.074, Fig. 4). Selectively decreased reaction times for the target phobia indicate that attention is captured less by the target phobia following neuro-reinforcement. However, the magnitude of this decreased reaction time was not directly predicted by the magnitude of decreased amygdala response during the fear test, as assessed with Spearman’s rank correlation (*r*(12)=0.19, *p*=0.51).

Discussion

In a double-blind placebo-controlled clinical trial, I investigated whether multi-voxel neuro-reinforcement could nonconsciously intervene on specific phobia. I found evidence of specific reduction in amygdala reactivity to the target phobia supporting previous findings (Taschereau-Dumouchel et al., 2018) as well as significantly reduced attentional capture by the target phobia following neuro-reinforcement. However, it should be noted that this trending effect for amygdala response reduction did not reach true significance. This can likely be attributed to a lack of statistical power. In the future, a larger trial should be conducted to see if this effect replicates with sufficient power. Additionally, there was a marked reduction in amygdala response to the control phobia following multi-voxel neuro-reinforcement.

This generalized reduction in amygdala response may be due to similarities between phobic categories for patients or due to limitations in the category classification performance of the decoder. In multiple cases, target and control phobias appeared superficially related despite belonging to distinct animal species (e.g. spiders/cockroaches/beetles, chickens/peacocks). Although the goal was to demonstrate unchanged control phobia responding following neuro-reinforcement, generalized reduction in threat responding following neuro-reinforcement may still be therapeutically meaningful. Ultimately, the goal of treating people with multiple phobias would be to have threat responding to all phobias reduced. If non-specific reduction in threat responding is related to conceptual or perceptual overlap between target and control phobias for patients, it would only be more cost effective from a therapeutic angle to be able to reduce threat responding to both phobias simultaneously. Future studies should employ measures of representational similarity between target and control phobia multi-voxel patterns in order to assess the relationship between target and control similarity and effect specificity following neuro-reinforcement.   It should be noted that this reduction in amygdala response was specific to phobic images as amygdala response to neutral non-feared animals was used as a baseline in these comparisons. If amygdala response was similarly reduced for both feared and non-feared images there would have been no observed difference in amygdala responding following neuro-reinforcement.

Despite evidence of decreased amygdala response to phobia following neuro-reinforcement, I did not observe any decrease in self-reported fear levels. This also matches previous findings (Taschereau-Dumouchel et al., 2018) indicating non-conscious neuro-reinforcement does not seem to alter subjective fear experience. This discordance is consistent with a higher-order theory of emotion in which subjective mental experience operates via different mechanisms than physiological threat responses (Taschereau-Dumouchel et al., 2022). It may be that the kind of nonconscious exposure employed in this neuro-reinforcement design is more effective at targeting physiological threat responses rather than subjective fear experiences. While an effective treatment would ultimately aim to reduce subjective fear experiences when confronting phobic stimuli, neuro-reinforcement could represent an important first step in reducing subjective discomfort during traditional exposure treatments. Despite similar levels of self-reported fear, patients may be more willing to engage with phobic stimuli following neuro-reinforcement or be less behaviorally averse.

This notion is supported by the results from the affective stroop task. Following neuro-reinforcement, reaction times were significantly decreased specifically for the target phobia. In addition to providing further support for specific target engagement by neuro-reinforcement, this result suggests that patients may be less reflexively avoidant to their phobia following neuro-reinforcement. If this is the case, patients may find traditional behavioral exposure treatments less aversive following neuro-reinforcement leading to lower rates of attrition.

To test this hypothesis, future studies should complement neuro-reinforcement with a behavioral-approach task to investigate whether physiological symptoms are decreased when approaching the target phobia following neuro-reinforcement. If patients are more willing to approach the feared animal following neuro-reinforcement, then neuro-reinforcement may be a good complementary treatment alongside traditional exposure for ensuring the most comfortable treatment regimen possible.

It is also worth noting this procedure was able to be conducted in a double-blind fashion. This provides the utmost level of rigor for testing the efficacy of neuro-reinforcement as a clinical intervention. Other forms of neurofeedback are not so amenable to double-blind testing due to technical constraints and psychological interventions are not always tested at such a rigorous level.

In summary, this study represents the first clinical trial of multi-voxel neuro-reinforcement for nonconscious brain-based psychotherapy. This procedure demonstrated the ability to lessen physiological, reflexive responses to specific phobia through reduced amygdala activation as well as less attentional capture by phobic stimuli. These findings provide a promising foundation to attempt larger-scale replications in clinical cohorts. Through advents in virtual reality, these responses can also be investigated in future studies using more realistic and immersive stimuli. This nonconscious procedure produces minimal discomfort in patients with very low rates of attrition. Consequently, neuro-reinforcement may serve to complement current conventional psychotherapy approaches while providing a more tolerable experience for patients seeking treatment.

Chapter 3. Classifying the subjective awareness of threat from multi-voxel patterns.

Introduction

Through decades of effort and billions of dollars in funding and societal cost, the treatment of mental health disorders has remained difficult. This difficulty persists, in part, due to models that have focused on physiological or behavioral models to explain mental health symptoms. The temptation to do so is understandable given the ease of adapting these dimensions to objective operationalizations and translational models in primates, mice, and other non-human organisms. However, these endeavors have left much to be desired as mental health continues to grow as a public health crisis (Insel, 2019). Recently, a push has begun for mental health models to focus specifically on the subjective experience of the patient (Taschereau-Dumouchel et al., 2022; Whiteley, 2021). After all, the patient's subjective experience is the ultimate metric by which they will judge their treatment's success.

Even the most classic paradigms, such as the Pavlovian threat conditioning paradigm are in need of a fresh perspective when subjective experience is properly taken into account. Commonly referred to as the "fear conditioning" paradigm, this alternative name highlights just how for granted subjective experience is taken in the modeling of behavior. Most of what has been studied using this paradigm has not explicitly been fear but a variety of reflexive and physiological threat responses and behaviors (LeDoux & Hofmann, 2018). While the domains share obvious correlations, discordance between the two is critical to consider given that it may be in these various discordances that mental health disorders arise (Taschereau-Dumouchel et al., 2022).

Additionally, in the exceptionally complex endeavor of explaining behavior and subjective experience from neuroimaging brain patterns, it is essential to understand what processes are actually being studied. For example, the brain region perhaps most famously associated with the Pavlovian threat conditioning paradigm, the amygdala, has recently been shown to be more responsible for generating physiological threat responses rather than subjective fear responses (Taschereau-Dumouchel et al., 2019). Furthermore, through neuro-reinforcement, it has been shown that amygdala responses to feared stimuli can be selectively decreased while subjective fear of the feared stimulus remains unchanged (Taschereau-Dumouchel et al., 2018). While no singular study can control for every confound, it is important to be as conceptually precise as possible when investigating neural mechanisms where the temptation to impose high-level concepts on low-level processes remains high.

Here, I investigate the brain patterns associated with subjective threat awareness using machine learning techniques (Ashar et al., 2017; Chang et al., 2015; Eisenbarth et al., 2016; Taschereau-Dumouchel et al., 2019; Zhou et al., 2021). By leveraging machine-learning classifiers, threat responses can be classified based on whole-brain multi-voxel activation patterns according to participants' self-reported threat contingency awareness ratings. In addition to utilizing the full complex multivariate nature of the human brain's fMRI response, this analysis has the added benefit of utilizing threat awareness self reports taken after the task run. This results essentially in a no report paradigm (Tsuchiya et al., 2015) where brain responses from within the task are not confounded by participants needing to think about reporting their threat awareness during the actual task. By investigating a population that has been selected to represent a wide range of anxiety and neuroticism symptoms, I am able to explore what symptom dimensions correlate with learned fear expression. I also explore how results generalize to another large independent dataset using the same paradigm. By exploring brain patterns associated with subjective awareness of threat, I hope to elucidate which brain regions are important for threat awareness as opposed to simple reflexive threat responses to which other analyses bring focus.

Methods

Initial Dataset

Participants

As part of a larger study tracking development of psychological symptoms during emerging adulthood/late adolescence, 279 participants (183 females, mean age=19.65(.53)) were recruited at Northwestern University and University of California, Los Angeles. Participants were sampled from a larger recruited population of 2461 participants to select participants representing a broad distribution of self-reported reward sensitivity and threat neuroticism traits. Of these 279 participants, 273 went through Structured Clinical Interview for DSM-5 interviews. Of these 273 interviews, 64 participants met criteria for a current anxiety disorder but no depressive disorder, 19 met criteria for current comorbid depressive and anxiety disorders, and 4 met the criteria for a current depressive disorder but no anxiety disorder. Participants were excluded from all fMRI analysis if they demonstrated excessive motion (defined as >10% outlier scans) in any of the 3 task phases (acquisition, extinction, and extinction recall). After exclusions for motion and technical issues during data collection, 157 participants had usable data across all three task phases.

Threat Conditioning Task

Participants completed a three-phase (acquisition, extinction, and extinction recall) Pavlovian threat conditioning task while undergoing an fMRI scan (Fig. 5, Milad et al., 2009; Young et al., 2021). Extinction recall took place at least 48 hours after the initial acquisition and extinction session in a separate fMRI session (mean days apart=2.76(2.48)). Conditioned stimuli (CS) were colored lamp lights (blue, red, and yellow) presented visually within a broader context image (office or conference room, depending on task phase). The unconditioned stimulus (US) was an electric shock titrated for each participant as to be annoying but not painful.

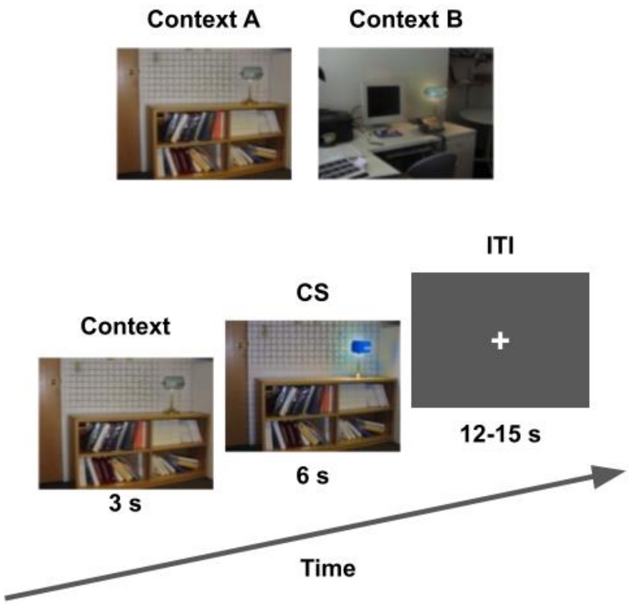

Figure 5. Task design. (Top) Context images used in acquisition, extinction, and extinction recall phases. Acquisition occurs in context A while extinction and extinction recall occur in context B. Context identities are counterbalanced across participants. (Bottom) Trial design for all task phases. Trials begin with a 3 second context presentation, followed by 6 seconds of the conditioned stimulus (CS) presented within the context image, followed by a 12-15 second inter-trial interval (ITI). During threat acquisition, unconditioned stimulus (US, electric shock) was delivered at the start of the ITI following CS offset.

During acquisition, participants were presented with 2 CS+ stimuli and 1 CS- stimulus (colored lamps counterbalanced across participants). Participants viewed 8 trials of each CS+ and 16 trials of the CS-. Each trial began with 3 seconds of the context image with no CS, then the CS was presented within the context image for 6 seconds followed by a variable inter-trial interval of 12-15 seconds. For trials where CS was paired with US, the removal of the context and CS+ image from the screen coincided with electric shock delivery. CS was reinforced at a rate of 62.5%. Trials were presented in a pseudorandom order such that shocks were delivered at the same timepoints across all participants while appearing random to the participants.

Extinction and extinction recall followed the same trial structure but with no US delivery. During extinction, participants viewed 16 trials of the extinguished CS+ (CS+E) and 16 CS- trials. During extinction recall, participants viewed 8 trials of CS+E, 8 trials of the unextinguished CS+

(CS+U), and 16 trials of the CS-. Context images were consistent within each phase such that each phase took place in a distinct context (office or conference room). Acquisition was always performed in one context while extinction and extinction recall were performed in the other, in an A-B-B fashion. The context identity (A/B) of office or conference room was counterbalanced across participants.

At the end of each phase, threat “contingency awareness” ratings were collected for each CS in order to explicitly assess learned CS-US associations. Participants were presented with an image of the CS and asked to rate the “likelihood of receiving a shock if you saw this image again” on a 3-point scale spanning “low”, “moderate”, and “high”.

MRI preprocessing

Structural T1 images were intensity normalized and the brain extracted using optiBET (Lutkenhoff et al., 2014). The brain extracted T1 image was then segmented into White Matter, Gray Matter, and Cerebrospinal Fluid using FAST (Zhang et al., 2001).

Functional images were preprocessed separately for each task phase. Motion outliers were calculated for functional images using *fsl_motion_outliers* before any preprocessing took place. Then, functional images were motion corrected, smoothed with a 4 mm FWHM kernel, and nonlinearly registered into the MNI152NLin6Asym standard space using 12 degrees of freedom with FSL’s FEAT (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). Motion components were then automatically detected and removed using ICA-Aroma (Pruim et al., 2015). ICA-aroma was used as it has been found to be one of the most effective methods of removing motion from fMRI data without discarding large amounts of data (Parkes et al., 2018). Following ICA-aroma, single-trial GLM estimates were computed using a Least-Squares Separate (LSS) approach (Turner et al., 2012) in pyMVPA/python (Hanke, Halchenko, Sederberg, Hanson, et al., 2009).

Concretely, a separate GLM was calculated for each experimental trial wherein the current trial was modeled with a moment-specific regressor while all other events were grouped under a singular "rest" regressor. During threat acquisition, events were modeled for each context presentation (office/conference room), CS presentation (2 CS+/1 CS-), and US presentation (shock). For threat extinction and extinction recall, task regressors included each context presentation and CS presentation. Each GLM estimation utilized a SPM-style hemodynamic response function model including a temporal derivative to account for slice acquisition timing differences and also included a 128s high-pass filter using a cosine drift model with nuisance regressors comprising motion outliers and 6 head motion parameters. LSS GLMs were calculated in participants' native space and then transformed into MNI space using the transformation parameters from the FSL registration.

MVPA analysis

Voxel-level MNI-space parameter estimates from single trial GLMs were averaged together for each stimulus type (CS+E, CS+U, and CS-) within runs and within participant. This resulted in 5 training exemplars for each participant, 3 from acquisition (CS+E, CS+U, and CS-) and 2 from extinction (CS+E and CS-). Only data from acquisition and extinction were used to achieve a balance between class labels. Class labels were assigned based on the threat contingency awareness ratings participants provided for each stimulus following each run. Initially collected on a 3-point scale ("high", "moderate", and "low"), ratings were binarized in order to create a more even distribution of classes as most participants did not utilize the full 3-point scale. High and moderate ratings were grouped together as one class indicating the stimulus was deemed a "threat" while low ratings comprised the other class indicating the stimulus posed "no threat". From 157 participants this resulted in 785 training exemplars of which 358 exemplars represented "threat" while 427 exemplars represented "no threat". For each decoding analysis (whole brain or ROI), features inputted to the classifier represent voxel-level GLM parameter estimates that are anatomically aligned across participants in MNI space.

Whole-brain threat awareness decoding

To assess threat awareness decodability at the whole-brain level, whole-brain parameter estimates in MNI space from the LSS GLM analysis were masked to the gray matter using a mask generated from a FAST segmentation of the FSL standard extracted brain. Whole-brain parameter estimates were then subjected to the following cross-validation and permutation testing.

Cross-validation testing

In order to assess whether threat awareness could be successfully decoded from the data, a 20-times repeated split-half cross validation scheme was used. Repeated split-half cross-validation schemes have been shown to have the most power despite resulting in a slightly lower classification accuracy (Valente et al., 2021). For each repeat, the data was randomly split in half while preserving participant grouping so that there were no influences of one participant on both the training and testing data to prevent overfitting. That is, the totality of each participant's data was either part of the training set or the testing set, but never both. Due to the uneven number of participants, this resulted in a split of 390 trials to 395 trials. For each half split, each half was used once as a training set and once as a testing set. Balanced accuracy scores as well as the area under the receiver operating characteristic curve (AUC) were calculated for each training and testing pair and then averaged first over half splits and then over the 20 repeats to assess cross-validation performance.

For each training set, class sizes were balanced, if necessary, by random undersampling of the overpopulated class. Training data was then submitted to a C-Support Vector Classifier with the default regularization parameter of C=1.0 and radial basis function kernel using the sklearn package in python (Pedregosa et al., 2011). Standardization parameters were calculated on the training set to perform feature-wise standardization by removing the mean and scaling to unit variance. Both training and testing data were transformed according to these standardization parameters, but importantly no overfitting resulted as the parameters were calculated using the training data only preventing any leaking of the testing data into the training data. Predictions and estimates of the classifier decision function were collected for each testing set to calculate balanced accuracy and AUC respectively.

Permutation testing

To determine the significance of the cross-validated classifier performance, non-parametric permutation testing was performed. To build a data-driven null distribution, 1000 permutation tests were performed over the full 20-repeat split-half cross-validation procedure. For each permutation, class labels were shuffled once and then performance of the permuted classifier was assessed over the full 20-repeat split-half cross validation. As with the real data, performance of the permutation classifier was averaged across half splits and then averaged across the 20 repeats to obtain a singular balanced accuracy and AUC estimate for each permutation. These null distributions were then used to set the critical value for significance. That is, for the real classifier’s performance to be considered statistically significant, it needed to be larger than 950 of the permutation estimates.

Within-region threat awareness decoding

To investigate which brain regions individually were sufficient to decode threat awareness, the same cross-validation and permutation testing routines were performed in each parcel of a 200 parcel parcellation derived from resting state functional connectivity patterns (Schaefer et al., 2018) as well as Amygdala and Hippocampus segmentations from the Harvard-Oxford Atlas

(Desikan et al., 2006) for a total of 204 brain regions. For each ROI analysis, features inputted to the classifier were the voxel-level MNI-space parameter estimates falling within the ROI mask. In order to correct for multiple comparisons across the 204 regions, a non-parametric p-value was estimated for each brain region. A gaussian kernel was fitted to the null distribution calculated from a region's permutation tests and the probability density function evaluated at the point of that region's true classifier performance. Then, non-parametric p-values across the 204 brain regions were FDR-corrected to control the false discovery rate with a q-value threshold of 0.001.

Generalization Dataset

Participants

Participants were recruited from two clinics at University of California San Diego Primary Care Clinics (N=101) and University of California Los Angeles Family Health Center (N=124). Participants were included if they had a score greater than or equal to 10 on the Patient Health Questionnaire-9 (PHQ-9, Kroenke et al., 2001) or greater than or equal to 8 on the Overall Anxiety and Impairment Scale (OASIS, Campbell-Sills et al., 2009). Participants were excluded for moderate to severe alcohol or cannabis use or any other mild substance use disorder. Additional exclusion criteria were a diagnosis of Bipolar or Psychotic disorders, moderate to severe traumatic brain injury, active suicidal ideation, or MRI contraindications. Participants were excluded from all fMRI analysis if they demonstrated excessive motion (defined as >10% outlier scans) in any of the 2 task phases (acquisition, extinction). After exclusions for motion and technical issues during data collection, 183 participants had usable data across both acquisition and extinction phases.

Task

Participants completed a slightly modified version of the same Pavlovian threat conditioning paradigm. Participants completed a singular fMRI session with acquisition and extinction phases only. As there was no extinction recall phase, participants only viewed one type of CS+ during the threat acquisition phase. All other task parameters and procedures are the same as the initial dataset described here.

MRI processing

Structural T1 images were intensity normalized and the brain extracted using optiBET (Lutkenhoff et al., 2014). Functional runs for the threat acquisition and threat extinction task phases were processed separately. Complete first-level analyses were processed using FSL's FEAT (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). FMRI data were motion corrected using MCFLIRT (Jenkinson et al., 2002), slice-time corrected with Fourier-space time-series phase-shifting, and brain extracted using BET (Smith, 2002). Images were spatially smoothed with a FWHM 4mm Gaussian kernel, grand-mean intensity normalized, and highpass filtered with Gaussian-weighted least-squares straight line fitting, sigma=50.0s.

Functional runs were registered to the standard space using FLIRT (Jenkinson et al., 2002; Jenkinson & Smith, 2001) and further refined with FNIRT nonlinear registration. First-level GLMs were calculated using FILM with local autocorrelation correction (Woolrich et al., 2001). Regressors were included for the onset of CS+, CS-, and context image presentation. Nuisance regressors included the 6 head motion parameters from motion correction and timepoint censoring for any motion outlier TRs.

First-level GLM estimates for the CS+ and CS- during threat acquisition and extinction were standardized within participant and paired with the behavioral contingency ratings for each run yielding 4 samples per participant for MVPA generalization analysis.

MVPA generalization analysis

To test how the machine-learning classifier developed here generalizes to independent datasets, a generalization analysis was performed by training a classifier on the initial dataset and testing it on the generalization dataset. This followed the same procedure as the MVPA analyses performed during the cross-validation procedure described above for the initial dataset but with the classifier being trained on the totality of the initial dataset and tested on the totality of the generalization dataset. Classifier generalization performance was assessed using AUC scores and balanced accuracy scores. Statistical significance of generalization performance was assessed using a permutation test in which class labels were shuffled for both the initial and generalization datasets with 1000 permutations.

Classifier evidence for threat awareness during extinction recall

To investigate how classifier evidence for threat awareness during extinction recall related to participant symptomatology in the initial dataset, a decoder was applied to average extinction recall trials. To prevent overfitting and maximize training data, the decoder was trained on the generalization dataset (in which there was no extinction recall phase) and applied to the initial dataset. Classifier estimates for each participant in the initial dataset were calculated for the CS+U and CS+E exemplars. Classifier estimates for CS+U were subtracted from CS+E and these differences in estimates were correlated with self-report questionnaire scores using Pearson’s correlation.

Trial-by-trial threat awareness classification estimates

To generate a time-resolved picture of how threat awareness develops in the Pavlovian threat conditioning task, classifier estimates were also calculated at the individual trial level. The same decoder trained on the generalization dataset was used to prevent overfitting. This decoder was then applied to the LSS single-trial data from the initial dataset. Trial-by-trial classifier estimates were collected for each conditioned stimulus across threat acquisition, extinction, and extinction recall task runs. Significant threat awareness was assessed by one-sample t-tests Bonferonni corrected for the number of trials tested within each task run. Results were only tested for being greater than 0 to assess at what time points significant threat awareness was detected.

Results

Whole-brain threat awareness decoding

To determine whether subjective threat awareness could be decoded at the whole-brain level, a C-support vector machine classifier was trained and tested on brain responses to conditioned stimuli during threat acquisition and extinction. Whole-brain decoding demonstrated high sensitivity and specificity for the awareness of threat during cross validation (average AUC=0.84, p<0.001, average bal. accuracy=0.79, p<0.001). This finding indicates that subjective awareness of threat can indeed be robustly decoded from whole-brain activation patterns.

Generalization of whole-brain decoding

To test how generalizable this whole-brain decoding of threat awareness was, this decoder was also applied to a large independent dataset. Testing on independent data demonstrated robust generalized performance of the decoder to a new dataset (average AUC=0.79, average bal. accuracy=0.69, p<0.001).

Within-region decoding

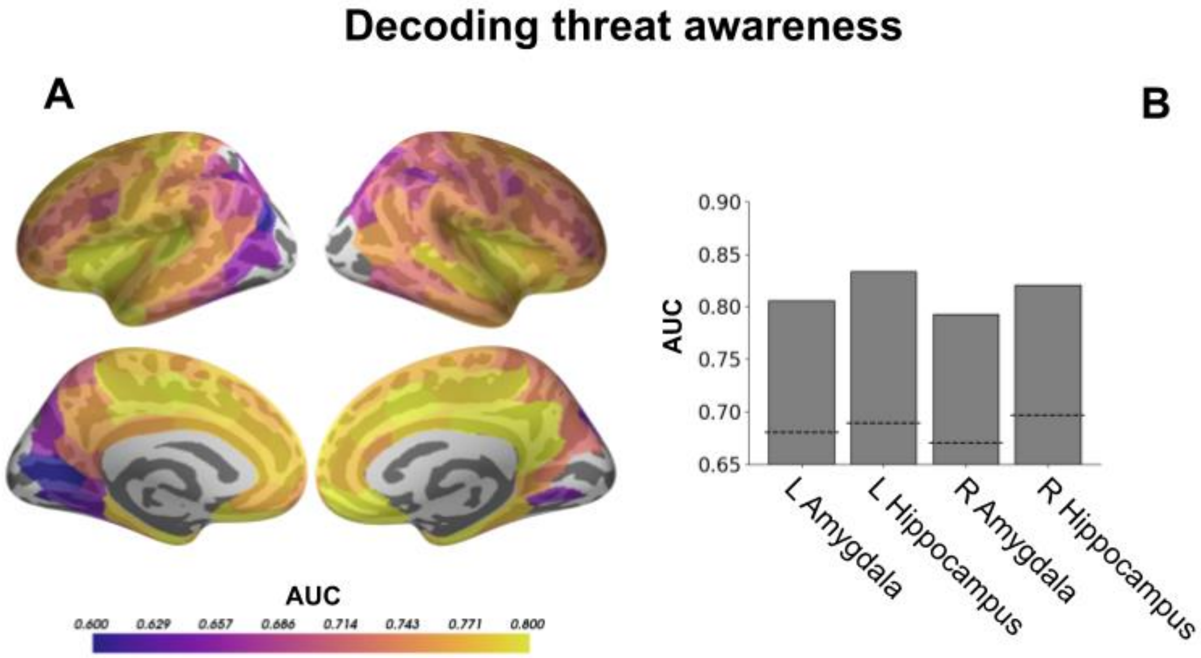

Figure 6. Within-region classification of subjective threat awareness. A C-support vector machine classifier was iteratively trained in each of 200 cortical parcellations as well as amygdala and hippocampus. Significance of within-region decoders was determined through non-parametric permutations within each region and non-parametric p-values were FDR-corrected for 204 comparisons with q=0.001. (A) Parcellations with significant subjective threat awareness classification are plotted with color corresponding to the region's average area under the receiver operating characteristic curve (AUC) over 20 repeated split-half cross validations. (B) Classifier performance in subcortical areas amygdala and hippocampus is plotted as assessed by AUC scores. Dashed lines represent a non-parametric threshold corresponding to p<0.001 uncorrected. Subjective threat awareness had a distributed signature with increased threat decodability traveling along the visual processing hierarchy. A subset of regions including posterior cingulate, retrosplenial cortex, vmPFC, inferior frontal gyrus, and OFC as well as subcortical regions amygdala and hippocampus demonstrated particularly strong within-region decoding comparable to whole-brain classification performance.

To determine which regions of the brain contribute to the decodability of subjective threat awareness, the same cross-validated classification procedure was performed iteratively over 200 cortical parcellations spanning the entirety of the cortex (Schaefer et al., 2018) as well as the subcortical regions amygdala and hippocampus. This analysis revealed a widely distributed signature for the subjective awareness of threat as threat awareness could be decoded from 188 of the 204 tested parcellations following FDR correction with q=0.001 (Fig. 6). Full results from all parcellations are reported in Appendix Table A1.

Within-region decoding results showed an hierarchical organization with early visual areas being unable to decode threat awareness. Threat awareness became more decodable at increased levels of the visual hierarchy. Threat awareness was most decodable from areas including posterior cingulate, retrosplenial cortex, vmPFC, inferior frontal gyrus, and OFC as well as subcortical regions amygdala and hippocampus (Fig. 6B). All these regions showed within-region decoding broadly comparable to the whole-brain signature with AUC's >0.8. These findings suggest threat awareness information is not present in early visual processing but rather critically relies on lower-level signals coming from areas like amygdala and hippocampus.

Threat awareness during extinction recall

As an assessment of the relationship between extinction memory and participant symptomology, I examined classifier evidence for threat awareness during extinction recall as a function of participant symptom scores. The trait of worry was significantly correlated with the level of classifier evidence for threat awareness in response to the CS+E compared to the CS+U during extinction recall (Fig. 7, *r*(155)=0.25, *p*=0.002). Participants with higher trait worry had greater classifier evidence for threat awareness in response to the CS+E compared to the

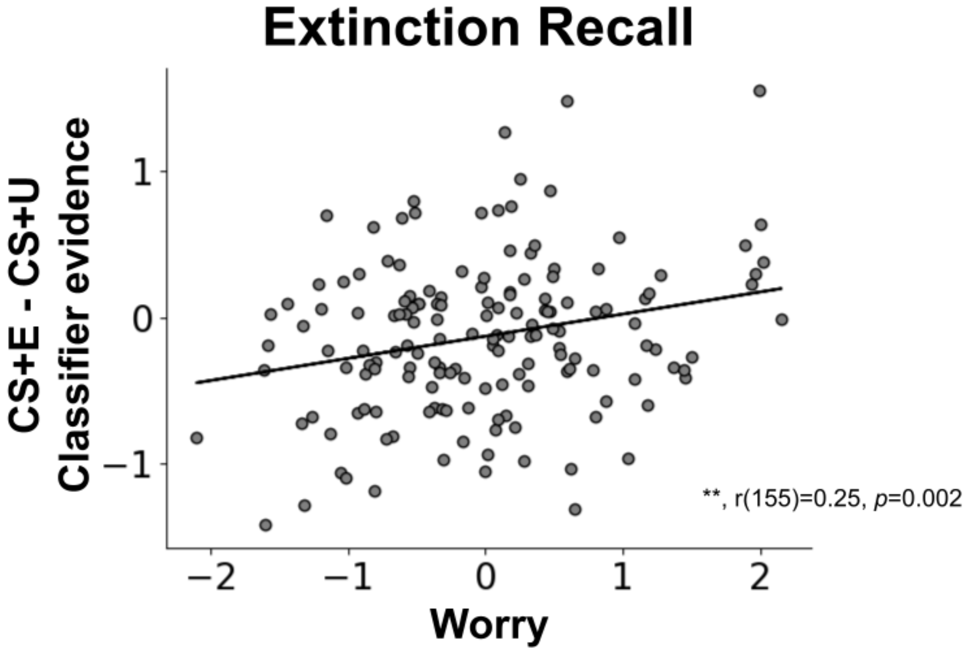

Figure 7. Classifier evidence for threat awareness during extinction memory recall is associated with trait worry. During extinction recall, predicted threat awareness for CS+E and CS+U is measured by classifier evidence from a classifier trained on acquisition and extinction trials only and tested on CS+E and CS+U trials during extinction recall. The difference in classifier evidence for threat awareness for the extinguished conditioned stimulus (CS+E) compared to the unextinguished conditioned stimulus (CS+U) is measured on the y-axis. Trait worry as assessed with the Penn State Worry Questionnaire (PSWQ37) is represented on the x-axis. Participants with higher trait worry had a greater threat signature for the CS+E compared to CS+U indicating a failure to retain extinction memory.
function of participant symptom scores.

CS+U while participants with low trait worry had greater classifier evidence for threat awareness for the CS+U compared to CS+E. That is, participants with higher trait worry demonstrated worse extinction recall as evidenced by classifier estimates of threat awareness for the extinguished conditioned stimulus being higher than the unextinguished conditioned stimulus, which is the opposite of what would be expected following successful extinction. This effect was primarily driven by higher classifier evidence for threat awareness in the response to the CS+E as worry significantly correlated with classifier evidence for threat awareness in response to the CS+E (*r*(155)=0.24, *p*=0.0024) but not CS+U (*r*(155)=-0.13, *p*=0.10). As there was no relation between trait worry and classifier evidence for threat awareness in response to CS+ vs CS- at the end of extinction (*r*(155)=-0.0088, *p*=.91), this indicates those with high trait worry have worse consolidation and retention of extinction memory rather than a failure to learn extinction in the first place. That is, worry is associated with greater threat return after 48 hours following extinction. Importantly, worry did not correlate with differences in self-reported threat contingency awareness for the CS+E and CS+U (*r*(155)=0.83, *p*=0.30). This finding highlights worry as a potential process critical in the relation between anxiety disorders and threat conditioning.

Trial-by-trial threat awareness decoder activation

In order to obtain a time-resolved measure of the emergence of subjective threat awareness in this threat conditioning paradigm, the threat awareness decoder was applied to single trial data across acquisition, extinction, and extinction recall. Classifier evidence for the subjective awareness of threat was significantly greater than chance for the last 3 trials of threat acquisition for the CS+E and for the last trial of CS+U (Fig. 8, one-sample t-tests Bonferonni corrected across time points, *p*<0.05). This finding indicates it takes at least 5 trials of partially-reinforced CS-US pairings for the brain signature of learned subjective threat to emerge. This late appearance of subjective threat awareness also supports the fact that the classifier developed here is detecting learned threat as opposed to simple threat response elicited by the US which would be evident in early trials. There were no significant activations of threat during

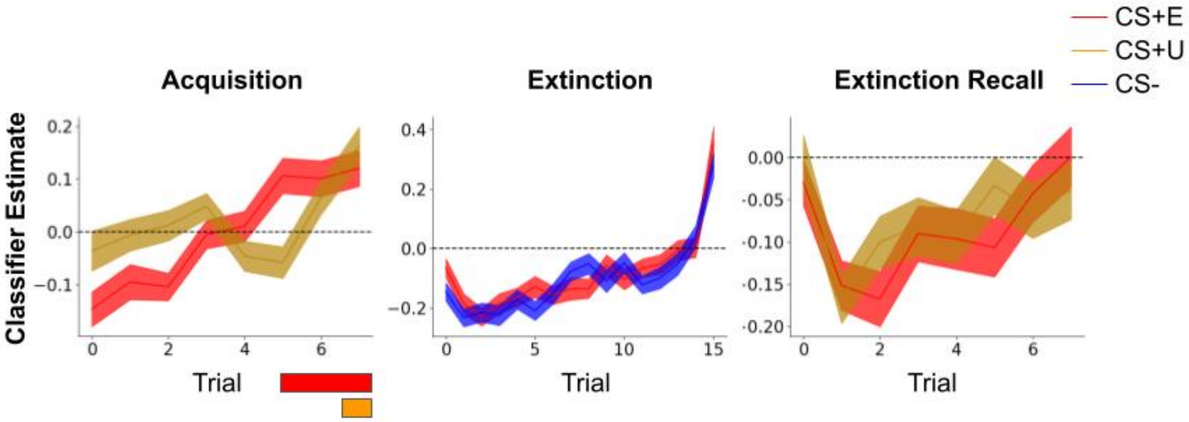

Figure 8. Trial-by-trial threat awareness decoder activation. The threat awareness decoder was applied to single-trial data in the threat acquisition, extinction, and extinction recall phases to determine at what time points subjective threat awareness emerged. Y-axis represents the classifier estimate for a given trial while the x-axis represents trials across time. Legend at top-right indicates which line colors track what conditioned stimuli. Shaded areas around line plots represent standard error of mean. Colored bars below x-axis represent timepoints with classifier evidence greater than 0 (dashed line) as determined by one-sample t-tests Bonferonni corrected across number of trials. Significant threat awareness emerged at the end of threat acquisition for both CS+ stimuli but was not observed in the extinction or extinction recall phases.

extinction and extinction recall phases at the single-trial level (Fig. 8). This could be due to the decoder lacking adequate performance at the single-trial level in these later phases in which no objective threat is present.

Discussion

For this study, the main goal was to find a whole-brain pattern for the subjective awareness of threat in a Pavlovian threat conditioning paradigm. This resulted in development of a whole-brain decoder that was able to detect subjective threat awareness developing in the final trials of threat acquisition with high sensitivity and specificity in a large cohort of over 150 participants. Results were highly generalizable to another large independent dataset of over 180 participants. This is indicative of a robust signature of threat awareness as it could be detected across a large number of participants collected across multiple geographic locations on multiple different scanner systems. Decoding results were also robust to differences in preprocessing and differences in single-trial and first-level GLM modeling.

Within-region decoding revealed a distributed pattern of threat awareness across the brain organized in a hierarchical fashion. Threat awareness was not decodable from early visual areas and then became more decodable as the visual hierarchy progressed. This revealed a sub-network of regions including posterior cingulate, retrosplenial cortex, vmPFC, inferior frontal gyrus, and OFC from which threat awareness could be decoded on par with the whole-brain signature. The inability of threat awareness to be decoded from early visual areas is supportive of a higher-order view of subjective awareness in which higher-order processes aside from early sensory areas are needed for content to enter subjective awareness (Brown et al., 2019; LeDoux & Hofmann, 2018).

I also investigated subcortical regions amygdala and hippocampus which were critical in the decoding of threat awareness with decoder performance comparable to the whole-brain level. These findings are broadly consistent with other recent investigations of whole-brain signatures of subjective fear (Taschereau-Dumouchel et al., 2019; Zhou et al., 2021) which have indicated areas around PFC, OFC, and cingulate cortex as areas important for the experience of subjective fear. Importantly, the results presented here do not have anything to say about the emotional subjective experience of fear per se. The current study examined subjective awareness of threat without respect to the potential emotional experience involved. Moreover, the current study was limited in its ability to disentangle subjective threat awareness from more implicit physiological threat responding. This could explain high threat awareness decoding in subcortical regions like amygdala and hippocampus which are found to be more predictive of physiological threat than subjective fear (Taschereau-Dumouchel et al., 2019). Presumably, physiological threat signatures and subjective awareness of threat are highly correlated in the current study.

I also examined classifier evidence for threat awareness during an extinction recall period following 48 hours after the initial threat acquisition and extinction as a function of participant symptomatology. Participants with higher trait worry had higher expression of threat awareness for the previously extinguished conditioned stimulus compared to an unextinguished conditioned stimulus. This indicates greater fear renewal for those with high trait worry. Consequently, worry should be explored in future studies as a process potentially linking anxiety disorders and aberrant threat conditioning (Craske et al., 2017; Fenster et al., 2018). As worry is characteristic of intrusive thought, it could also be related to intrusive imagery that is characteristic of fear-related disorders like post-traumatic stress disorder (Brewin et al., 2010). PTSD is similarly characterized by deficient extinction recall (Garfinkel et al., 2014). Future studies should investigate the link between intrusive thoughts, images, and extinction recall abilities to see if a foundational mechanism links these various intrusive experiences.

In summary, threat awareness is coded by a distributed brain signature that can be observed across differences in multiple scan sites and geographic locations. Future studies will be needed to disentangle subjective threat awareness from more physiological implicit threat processes. Future investigation will also be needed to differentiate between things like subjective fear and subjective awareness of threat. The additional disentanglement between things like subjective fear and awareness of threat will help to identify brain regions and processes responsible for the emotional experience underlying many fear-related disorders like anxiety and PTSD.

Chapter 4. Investigating brain networks involved in acquisition, extinction, and recall of learned threat.

Introduction

Responses to threat need to be both fast and accurate to ensure lasting survival in an environment. Direct threats need to be quickly identified and reflexively responded to when danger comes our way. However, maladaptive threat responses also need to be properly inhibited to prevent context-inappropriate reactions such as running away from someone who is seeking your help. Deficits in either of these processes can lead to behavioral outcomes resembling an anxiety disorder. Consequently, fear and anxiety disorders are often thought to be characterized by aberrations in threat processing and conditioning (Craske et al., 2017; Fenster et al., 2018). For example, anxiety disorders are associated with increased threat acquisition and impoverished threat extinction (Pittig et al., 2018). Anxiety is also thought to be related to how learned threat generalizes to new stimuli beyond the initial threat learning episode (Dunsmoor & Paz, 2015). As such, the Pavlovian threat conditioning paradigm has become a pillar in studies examining fear and threat processing due to its simplicity and utility in translational research from animal models to human participants.

Building from animal models, critical brain areas for threat conditioning have been identified such as amygdala and hippocampus (Phillips & LeDoux, 1992). Functioning in these central regions is certainly informative for anxiety disorders as increased anxiety is associated both with facilitated acquisition of threat response as well as increased response to clear threats and impoverished response to ambiguous threats (Im et al., 2017; Pittig et al., 2018). However, it is important to situate these critical nodes within the larger functional networks in which they operate. Especially at the human level, it is likely that there are multiple circuits operating in parallel in what we would typically label the “fear” response (LeDoux & Hofmann, 2018). Understanding the mechanisms of the subjective experience of fear in humans is critical to helping treat anxiety and fear-related disorders (Taschereau-Dumouchel et al., 2022). Focusing on changing behavioral or physiological outcomes has led to a dearth of effective treatments as these things can be changed without any change in the subjective distress experienced (Taschereau-Dumouchel et al., 2018).

Much of the previous human neuroimaging work using the Pavlovian threat conditioning paradigm has used simple univariate contrasts to identify threat-sensitive brain regions (Fullana et al., 2016). However, whole-brain connectivity is beginning to be used to understand the broader dynamics at play in human threat conditioning (Berg et al., 2021; Wen et al., 2021). Network analyses have become increasingly popular as a way to understand how distributed regions across the entire brain organize their activity in coordinated functions (Sporns, 2014). Distinct networks have been shown to track overgeneralization of conditioned threat in post-traumatic stress disorder (Berg et al., 2021). Extinction of conditioned threat has also been shown to modulate brain connectivity in areas associated with default mode, frontoparietal, and ventral attention networks (Wen et al., 2021). Here, I utilize group network analysis methods to investigate brain networks involved in the acquisition and extinction of conditioned threat as well as the recall of extinction memory after a 48-hour consolidation period.

Methods

Participants

Participants for this dataset come from the same sample of 279 participants described in Chapter 3. Participants were excluded from all fMRI analysis if they demonstrated excessive motion (defined as >10% outlier scans) in any of the 3 task phases (acquisition, extinction, and extinction recall). After exclusions for motion and technical issues during data collection, 223 participants were analyzed for acquisition and 208 participants were analyzed for extinction and extinction recall.

Task

Participants completed the same three-phase (acquisition, extinction, and extinction recall) Pavlovian threat conditioning task described in Chapter 3 while undergoing an fMRI scan.

MRI preprocessing

Structural T1 images were intensity normalized and the brain extracted using optiBET (Lutkenhoff et al., 2014). The brain extracted T1 image was then segmented into White Matter, Gray Matter, and Cerebrospinal Fluid using FAST (Zhang et al., 2001).

Functional images were preprocessed separately for the acquisition, extinction, and extinction recall runs. Motion outliers were calculated for functional images using *fsl_motion_outliers* before any preprocessing took place. Then, functional images were motion corrected, smoothed with a 4 mm FWHM kernel, and nonlinearly registered into the MNI152NLin6Asym standard space using 12 degrees of freedom with FSL's FEAT (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). Motion components were then automatically detected and removed using ICA-Aroma (Pruim et al., 2015). ICA-aroma was used as it has been found to be one of the most effective methods of removing motion from fMRI data without discarding large amounts of data (Parkes et al., 2018). Following ICA-aroma, data had linear and quadratic trends removed along with 6 head motion parameters from FSL and white matter and cerebrospinal fluid time courses using AFNI (Cox, 1996; Cox & Hyde, 1997). A highpass filter was also applied to remove frequencies below .008 Hz during this step to remove potential sources of noise. A low-pass filter was not applied to prevent the filtering of any task-relevant content contained in the high frequencies. Finally functional data were transformed into the standard space using parameters from the FSL registration.

Group ICA and dual regression

Similar to preprocessing, a separate group ICA analysis was performed for the acquisition, extinction, and extinction recall fMRI runs. Preprocessed functional data for all participants were submitted to group ICA in Melodic from FSL using the temporal concatenation method. Group ICAs were limited to 20 ICs based on previous work (Webb et al., 2016) and to limit the number of multiple comparisons. The group ICA analysis was masked to brain tissue only using the FSL standard brain mask. Group IC maps were thresholded at Z=4 for generation of figures and identification of key network regions. Any components that resembled physiological noise or did not have significant contributions from the cortex were discarded from further analysis. This resulted in 16 networks analyzed during each task phase. Dual regression was also performed using FSL to obtain participant-specific time course contributions to each IC.

Modeling IC response to task conditions

To find how each IC responded to the conditions of the task, a GLM was fitted to each participant-specific time course for each IC using pyMVPA in python (Hanke, Halchenko, Sederberg, Hanson, et al., 2009). This process is identical to modeling a typical univariate whole-brain GLM but rather than a voxel time course being used as a dependent variable, the time course of an IC is used. Specifically, the conditions of CS+ and CS- were modeled separately for the early and late phases of each run for acquisition and extinction. The unconditioned stimulus (US) was also modeled during acquisition. For extinction recall, early and late unextinguished CS+ (CSU), extinguished CS+ (CSE), and CS- were modeled. Early and late represented the first and last 4 trials of each stimulus type, respectively. A regressor was also included for context presentation (office or conference room image) during each experimental phase. Motion outliers identified in preprocessing were included as additional regressors in the GLM to minimize effects of movement on parameter estimation. GLMs were modeled with an SPM-style hemodynamic response function model including a temporal derivative to account for temporal differences in slice acquisition. All regressors were specified with a duration of 0 seconds for an event-related response over epoch response in order to minimize influence of temporally adjacent events.

Statistical Analysis

To identify which IC networks were sensitive to task conditions, a mixed linear model was calculated for each IC within each task using the statsmodels package in python. In acquisition and extinction, the dependent variable of IC Beta estimate was tested for the interaction of fixed effects condition (CS+/CS-) and time (early/late) for each IC. For extinction recall, we were only interested in the early phase of the task so a model of fixed effect of condition (early CSU vs. early CSE) was tested. Participant and scan site (UCLA/NU) were included as random effects with participant nested within scan site for each model. With a model for each IC not discarded for resembling noise/non-cortical sources, 16 total models were tested for each task phase. Results of each model were bonferroni corrected for multiple comparisons.

Results

Threat Acquisition

During threat acquisition there was a significant interaction between CS-type and Time (*F*(1,888)=15.89, *p*=0.0012 Bonferroni corrected) in a network including vmPFC, OFC, hippocampus, angular gyrus, posterior cingulate, and retrosplenial cortex (Fig. 9A). This interaction was characterized by no initial difference between CS+ and CS- during early trials (*t*(222)=0.99, *p*=0.32) while in late trials network activity was significantly greater for CS- compared to CS+ (*t*(222)=6.71, *p*<0.001). This finding corroborates previous findings of decreased whole-brain connectivity for CS+ compared to CS- in late threat acquisition (Wen et al., 2021) while adding spatial specificity of the involved network. Such late decreases in

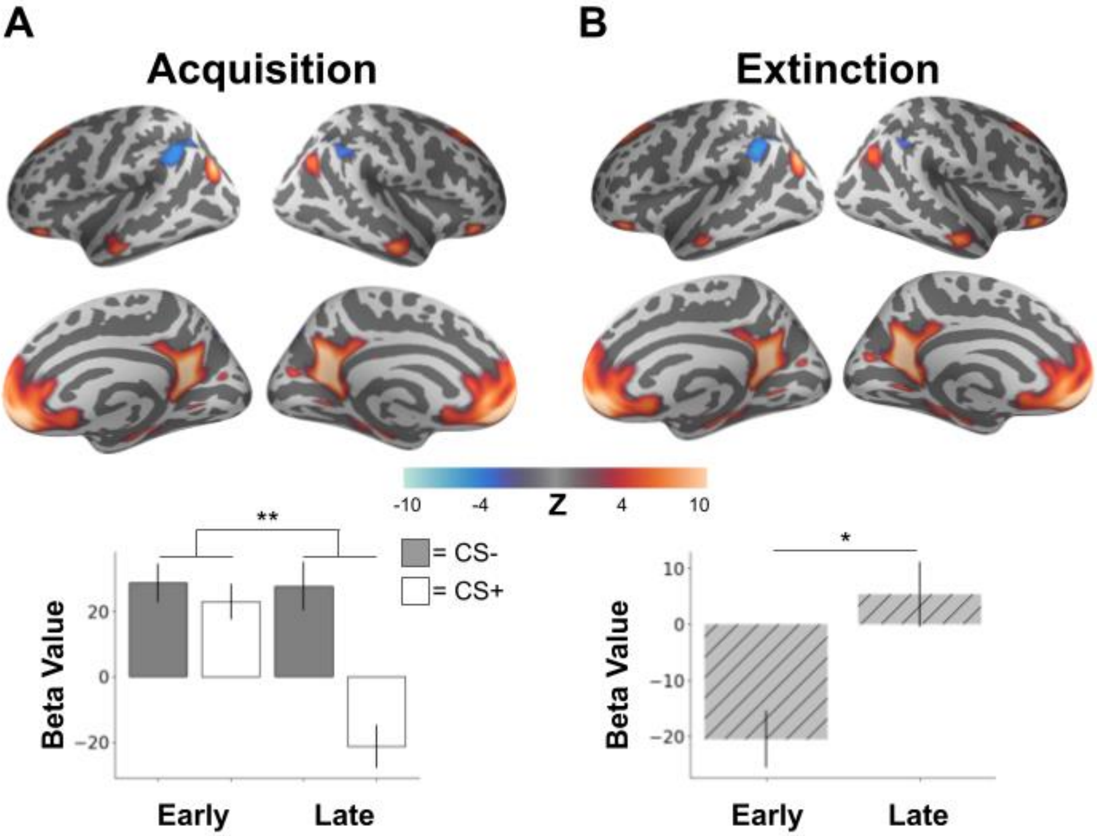

Figure 9. Brain network demonstrating acquisition and extinction of learned threat. (A) A distributed brain network involving bilateral hippocampus, vmPFC, and posterior cingulate demonstrated a significant interaction between CS-type (CS-/CS+) and time (early/late) during threat acquisition. Brain plots show thresholded independent component (IC) spatial maps. Bar plots show IC-specific GLM parameter estimates. During late acquisition, the network demonstrated increased connectivity to the CS- compared to the CS+. (B) This same brain network involving bilateral hippocampus, vmPFC, and posterior cingulate was observed during threat extinction with connectivity in the network increasing from early to late extinction. * p<0.05, ** p<0.01

response to CS+ are thought to identify regions important for threat extinction (Garcia et al., 1999; Hennings et al., 2020; Phelps et al., 2004).

One other network demonstrated a main effect of CS-type with a similar pattern of CS- response being significantly greater than CS+ response (F(1,888)=14.13, p<0.001 Bonferonni corrected). The network consisted of the precentral and postcentral gyri and insular cortex (Fig. 11A).

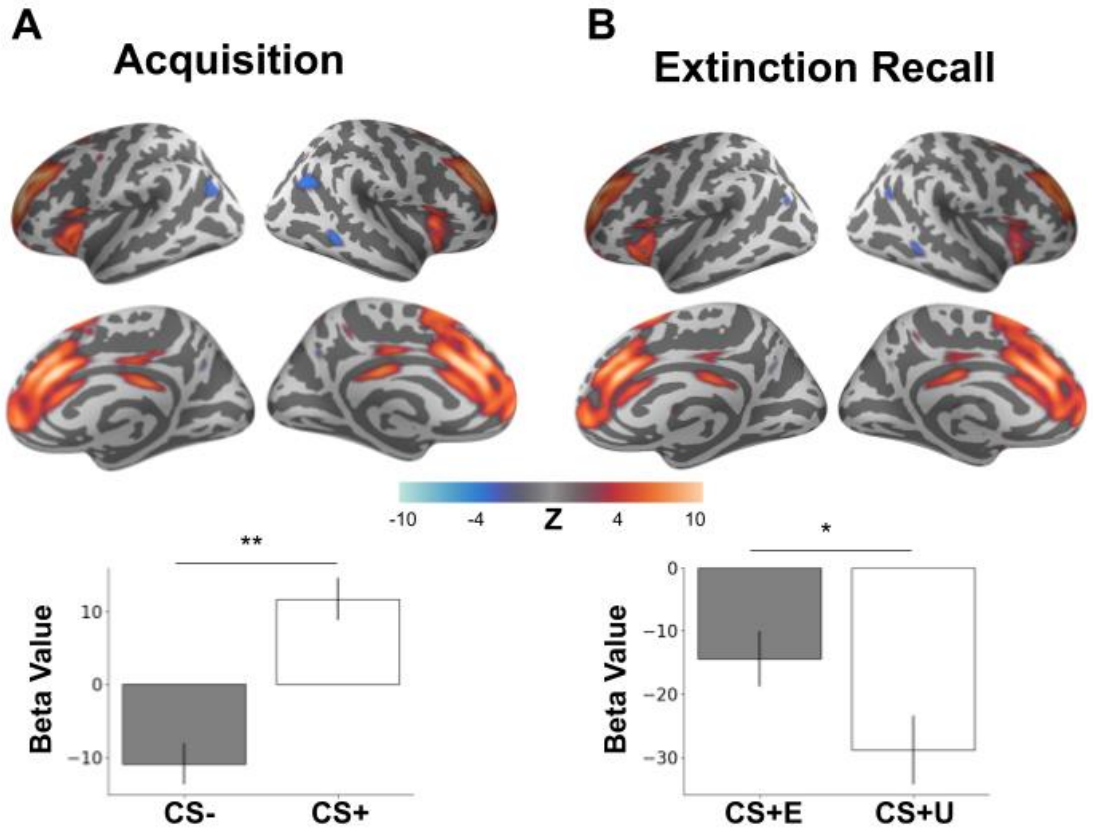

Figure 10. Brain network demonstrating acquisition of learned threat and recall of extinction memory. Brain plots show thresholded independent component (IC) spatial maps. Bar plots show IC-specific GLM parameter estimates. (A) A distributed brain network consisting of dorsal anterior cingulate cortex (dACC), mPFC, and inferior frontal gyrus demonstrated an effect of CS-type during threat acquisition with greater connectivity elicited by the CS+ compared to the CS-. (B) This same brain network was observed during extinction recall with significantly decreased connectivity elicited by the unextinguished CS+ (CS+U) compared to the extinguished CS+ (CS+E).

Finally, two networks also exhibited a main effect of CS-type. These networks demonstrated the canonical threat acquisition response of CS+>CS- indicating successful acquisition of threat-related response. One network spanned the entirety of insular cortex including anterior insula along with cingulate gyrus, inferior frontal gyrus and OFC (*F*(1,888)=109.91, *p*<0.001 Bonferonni corrected, Fig. 11B). The other network included dorsal anterior cingulate cortex

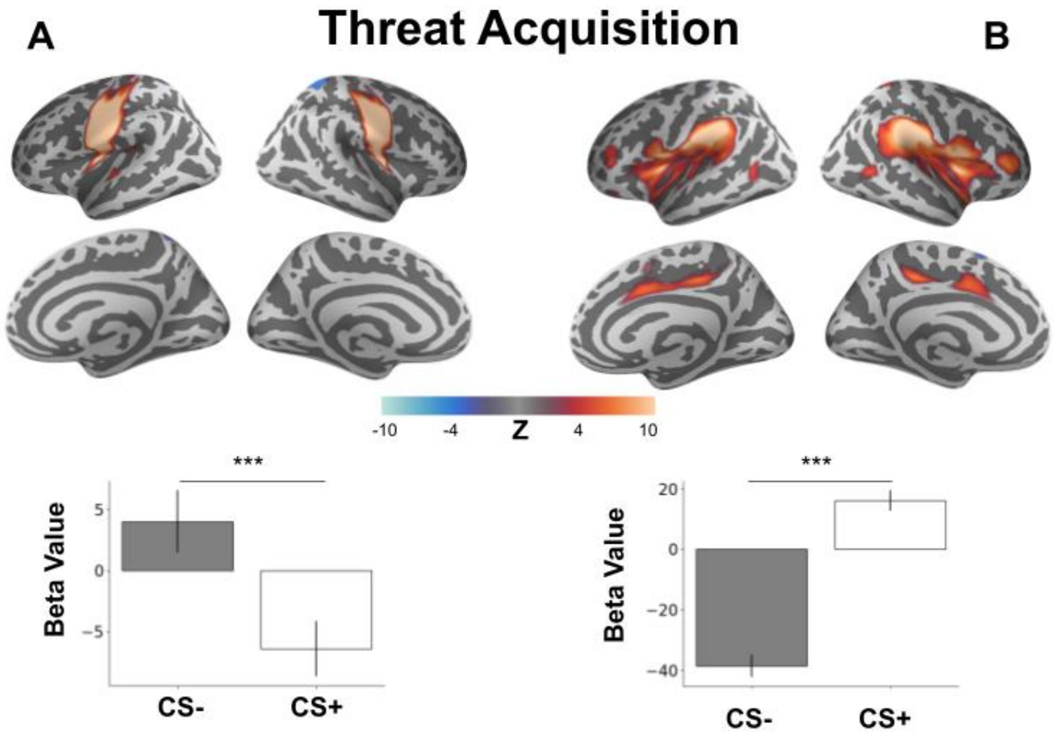

Figure 11. Brain networks involved in learned threat acquisition. Brain plots show thresholded independent component (IC) spatial maps. Bar plots show IC-specific GLM parameter estimates. (A) A brain network involving precentral and postcentral gyri as well as the left insular cortex demonstrated increased connectivity elicited by the CS- and decreased connectivity elicited by the CS+ during threat acquisition. (B) A brain network involving the insula, middle frontal gyrus, cingulate gyrus, and OFC. This network demonstrated acquired threat response with increased connectivity induced by the CS+ and decreased connectivity elicited by the CS-. *** *p*<0.001

(dACC), mPFC, and inferior frontal gyrus (*F*(1,888)=15.70, *p*=0.0013 Bonferonni corrected, Fig. 10A). Both networks showing threat sensitivity overlap with the salience network, indicating the salience network's involvement in learned threat detection.

Threat Extinction

Perplexingly, there was no CS-type sensitivity during threat extinction despite immediately following threat acquisition. No IC networks demonstrated interactions of CS-type and time or main effects of CS-type. However, extinction processes could be tracked through observed main effects of time.

Most saliently, I observed a main effect of time in the same network involving vmPFC, hippocampus, and posterior cingulate from threat acquisition (Fig. 9A) during extinction ($F(1,828)=11.42$, $p=0.012$ Bonferonni corrected, Fig. 9B). This network demonstrated decreased connectivity in response to CS stimuli during early extinction that increased to positive connectivity during late extinction. This change in connectivity likely tracked extinction learning due to the positive increase over the extinction period as opposed to a decrease over the extinction period as might be expected from habituation or unrelated processes. Another network involving lateral occipital cortex and fusiform cortex showed the same connectivity increase from negative to positive over the extinction period ($F(1,828)=30.04$, $p<0.001$ Bonferonni corrected, Fig. 12A). Though this network was not observed during the threat acquisition process, this network activity also likely tracks extinction learning due to the increase of connectivity over the task period.

Finally, I observed a main effect of a time in a third network involving angular gyrus, frontal gyrus, and posterior cingulate ($F(1,828)=21.49$, $p<0.001$ Bonferonni corrected, Fig. 12B). However, as the connectivity pattern of this network decreased from early extinction to late extinction, it is difficult to attribute this result to extinction learning without stimulus-specific evidence as it could also represent stimulus habituation or other unrelated processes.

Extinction Recall

As only the early period of extinction recall was of experimental interest, I ran a model examining only the main effect of CS-type in the early phase of the task (early CS+E vs. early

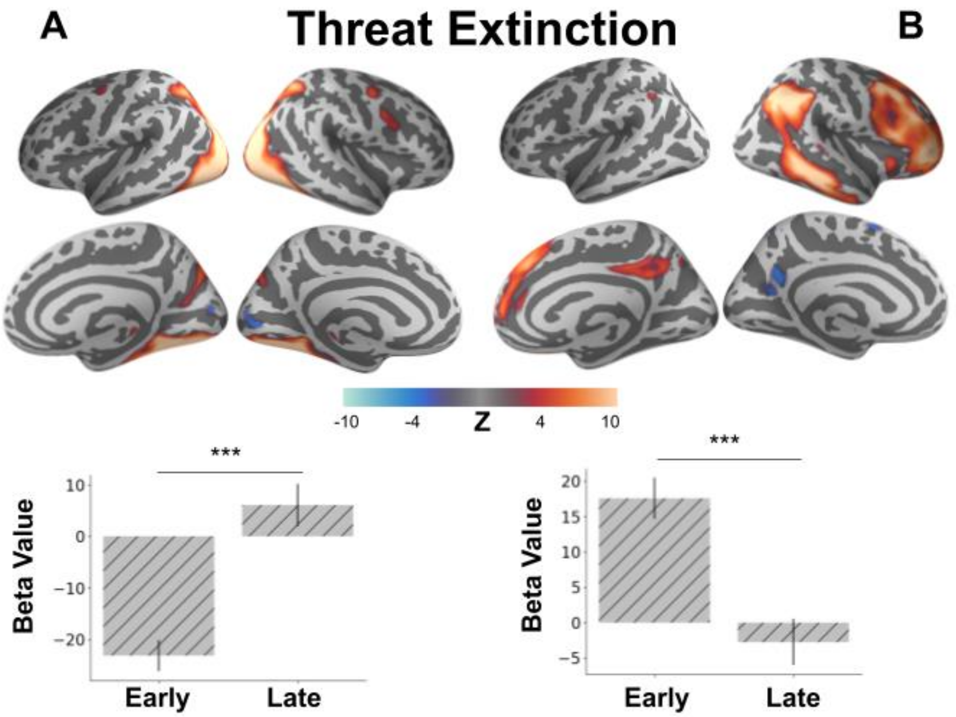

Figure 12. Brain networks involved in the extinction of learned threat response. Brain plots show thresholded independent component (IC) spatial maps. Bar plots show IC-specific GLM parameter estimates. (A) A brain network involving lateral occipital cortex as well as fusiform cortex demonstrates a significant increase in connectivity in response to CS stimuli from early to late trials in the extinction learning phase. (B) A brain network involving angular gyrus, frontal gyrus, and posterior cingulate showed significantly reduced connectivity in response to CS stimuli from early to late trials in the extinction learning phase. ***p<0.001

CS+U). This revealed a singular network that exhibited a greater decrease in connectivity (*F*(1,412)=10.15, *p*=0.025 Bonferonni corrected) for the unextinguished CS+ (CS+U) compared to the extinguished CS+ (CS+E, Fig. 10B). In spatial structure, this network strongly resembled the same network that showed threat sensitivity during acquisition (Fig. 10A) with regions including dACC, mPFC, and inferior frontal gyrus. These results showcase the importance of this network both in the acquisition of threat learning as well as the expression of extinction memory.

Discussion

In this study, I examined brain connectivity networks during a 2-day Pavlovian threat conditioning paradigm. Using group independent component analysis, I compared how independent brain networks responded to CS stimuli during threat acquisition and extinction as well as extinction recall a full 48 hours later. This revealed multiple distinct brain networks involved in the acquisition, extinction, and recall of extinction memory for learned threat. A stable network overlapping with the Default Mode Network involving hippocampus, vmPFC, and posterior cingulate was involved in both the acquisition and extinction of learned threat. An additional persisting network overlapping with the Salience Network involving dACC, mPFC, and inferior frontal gyrus was involved in the acquisition of learned threat as well as the expression of extinction memory. A number of other networks were independently involved in the acquisition and extinction of learned threat.

The finding of a network involving hippocampus and vmPFC is consistent with previous work that has focused on these regions as part of a 'network' that consistently responds to threat conditioning paradigms (Giustino & Maren, 2015; Picó-Pérez et al., 2019). It is worth noting that these regions have been identified in the current work without the imposition of a model through the a priori selection of regions of interest. This strengthens the argument for them as canonical threat learning regions while demonstrating the coordination of these regions in a self-contained connectivity network. Responses specific to the conditioned cue developed late in the threat acquisition phase, indicating this network's involvement in the learning aspect of threat acquisition. Connectivity in this network then increased during the extinction process, again indicating a learning process over the extinction period. As the observed network partially overlaps with the canonical default mode network, this adds to a building body of evidence in human neuroimaging implicating the default mode network in threat learning (Berg et al., 2021; Wen et al., 2021; Zidda et al., 2018).

I also observed a stable network across threat acquisition and extinction recall in dACC, inferior frontal gyrus, and mPFC overlapping partly with the salience network. As the observed stimulus specificity during acquisition in this network was not specific to the late acquisition period, it is possible that this network detects learned threats at a rapid rate (e.g. one-shot learning). This network was also the only network to show lasting conditioned stimulus specificity between the extinguished and unextinguished conditioned stimuli a full 48 hours after the threat acquisition and extinction periods. This underscores the importance of this network of regions in the threat learning process as a site of potentially rapid and lasting threat memory acquisition. This may be of broad clinical significance in understanding anxiety and fear-related disorders as the salience network itself has been found to track fear generalization and symptom severity in post-traumatic stress disorder (Berg et al., 2021). Perhaps a rapid learning rate within this network predisposes it to overgeneralized threat.

Rapid threat learning within the salience network is further supported by the current work with the finding of an additional insula-centered network during threat acquisition where conditioned stimulus specificity emerged early. It is unlikely that these early stimulus differences are driven primarily by confounding of the US during threat acquisition due to the slow trial structure, event-related duration modeling, and explicit modeling of the US in the independent component GLMs. Additionally, results in both of these networks are characterized by a substantial decrease in connectivity in response to the CS- as opposed to increased connectivity to CS+ stimuli and there is no possibility of US contamination on CS- trials.

Interestingly, there was no conditioned stimulus specificity exhibited during the extinction period in this network or any examined network. This is all the more perplexing given the extinction phase immediately followed the threat acquisition phase, in which conditioned stimulus specificity was widely observed. This broadly matches other recent whole-brain connectivity findings (Wen et al., 2021) which also observed no conditioned stimulus specific response early in extinction (though specificity was observed by the end of extinction). While it is difficult to be certain why this was the case within the current study, it may be related to the presentation of the conditioned stimuli within varying contexts across phases. Threat extinction always took place in a different context than threat acquisition. So, it may be that the new context presented in threat extinction was a sufficient safety signal over the extinction period, that participants did not pay as much attention to specific conditioned stimulus identities. However, as there was conditioned stimulus specificity during extinction recall in the same extinction context 48 hours later, clearly context presentation alone does not sufficiently explain the lack of stimulus specificity in extinction. It could be additionally related to consolidation or habituation processes that come with examining stimulus differences immediately following learning and after multiple days of memory consolidation.

It should also be noted that the results from the current study come from a large sample size of more than 200 participants. As neuroimaging frequently suffers from low sensitivity and under-powered sample sizes (Thirion et al., 2007), the findings here can be considered robust. Despite this high-powered sample, significant amygdala involvement was conspicuously missing from any of the networks found in this current analysis. This matches large meta-analyses as well as other findings that find a minimal role (if any) for the amygdala in human threat conditioning paradigms (Fullana et al., 2016; Visser et al., 2021).

In summary, the findings of this study indicate a distributed response from multiple independent brain networks during the acquisition, extinction, and recall of learned threat memories spanning canonical networks like the default mode and salience networks. This echoes other recent findings from whole-brain connectivity analyses of threat conditioning that demonstrate the need for researchers to move beyond the previously focal region of interest analyses of threat conditioning paradigms in order to understand the dynamic nature of the human brain's response in threat learning (Wen et al., 2021). The present work has the added benefit of refining these whole-brain connectivity patterns into independent networks to identify which regions work directly in concert as well as which connectivity networks are involved in different aspects of threat learning and extinction. Future work will need to disentangle which of these networks are involved in automatic defensive responses to threat and which contribute to the actual subjective experience of fear and threat in the human brain in order to most appropriately target clinical interventions for fear-related disorders.

Chapter 5. General Discussion.

In this thesis, I investigated threat and fear processes in the human brain across three chapters. In Chapter 2, I used real-time fMRI neuro-reinforcement to causally intervene on specific phobia in a randomized double-blind placebo-controlled clinical trial. Then, in Chapter 3, I leveraged machine-learning techniques on two independent datasets using a Pavlovian threat conditioning paradigm to identify neural patterns associated with the subjective awareness of threat. Lastly, to further understand whole-brain dynamics behind human response to learned threat, I explored whole-brain connectivity networks during the same Pavlovian threat conditioning paradigm in Chapter 4.

Results from across these 3 studies came together to demonstrate a broadly distributed response to threat and fear across the human brain. This response was composed of multiple parallel but interacting processes leading to the expression of threat and fear responses through the domains of subjective experience, behavior, and physiology. Findings in Chapter 2 demonstrated that implicit activation of feared visual representations at a nonconscious level has the potential to reduce threat responses both neurally and behaviorally in specific phobia. Following multi-voxel neuro-reinforcement there was evidence of reduced amygdala response to phobic images as well as reduced attentional capture by targeted phobias. Interestingly, despite both these observed changes, there were no changes in the self-reported levels of fear, indicating subjective experience may not be affected by modulating threat processes at an implicit nonconscious level. These findings broadly match our previous proof-of-concept study in a non-clinical population (Taschereau-Dumouchel et al., 2018). Importantly, the observed changes were obtained with no distressing conscious exposures or otherwise distressing or panic-inducing sensations. This highlights multi-voxel neuro-reinforcement as a promising intervention for specific phobia which may facilitate more traditional behavioral exposure treatments by reducing physiological and behavioral expressions of threat and fear. Following multi-voxel neuro-reinforcement, traditional exposure may lead to less attrition if reflexive responses are dampened and the instinct to avoid the feared stimulus is reduced. The discordance in the observed changes following multi-voxel neuro-reinforcement underscores the importance of understanding the independent processes contributing to fear and threat responses.

In Chapter 3 I found robust whole-brain classification of the subjective awareness of threat in a Pavlovian threat conditioning task using a machine-learning classifier. Classifier performance generalized well to a new independent dataset. Iterative within-region classification revealed a distributed fingerprint for subjective threat awareness organized in an hierarchical fashion across the cortex. Subjective threat awareness could be not classified from early visual regions but became more classifiable further up the visual hierarchy. Brain regions containing the most information about subjective awareness of threat were posterior cingulate, retrosplenial cortex, vmPFC, inferior frontal gyrus, OFC, and hippocampus. These results help inform which brain regions are critical for the actual subjective experience of threat while most previous neuroimaging analyses of threat conditioning have failed to take subjective threat learning into account in place of assumed perfect threat learning (Fullana et al., 2016).

The connectivity network results from Chapter 4 collaborate this multifaceted view of human fear and threat responding. Group independent component analysis performed on fMRI data from threat acquisition, extinction, and extinction recall in a Pavlovian threat conditioning paradigm revealed a number of independent connectivity networks involved in the acquisition and extinction of learned threat memory. These independent networks followed a variety of response patterns with increasing and decreasing connectivity as well as opposing stimulus specificity at overlapping timepoints during threat and safety learning. I found a stable network involved in threat acquisition and extinction overlapping with the Default Mode Network including the hippocampus, vmPFC, and posterior cingulate. I also observed a network persisting across threat acquisition and extinction recall overlapping with the Salience network including dACC, mPFC, and inferior frontal gyrus. These networks demonstrated opposing stimulus specificities during threat acquisition with the Default Mode Network-overlapping network responding more to the safety stimulus while the Salience Network-overlapping network responded more to the threat stimulus. These findings add to a growing body of literature examining connectivity during threat conditioning (Berg et al., 2021; Wen et al., 2021) while adding subnetwork specificity (rather than examining the average connectivity across the whole-brain). As such a variety of responses can be observed in parallel, these results hazard against the oversimplification of the brain's response to threat as a singular process.

However, these findings are not without their limitations. In Chapter 2, the finding of reduced amygdala activation to the target phobia only reached trending significance. This is most likely due to insufficient power as the sample size of 18 participants is small compared to modern fMRI standards (Turner et al., 2018). This power issue unfortunately can not be combated with increased trial counts per participant as the studied amygdala response is transient, habituating after just a few trials. As such, only the first two trials are examined per participant following previous methodologies (Koizumi et al., 2017; Schiller et al., 2010) leaving a greater possibility of inconsistent measurement as compared to other fMRI studies with little to be done to combat these issues within participant. Nonetheless, target-specific engagement of the amygdala had been found in a similarly sized non-clinical population in the proof-of-concept study (Taschereau-Dumouchel et al., 2018). This issue highlights the difficulty in demonstrating the effect of multi-voxel neuro-reinforcement with primarily neural outcomes as this modest sample size still represents a significant effort and cost totaling over 100 separate fMRI sessions due to the multi-session nature of the intervention. Reaching a sample size of 100+ participants within a single multi-voxel neuro-reinforcement study would be an intractable and prohibitively costly endeavor.

A preferable alternative would be an effect demonstrated through a behavioral paradigm in which power issues are not so prevalent in smaller sample sizes. This is a matter of finding the right task as subjective fear ratings did not change in the results reported in Chapter 2 or in the proof-of-concept study (Taschereau-Dumouchel et al., 2018). The affective stroop task deployed in Chapter 2 seems to be a promising addition as this effect did reach statistical significance for the targeted phobia. So despite the fact that patients may report the same subjective level of fear following multi-voxel neuro-reinforcement, behavior may be significantly changed as measured by tasks involving reflexive and automatic responses. In future studies, besides a larger sample size, it would be informative to see other behavioral tasks implemented in addition to the affective stroop. With the increasing popularity of virtual reality, future studies should see if patients demonstrate altered approach or avoidance in a behavioral approach task performed in virtual space despite reporting the same subjective level of fear.

I sought to address some of the limitations in interpretation from previous fMRI studies of threat conditioning with the analysis reported in Chapter 3. By incorporating participants' self-reported threat-stimulus contingency awareness ratings into the analysis, the results were able to account for individual differences in threat learning such as those that failed to properly identify the CS+ with threat or those that overgeneralized to view the CS- as threatening. However, the results in Chapter 3 were still limited by a number of factors. Most critically, despite explicitly analyzing the subjective awareness of threat, subjective awareness could not be properly dissociated from other correlated confounding processes like physiological and nonconscious threat responding. Consequently, it is difficult to know if high classification accuracy of subjective awareness of threat in subcortical regions like hippocampus and amygdala is due to these regions contributing to actual subjective awareness or other lower-level threat processes not directly contributing to subjective awareness. Additionally, as contingency ratings were collected at the end of each task block rather than after each stimulus presentation, it is difficult to know how threat awareness developed and diminished over time during threat acquisition and extinction.

In Chapter 4, findings were limited in interpretation due to the nature of the Pavlovian threat conditioning task. Despite its simple form and prolific use, it is not clear exactly what processes are tracked through the task. This analysis, like the vast majority of analyses of the task (Fullana et al., 2016), relied on contrasts between conditioned stimulus types (CS+E, CS+U, and CS-) to derive meaning. However, exactly what differences between these stimuli reflect is not always clear. The meaning is perhaps most clear during the threat acquisition phase where some sort of “threat” response is expected to be acquired, measured by the difference between CS+ and CS-. However, oftentimes the inverse is found with CS- acquiring a response relative to CS+ as it was in some networks reported in Chapter 4 (Fullana et al., 2016; Wen et al., 2021). This has led to theorizing around what may be a “safety” signal learned in the task but little work has been done examining the overlap and differences between a “threat” versus “safety” signal in the classic conditioning model (Fullana et al., 2016). These conceptual difficulties only compound when considering the extinction and extinction recall phases. Beyond detection and removal of the original “threat” signal in extinction, it is not clear what an expected result should be. By this standard, the desired result would be a null result at the end of extinction. Conversely, perhaps stimulus specificity emerges at the end of extinction. If it was not present at the end of acquisition, it is not clear what this stimulus specificity indicates between learned safety or persisting threat. Finally, things become all the more unclear in extinction recall where stimulus differences can not be adequately discriminated between signals of renewed fear, original fear, or expression of an “extinction memory”. Without an understanding of what a safety signal or extinction memory is, it is difficult to distinguish extinction recall results from simply lingering threat acquisition.

In future studies, explicit consideration of the differing processes contributing to subjective awareness, behavior, and physiology surrounding threat should be given in the task design. Rather than passive stimulus viewing, a behavioral component could be integrated where potentially threatening stimuli or approached or avoided with differing consequences for unconditioned stimulus delivery. Additionally, it would be informative to add explicit fear ratings for conditioned stimuli. Subjective awareness of threat is likely not a homogenous process itself with different regions potentially being implicated for the simple awareness of threat compared to the emotional experience of fear in the face of threat. Moreover, understanding the nonconscious processes of threat responding is also critical. A future experiment in which nonconscious presentations of conditioned threat stimuli (using masking or near-threshold presentations) are assessed for both neural and behavioral response properties would be very informative of which threat responses and behaviors require subjective awareness. Lastly, while fMRI has superior spatial resolution, a time-resolved analysis of these paradigms with a modality such as magnetoencephalography or electroencephalography would be greatly beneficial. This would allow the temporal evolution of learned threat and fear responses to be assessed within trial, with reflexive and automatic responses occurring within the first 200 milliseconds while subjective awareness processes would emerge over a longer timescale.

In closing, this thesis investigated human fear and threat responses using both traditional fMRI and closed-loop real-time fMRI feedback. The findings of these studies highlighted the importance of understanding the human fear response through the multiple domains through which it expresses itself such as subjective experience, behavior, and physiology. While these findings demonstrate promise both for the understanding of as well as clinical intervention in the human experience of fear, there is still much to be done in future studies to understand the full complexity of human threat and fear responses.

Appendices

Appendix A. Within-region permutation test results

Table A1. Non-parametric permutation test results for within-region decoding analyses. A machine learning classifier was trained within each region to classify subjective threat awareness. Label names are reported in the left column with parcellations coming from the 200 parcel Schaefer atlas (Schaefer et al., 2018) and amygdala and hippocampus segmentations from the Harvard-Oxford atlas (Destrieux et al., 2010). Classifier performance was assessed with area under the receiver operating characteristic curve (AUC) across 20 repeated split-half cross-validations. Non-parametric null distribution was calculated by performing full cross-validation on 1000 iterations of data with permuted class labels. Non-parametric p values were calculated by fitting a non-parametric gaussian kernel to the null distribution and estimating the density function for the observed AUC. Significance was corrected for multiple comparisons by False-Discovery Rate correction with threshold q=0.001.

Label	AUC	*p*-value	Significance
7Networks_LH_Vis_1	0.632766	1.20E-14	TRUE
7Networks_LH_Vis_2	0.562645	1	FALSE
7Networks_LH_Vis_3	0.584269	0.114602	FALSE
7Networks_LH_Vis_4	0.617315	2.19E-09	TRUE
7Networks_LH_Vis_5	0.577447	0.471595	FALSE
7Networks_LH_Vis_6	0.661548	1.53E-32	TRUE
7Networks_LH_Vis_7	0.60672	0.06665	FALSE
7Networks_LH_Vis_8	0.655234	3.85E-34	TRUE
7Networks_LH_Vis_9	0.553134	1	FALSE
7Networks_LH_Vis_10	0.589952	2.31E-05	TRUE
7Networks_LH_Vis_11	0.551608	1	FALSE
7Networks_LH_Vis_12	0.660436	8.99E-26	TRUE
7Networks_LH_Vis_13	0.593992	0.002726	FALSE
7Networks_LH_Vis_14	0.576287	0.305778	FALSE
7Networks_LH_SomMot_1	0.771042	8.05E-104	TRUE
7Networks_LH_SomMot_2	0.799083	7.64E-148	TRUE
7Networks_LH_SomMot_3	0.787974	2.38E-123	TRUE
7Networks_LH_SomMot_4	0.785689	1.43E-164	TRUE

7Networks_LH_SomMot_5	0.775743	8.02E-107	TRUE
7Networks_LH_SomMot_6	0.775704	6.75E-162	TRUE
7Networks_LH_SomMot_7	0.753579	3.03E-85	TRUE
7Networks_LH_SomMot_8	0.804818	2.22E-163	TRUE
7Networks_LH_SomMot_9	0.754514	5.36E-121	TRUE
7Networks_LH_SomMot_10	0.788915	4.82E-148	TRUE
7Networks_LH_SomMot_11	0.734933	2.27E-71	TRUE
7Networks_LH_SomMot_12	0.780689	2.03E-129	TRUE
7Networks_LH_SomMot_13	0.784502	2.58E-162	TRUE
7Networks_LH_SomMot_14	0.775841	7.87E-102	TRUE
7Networks_LH_SomMot_15	0.778505	1.84E-151	TRUE
7Networks_LH_SomMot_16	0.764274	4.97E-114	TRUE
7Networks_LH_DorsAttn_Post_1	0.671068	1.58E-36	TRUE
7Networks_LH_DorsAttn_Post_2	0.654815	1.07E-09	TRUE
7Networks_LH_DorsAttn_Post_3	0.677605	3.33E-14	TRUE
7Networks_LH_DorsAttn_Post_4	0.740764	1.68E-63	TRUE
7Networks_LH_DorsAttn_Post_5	0.745178	2.37E-102	TRUE
7Networks_LH_DorsAttn_Post_6	0.687799	1.17E-27	TRUE
7Networks_LH_DorsAttn_Post_7	0.64221	1.16E-08	TRUE
7Networks_LH_DorsAttn_Post_8	0.612601	0.002658	FALSE
7Networks_LH_DorsAttn_Post_9	0.709223	6.94E-53	TRUE
7Networks_LH_DorsAttn_Post_10	0.723036	8.90E-71	TRUE
7Networks_LH_DorsAttn_FEF_1	0.759111	9.37E-88	TRUE
7Networks_LH_DorsAttn_FEF_2	0.77028	4.26E-110	TRUE
7Networks_LH_DorsAttn_PrCv_1	0.702639	3.72E-35	TRUE
7Networks_LH_SalVentAttn_ParOper_1	0.758364	1.19E-86	TRUE
7Networks_LH_SalVentAttn_ParOper_2	0.763731	2.97E-106	TRUE
7Networks_LH_SalVentAttn_ParOper_3	0.72553	6.79E-69	TRUE
7Networks_LH_SalVentAttn_FrOper_1	0.800629	1.02E-144	TRUE
7Networks_LH_SalVentAttn_FrOper_2	0.807903	1.89E-146	TRUE

7Networks_LH_SalVentAttn_FrOper_3	0.801054	4.25E-147	TRUE
7Networks_LH_SalVentAttn_FrOper_4	0.775122	4.65E-144	TRUE
7Networks_LH_SalVentAttn_PFCl_1	0.769382	6.44E-96	TRUE
7Networks_LH_SalVentAttn_Med_1	0.789082	2.70E-134	TRUE
7Networks_LH_SalVentAttn_Med_2	0.792614	8.36E-143	TRUE
7Networks_LH_SalVentAttn_Med_3	0.761205	8.30E-96	TRUE
7Networks_LH_Limbic_OFC_1	0.796754	2.85E-135	TRUE
7Networks_LH_Limbic_OFC_2	0.797986	8.86E-137	TRUE
7Networks_LH_Limbic_TempPole_1	0.804719	5.72E-148	TRUE
7Networks_LH_Limbic_TempPole_2	0.741119	2.18E-97	TRUE
7Networks_LH_Limbic_TempPole_3	0.78426	4.04E-153	TRUE
7Networks_LH_Limbic_TempPole_4	0.779584	8.50E-154	TRUE
7Networks_LH_Cont_Par_1	0.710183	2.42E-67	TRUE
7Networks_LH_Cont_Par_2	0.651319	2.24E-15	TRUE
7Networks_LH_Cont_Par_3	0.696257	4.25E-44	TRUE
7Networks_LH_Cont_Temp_1	0.700909	1.22E-53	TRUE
7Networks_LH_Cont_PFCl_1	0.766817	6.78E-121	TRUE
7Networks_LH_Cont_PFCl_2	0.75645	2.83E-84	TRUE
7Networks_LH_Cont_PFCl_3	0.732034	7.39E-66	TRUE
7Networks_LH_Cont_PFCl_4	0.724721	2.30E-48	TRUE
7Networks_LH_Cont_PFCl_5	0.726062	3.61E-51	TRUE
7Networks_LH_Cont_PFCl_6	0.748734	8.21E-96	TRUE
7Networks_LH_Cont_pCun_1	0.661642	4.38E-31	TRUE
7Networks_LH_Cont_Cing_1	0.758878	4.56E-94	TRUE
7Networks_LH_Cont_Cing_2	0.775324	2.40E-133	TRUE
7Networks_LH_Default_Temp_1	0.794165	1.79E-146	TRUE
7Networks_LH_Default_Temp_2	0.734247	3.76E-55	TRUE
7Networks_LH_Default_Temp_3	0.767173	1.46E-114	TRUE
7Networks_LH_Default_Temp_4	0.756753	5.07E-76	TRUE
7Networks_LH_Default_Temp_5	0.754365	3.14E-122	TRUE

7Networks_LH_Default_Temp_6	0.732778	3.02E-74	TRUE
7Networks_LH_Default_Temp_7	0.604135	3.02E-07	TRUE
7Networks_LH_Default_Temp_8	0.699805	1.99E-61	TRUE
7Networks_LH_Default_Temp_9	0.68167	1.56E-39	TRUE
7Networks_LH_Default_PFC_1	0.791686	4.58E-149	TRUE
7Networks_LH_Default_PFC_2	0.768361	6.73E-86	TRUE
7Networks_LH_Default_PFC_3	0.742701	4.58E-92	TRUE
7Networks_LH_Default_PFC_4	0.750188	2.16E-74	TRUE
7Networks_LH_Default_PFC_5	0.754086	3.55E-118	TRUE
7Networks_LH_Default_PFC_6	0.781946	4.70E-163	TRUE
7Networks_LH_Default_PFC_7	0.772326	4.71E-113	TRUE
7Networks_LH_Default_PFC_8	0.77706	1.90E-112	TRUE
7Networks_LH_Default_PFC_9	0.771017	5.05E-101	TRUE
7Networks_LH_Default_PFC_10	0.761008	7.34E-127	TRUE
7Networks_LH_Default_PFC_11	0.740845	7.79E-67	TRUE
7Networks_LH_Default_PFC_12	0.781312	1.28E-119	TRUE
7Networks_LH_Default_PFC_13	0.781377	6.55E-115	TRUE
7Networks_LH_Default_PCC_1	0.722017	3.07E-72	TRUE
7Networks_LH_Default_PCC_2	0.760726	2.68E-121	TRUE
7Networks_LH_Default_PCC_3	0.795239	1.52E-148	TRUE
7Networks_LH_Default_PCC_4	0.75216	1.37E-73	TRUE
7Networks_LH_Default_PHC_1	0.78573	1.26E-101	TRUE
7Networks_RH_Vis_1	0.730297	3.95E-64	TRUE
7Networks_RH_Vis_2	0.783019	7.54E-120	TRUE
7Networks_RH_Vis_3	0.590723	0.048796	FALSE
7Networks_RH_Vis_4	0.645717	7.61E-22	TRUE
7Networks_RH_Vis_5	0.591284	0.107845	FALSE
7Networks_RH_Vis_6	0.543279	1	FALSE
7Networks_RH_Vis_7	0.757153	1.86E-115	TRUE
7Networks_RH_Vis_8	0.535324	1	FALSE

7Networks_RH_Vis_9	0.600959	0.040507	FALSE
7Networks_RH_Vis_10	0.731587	1.05E-81	TRUE
7Networks_RH_Vis_11	0.608348	0.026187	FALSE
7Networks_RH_Vis_12	0.557101	0.110835	FALSE
7Networks_RH_Vis_13	0.722415	6.80E-85	TRUE
7Networks_RH_Vis_14	0.635928	2.64E-12	TRUE
7Networks_RH_Vis_15	0.669223	2.76E-16	TRUE
7Networks_RH_SomMot_1	0.769863	9.98E-111	TRUE
7Networks_RH_SomMot_2	0.73957	8.25E-85	TRUE
7Networks_RH_SomMot_3	0.795122	3.40E-131	TRUE
7Networks_RH_SomMot_4	0.771923	1.92E-132	TRUE
7Networks_RH_SomMot_5	0.750952	2.75E-62	TRUE
7Networks_RH_SomMot_6	0.764146	2.09E-92	TRUE
7Networks_RH_SomMot_7	0.722627	1.43E-62	TRUE
7Networks_RH_SomMot_8	0.788706	1.56E-138	TRUE
7Networks_RH_SomMot_9	0.697598	6.25E-52	TRUE
7Networks_RH_SomMot_10	0.722756	1.57E-54	TRUE
7Networks_RH_SomMot_11	0.787032	5.18E-105	TRUE
7Networks_RH_SomMot_12	0.741859	1.48E-141	TRUE
7Networks_RH_SomMot_13	0.710141	3.49E-54	TRUE
7Networks_RH_SomMot_14	0.759057	1.83E-85	TRUE
7Networks_RH_SomMot_15	0.717856	8.83E-35	TRUE
7Networks_RH_SomMot_16	0.749068	1.56E-102	TRUE
7Networks_RH_SomMot_17	0.737918	7.67E-76	TRUE
7Networks_RH_SomMot_18	0.761613	1.47E-139	TRUE
7Networks_RH_SomMot_19	0.719278	3.23E-31	TRUE
7Networks_RH_DorsAttn_Post_1	0.699742	5.11E-14	TRUE
7Networks_RH_DorsAttn_Post_2	0.753189	5.11E-86	TRUE
7Networks_RH_DorsAttn_Post_3	0.728684	1.11E-83	TRUE
7Networks_RH_DorsAttn_Post_4	0.664122	1.07E-12	TRUE

7Networks_RH_DorsAttn_Post_5	0.716626	2.27E-40	TRUE
7Networks_RH_DorsAttn_Post_6	0.666373	8.55E-24	TRUE
7Networks_RH_DorsAttn_Post_7	0.71081	5.51E-49	TRUE
7Networks_RH_DorsAttn_Post_8	0.670391	9.40E-28	TRUE
7Networks_RH_DorsAttn_Post_9	0.729848	9.83E-64	TRUE
7Networks_RH_DorsAttn_Post_10	0.725605	3.60E-68	TRUE
7Networks_RH_DorsAttn_FEF_1	0.74574	8.80E-80	TRUE
7Networks_RH_DorsAttn_FEF_2	0.766602	2.29E-85	TRUE
7Networks_RH_DorsAttn_PrCv_1	0.741432	2.26E-69	TRUE
7Networks_RH_SalVentAttn_TempOccPar_1	0.743861	7.42E-109	TRUE
7Networks_RH_SalVentAttn_TempOccPar_2	0.737418	2.09E-60	TRUE
7Networks_RH_SalVentAttn_TempOccPar_3	0.746459	2.21E-74	TRUE
7Networks_RH_SalVentAttn_PrC_1	0.685396	1.55E-39	TRUE
7Networks_RH_SalVentAttn_FrOper_1	0.778321	4.56E-102	TRUE
7Networks_RH_SalVentAttn_FrOper_2	0.794254	1.20E-134	TRUE
7Networks_RH_SalVentAttn_FrOper_3	0.775576	2.78E-122	TRUE
7Networks_RH_SalVentAttn_FrOper_4	0.790091	5.31E-156	TRUE
7Networks_RH_SalVentAttn_Med_1	0.804017	5.46E-168	TRUE
7Networks_RH_SalVentAttn_Med_2	0.792857	9.84E-135	TRUE
7Networks_RH_SalVentAttn_Med_3	0.769307	5.81E-120	TRUE
7Networks_RH_Limbic_OFC_1	0.800095	7.27E-130	TRUE
7Networks_RH_Limbic_OFC_2	0.790166	2.19E-105	TRUE
7Networks_RH_Limbic_OFC_3	0.753318	1.46E-113	TRUE
7Networks_RH_Limbic_TempPole_1	0.784085	3.85E-125	TRUE
7Networks_RH_Limbic_TempPole_2	0.764653	9.56E-95	TRUE
7Networks_RH_Limbic_TempPole_3	0.809242	8.51E-171	TRUE
7Networks_RH_Cont_Par_1	0.705036	2.73E-41	TRUE
7Networks_RH_Cont_Par_2	0.691208	2.74E-35	TRUE
7Networks_RH_Cont_Par_3	0.709573	1.13E-33	TRUE
7Networks_RH_Cont_Temp_1	0.721735	4.77E-50	TRUE

7Networks_RH_Cont_PFCv_1	0.774332	1.21E-126	TRUE
7Networks_RH_Cont_PFCl_1	0.760986	2.89E-90	TRUE
7Networks_RH_Cont_PFCl_2	0.740101	4.45E-91	TRUE
7Networks_RH_Cont_PFCl_3	0.710852	8.22E-42	TRUE
7Networks_RH_Cont_PFCl_4	0.7252	5.94E-46	TRUE
7Networks_RH_Cont_PFCl_5	0.734115	3.73E-96	TRUE
7Networks_RH_Cont_PFCl_6	0.721183	6.96E-49	TRUE
7Networks_RH_Cont_PFCl_7	0.746382	1.93E-89	TRUE
7Networks_RH_Cont_pCun_1	0.73204	9.70E-69	TRUE
7Networks_RH_Cont_PFCmp_1	0.798711	2.31E-130	TRUE
7Networks_RH_Cont_PFCmp_2	0.785594	3.02E-129	TRUE
7Networks_RH_Cont_PFCmp_3	0.78853	1.01E-126	TRUE
7Networks_RH_Cont_PFCmp_4	0.767855	1.89E-96	TRUE
7Networks_RH_Default_Par_1	0.682767	1.26E-28	TRUE
7Networks_RH_Default_Par_2	0.747163	1.57E-81	TRUE
7Networks_RH_Default_Par_3	0.708224	9.47E-17	TRUE
7Networks_RH_Default_Temp_1	0.756282	9.81E-76	TRUE
7Networks_RH_Default_Temp_2	0.749298	2.64E-73	TRUE
7Networks_RH_Default_Temp_3	0.788037	2.32E-149	TRUE
7Networks_RH_Default_Temp_4	0.722383	1.19E-91	TRUE
7Networks_RH_Default_Temp_5	0.787762	4.21E-176	TRUE
7Networks_RH_Default_PFCv_1	0.749247	3.11E-73	TRUE
7Networks_RH_Default_PFCm_1	0.807987	1.84E-151	TRUE
7Networks_RH_Default_PFCm_2	0.77118	3.73E-132	TRUE
7Networks_RH_Default_PFCm_3	0.754254	2.50E-95	TRUE
7Networks_RH_Default_PFCm_4	0.781183	4.03E-114	TRUE
7Networks_RH_Default_PFCm_5	0.7541	8.10E-70	TRUE
7Networks_RH_Default_PFCm_6	0.744002	3.21E-79	TRUE
7Networks_RH_Default_PFCm_7	0.737682	1.04E-70	TRUE
7Networks_RH_Default_PCC_1	0.807347	7.18E-186	TRUE

7Networks_RH_Default_PCC_2	0.819644	9.32E-145	TRUE
7Networks_RH_Default_PCC_3	0.740253	1.15E-75	TRUE
HOSPA_desc-Left_Amygdala-th50	0.805515	6.95E-152	TRUE
HOSPA_desc-Left_Hippocampus-th50	0.833426	3.01E-212	TRUE
HOSPA_desc-Right_Amygdala-th50	0.792829	2.14E-156	TRUE
HOSPA_desc-Right_Hippocampus-th50	0.82053	5.57E-163	TRUE

References

Anderson, A. K., & Phelps, E. A. (2002). Is the human amygdala critical for the subjective experience of emotion? Evidence of intact dispositional affect in patients with amygdala lesions. *Journal of Cognitive Neuroscience*, *14*(5), 709–720. https://doi.org/10.1162/08989290260138618

Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, *20*(2), 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7

Ashar, Y. K., Andrews-Hanna, J. R., Dimidjian, S., & Wager, T. D. (2017). Empathic Care and Distress: Predictive Brain Markers and Dissociable Brain Systems. *Neuron*, *94*(6), 1263-1273.e4. https://doi.org/10.1016/j.neuron.2017.05.014

Berg, H., Ma, Y., Rueter, A., Kaczkurkin, A., Burton, P. C., DeYoung, C. G., MacDonald, A. W., Sponheim, S. R., & Lissek, S. M. (2021). Salience and central executive networks track overgeneralization of conditioned-fear in post-traumatic stress disorder. *Psychological Medicine*, *51*(15), 2610–2619. https://doi.org/10.1017/S0033291720001166

Bremner, J. D., Vermetten, E., Schmahl, C., Vaccarino, V., Vythilingam, M., Afzal, N., Grillon, C., & Charney, D. S. (2005). Positron emission tomographic imaging of neural correlates of a fear acquisition and extinction paradigm in women with childhood sexual-abuse-related post-traumatic stress disorder. *Psychological Medicine*, *35*(6), 791–806. https://doi.org/10.1017/S0033291704003290

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox. *NeuroImage*, *16*(2), 497.

Brewin, C. R., Gregory, J. D., Lipton, M., & Burgess, N. (2010). Intrusive Images in Psychological Disorders: Characteristics, Neural Mechanisms, and Treatment Implications. *Psychological Review*, *117*(1), 210–232. https://doi.org/10.1037/a0018113

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Brown, T., & Barlow, D. (2014). *Anxiety and Related Disorders Interview Schedule for DSM-5 (ADIS-5)® - Adult Version: Client Interview Schedule 5-Copy Set*. Oxford University Press.

Campbell-Sills, L., Norman, S. B., Craske, M. G., Sullivan, G., Lang, A. J., Chavira, D. A., Bystritsky, A., Sherbourne, C., Roy-Byrne, P., & Stein, M. B. (2009). Validation of a Brief Measure of Anxiety-Related Severity and Impairment: The Overall Anxiety Severity and Impairment Scale (OASIS). *Journal of Affective Disorders*, *112*(1–3), 92–101. https://doi.org/10.1016/j.jad.2008.03.014

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLOS Biology*, *13*(6), e1002180. https://doi.org/10.1371/journal.pbio.1002180

Chiba, T., Kanazawa, T., Koizumi, A., Ide, K., Taschereau-Dumouchel, V., Boku, S., Hishimoto, A., Shirakawa, M., Sora, I., Lau, H., Yoneda, H., & Kawato, M. (2019). Current Status of Neurofeedback for Post-traumatic Stress Disorder: A Systematic Review and the Possibility of Decoded Neurofeedback. *Frontiers in Human Neuroscience*, *13*(233). https://doi.org/10.3389/fnhum.2019.00233

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4–5), 171–178. https://doi.org/10.1002/(sici)1099-1492(199706/08)10:4/5<171::aid-nbm453>3.0.co;2-l

Craske, M. G., Stein, M. B., Eley, T. C., Milad, M. R., Holmes, A., Rapee, R. M., & Wittchen, H.-

U. (2017). Anxiety disorders. *Nature Reviews Disease Primers*, *3*(1), 1–19. https://doi.org/10.1038/nrdp.2017.24

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Dunsmoor, J. E., & Paz, R. (2015). Fear Generalization and Anxiety: Behavioral and Neural Mechanisms. *Biological Psychiatry*, *78*(5), 336–343. https://doi.org/10.1016/j.biopsych.2015.04.010

Eisenbarth, H., Chang, L. J., & Wager, T. D. (2016). Multivariate Brain Prediction of Heart Rate and Skin Conductance Responses to Social Threat. *Journal of Neuroscience*, *36*(47), 11987–11998. https://doi.org/10.1523/JNEUROSCI.3672-15.2016

Feinstein, J. S., Buzza, C., Hurlemann, R., Follmer, R. L., Dahdaleh, N. S., Coryell, W. H., Welsh, M. J., Tranel, D., & Wemmie, J. A. (2013). Fear and panic in humans with bilateral amygdala damage. *Nature Neuroscience*, *16*(3), 270–272. https://doi.org/10.1038/nn.3323

Fenster, R. J., Lebois, L. A. M., Ressler, K. J., & Suh, J. (2018). Brain circuit dysfunction in post-traumatic stress disorder: From mouse to man. *Nature Reviews Neuroscience*, *19*(9), 535–551. https://doi.org/10.1038/s41583-018-0039-7

Fischl, B., Salat, D. H., Van Der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*(SUPPL. 1), S69–S84.

https://doi.org/10.1016/j.neuroimage.2004.07.016

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*, *21*(4), 500–508. https://doi.org/10.1038/mp.2015.88

Garcia, R., Vouimba, R.-M., Baudry, M., & Thompson, R. F. (1999). The amygdala modulates prefrontal cortex activity relative to conditioned fear. *Nature*, *402*(6759), 294–296. https://doi.org/10.1038/46286

Garfinkel, S. N., Abelson, J. L., King, A. P., Sripada, R. K., Wang, X., Gaines, L. M., & Liberzon, I. (2014). Impaired Contextual Modulation of Memories in PTSD: An fMRI and Psychophysiological Study of Extinction Retention and Fear Renewal. *Journal of Neuroscience*, *34*(40), 13435–13443. https://doi.org/10.1523/JNEUROSCI.4287-13.2014

Giustino, T. F., & Maren, S. (2015). The Role of the Medial Prefrontal Cortex in the Conditioning and Extinction of Fear. *Frontiers in Behavioral Neuroscience*, *9*. https://www.frontiersin.org/article/10.3389/fnbeh.2015.00298

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Hanke, M., Halchenko, Y., Sederberg, P., Olivetti, E., Fründ, I., Rieger, J., Herrmann, C., Haxby, J., Hanson, S., & Pollmann, S. (2009). PyMVPA: A unifying approach to the analysis of neuroscientific data. *Frontiers in Neuroinformatics*, *3*. https://www.frontiersin.org/article/10.3389/neuro.11.003.2009

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*. https://doi.org/10.1016/j.neuron.2011.08.026

Hennings, A. C., McClay, M., Lewis-Peacock, J. A., & Dunsmoor, J. E. (2020). Contextual reinstatement promotes extinction generalization in healthy adults but not PTSD. *Neuropsychologia*, *147*, 107573. https://doi.org/10.1016/j.neuropsychologia.2020.107573

Im, H. Y., Adams, R. B., Boshyan, J., Ward, N., Cushing, C. A., & Kveraga, K. (2017). Observer’s anxiety facilitates magnocellular processing of clear facial threat cues, but impairs parvocellular processing of ambiguous facial threat cues. *Scientific Reports*, *7*(1), 15151–15151. https://doi.org/10.1038/s41598-017-15495-2

Insel, T. R. (2019). Bending the Curve for Mental Health: Technology for a Public Health Approach. *American Journal of Public Health*, *109*(S3), S168–S170. https://doi.org/10.2105/AJPH.2019.305077

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/s1361-8415(01)00036-6

Koizumi, A., Amano, K., Cortese, A., Shibata, K., Yoshida, W., Seymour, B., Kawato, M., & Lau, H. (2017). Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-016-0006

Krishnapuram, B., Carin, L., Figueiredo, M. A. T., & Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(6), 957–968. https://doi.org/10.1109/TPAMI.2005.127

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal*

Medicine, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lang, P. J., Miller, G. A., & Levin, D. N. (1983). Anxiety and Fear. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation: Volume 3: Advances in Research and Theory* (pp. 123–151). Springer US. https://doi.org/10.1007/978-1-4615-9317-1_4

Lange, I., Goossens, L., Bakker, J., Michielse, S., Marcelis, M., Wichers, M., van Os, J., van Amelsvoort, T., & Schruers, K. (2019). Functional neuroimaging of associative learning and generalization in specific phobia. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *89*, 275–285. https://doi.org/10.1016/j.pnpbp.2018.09.008

Lau, H. (2022). *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford University Press.

LeDoux, J. E., & Hofmann, S. G. (2018). The subjective experience of emotion: A fearful view. *Current Opinion in Behavioral Sciences*, *19*, 67–72. https://doi.org/10.1016/j.cobeha.2017.09.011

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J., & Craske, M. G. (2015). Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clinical Psychology Review*, *42*, 72–82. https://doi.org/10.1016/j.cpr.2015.08.004

Lutkenhoff, E. S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J. D., Owen, A. M., & Monti, M. M. (2014). Optimized Brain Extraction for Pathological Brains (optiBET). *PLOS ONE*, *9*(12), e115551. https://doi.org/10.1371/journal.pone.0115551

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Marin, M.-F., Zsido, R. G., Song, H., Lasko, N. B., Killgore, W. D. S., Rauch, S. L., Simon, N. M., & Milad, M. R. (2017). Skin Conductance Responses and Neural Activations During Fear Conditioning and Extinction Recall Across Anxiety Disorders. *JAMA Psychiatry*, *74*(6), 622–631. https://doi.org/10.1001/jamapsychiatry.2017.0329

Milad, M. R., Pitman, R. K., Ellis, C. B., Gold, A. L., Shin, L. M., Lasko, N. B., Zeidan, M. A., Handwerger, K., Orr, S. P., & Rauch, S. L. (2009). Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biological Psychiatry*, *66*(12), 1075–1082. https://doi.org/10.1016/j.biopsych.2009.06.026

Parkes, L., Fulcher, B., Yücel, M., & Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, *171*, 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, *12*(null), 2825–2830.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction Learning in Humans: Role of the Amygdala and vmPFC. *Neuron*, *43*(6), 897–905. https://doi.org/10.1016/j.neuron.2004.08.042

Phillips, R. G., & LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience*, *106*(2), 274–285. https://doi.org/10.1037/0735-7044.106.2.274

Picó-Pérez, M., Alemany-Navarro, M., Dunsmoor, J. E., Radua, J., Albajes-Eizagirre, A., Vervliet, B., Cardoner, N., Benet, O., Harrison, B. J., Soriano-Mas, C., & Fullana, M. A. (2019). Common and distinct neural correlates of fear extinction and cognitive reappraisal: A meta-analysis of fMRI studies. *Neuroscience & Biobehavioral Reviews*, *104*, 102–115. https://doi.org/10.1016/j.neubiorev.2019.06.029

Pittig, A., Treanor, M., LeBeau, R. T., & Craske, M. G. (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research. *Neuroscience and Biobehavioral Reviews*, *88*, 117–140.

https://doi.org/10.1016/j.neubiorev.2018.03.015

Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, *112*, 267–277. https://doi.org/10.1016/j.neuroimage.2015.02.064

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, *463*(7277), 49–53. https://doi.org/10.1038/nature08637

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23 Suppl 1*, S208-219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Sporns, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience*, *17*(5), 652–660. https://doi.org/10.1038/nn.3690

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(13), 3470–3475. https://doi.org/10.1073/pnas.1721572115

Taschereau-Dumouchel, V., Kawato, M., & Lau, H. (2019). Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0520-3

Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G., & LeDoux, J. E. (2022). Putting the “mental” back in “mental disorders”: A perspective from research on fear and anxiety. *Molecular Psychiatry*, 1–9. https://doi.org/10.1038/s41380-021-01395-5

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, *35*(1), 105–120. https://doi.org/10.1016/j.neuroimage.2006.11.054

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. a., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: Judgements from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006.The

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

Turner, B. O., Mumford, J. A., Poldrack, R. A., & Ashby, F. G. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage*, *62*(3), 1429–1438. https://doi.org/10.1016/j.neuroimage.2012.05.057

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1), 1–10. https://doi.org/10.1038/s42003-018-0073-z

Valente, G., Castellanos, A. L., Hausfeld, L., De Martino, F., & Formisano, E. (2021). Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. *NeuroImage*, *238*, 118145. https://doi.org/10.1016/j.neuroimage.2021.118145

Visser, R. M., Bathelt, J., Scholte, H. S., & Kindt, M. (2021). Robust BOLD Responses to Faces

But Not to Conditioned Threat: Challenging the Amygdala's Reputation in Human Fear and Extinction Learning. *Journal of Neuroscience*, *41*(50), 10278–10292. https://doi.org/10.1523/JNEUROSCI.0857-21.2021

Webb, T. W., Igelström, K. M., Schurger, A., & Graziano, M. S. A. (2016). Cortical networks involved in visual awareness independent of visual attention. *Proceedings of the National Academy of Sciences*, *113*(48), 13923–13928. https://doi.org/10.1073/pnas.1611505113

Wen, Z., Chen, Z. S., & Milad, M. R. (2021). Fear extinction learning modulates large-scale brain connectivity. *NeuroImage*, *238*, 118261. https://doi.org/10.1016/j.neuroimage.2021.118261

Whiteley, C. M. K. (2021). Depression as a Disorder of Consciousness. *The British Journal for the Philosophy of Science*, 716838. https://doi.org/10.1086/716838

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

Young, K. S., Bookheimer, S. Y., Nusslock, R., Zinbarg, R. E., Damme, K. S. F., Chat, I. K.-Y., Kelley, N. J., Vinograd, M., Perez, M., Chen, K., Cohen, A. E., & Craske, M. G. (2021). Dysregulation of threat neurociruitry during fear extinction: The role of anhedonia. *Neuropsychopharmacology*, *46*(9), 1650–1657. https://doi.org/10.1038/s41386-021-01003-8

Zayfert, C., DeViva, J. C., Becker, C. B., Pike, J. L., Gillock, K. L., & Hayes, S. A. (2005). Exposure utilization and completion of cognitive behavioral therapy for PTSD in a "real world" clinical practice. *Journal of Traumatic Stress*, *18*(6), 637–645. https://doi.org/10.1002/jts.20072

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE*

Transactions on Medical Imaging, *20*(1), 45–57. https://doi.org/10.1109/42.906424

Zhou, F., Zhao, W., Qi, Z., Geng, Y., Yao, S., Kendrick, K. M., Wager, T. D., & Becker, B. (2021). A distributed fMRI-based signature for the subjective experience of fear. *Nature Communications*, *12*(1), 6643. https://doi.org/10.1038/s41467-021-26977-3

Zidda, F., Andoh, J., Pohlack, S., Winkelmann, T., Dinu-Biringer, R., Cavalli, J., Ruttorf, M., Nees, F., & Flor, H. (2018). Default mode network connectivity of fear- and anxiety-related cue and context conditioning. *NeuroImage*, *165*, 190–199. https://doi.org/10.1016/j.neuroimage.2017.10.024