Isolating and manipulating competing neural representations of fear and safety

I. Project science areas: 3 CTR; 7 NS

II. Project Description:

The goal of this proposal is to identify, isolate, and causally manipulate patterns of neural activity associated with fear or safety in the human brain. This clinical translational neuroscience research project may lead to discovering new approaches to innovate neuroscience-based treatments for mental health disorders characterized by excessive fear and anxiety. Anxiety Disorders and Posttraumatic Stress Disorder (PTSD) have a combined prevalence estimate of ~ 22% of the US adult population. Emerging research is detailing brain regions and functional networks affected across affective disorders [1]. This research reveals abnormalities in the amygdala, hippocampus, and ventromedial prefrontal cortex (vmPFC), regions crucial for learning to fear and learning to overcome fear [2,3]. Accordingly, translational neuroscience research has characterized brain abnormalities associated with deficits in fear extinction (i.e., learning that feared stimuli, environments, or activities are not threatening) across a number of psychiatric conditions [4]. Overall, experimental models of fear and extinction are invaluable for conceptualizing the etiology, maintenance, and relapse of a variety of psychiatric conditions [4,5], and provide a significant avenue to make advances in clinical translational neuroscience.

However, despite advancements in the neuroscience of fear and extinction, there are crucial gaps (one might even say chasms) in our understanding of how the human brain encodes, organizes, and retrieves these competing emotional memories. Specifically, it is extremely challenging to experimentally dissociate the encoding and retrieval of fear versus extinction memories in the brain. Although it is appreciated that contextual factors mediate expression of fear versus extinction [6], extinction memory retrieval is typically inferred by the *absence* of a behavioral fear response. Reliance on the absence of a behavior to infer the strength of a memory is problematic on scientific grounds. As the traditional aphorism goes, “absence of evidence is not evidence of absence.” Pioneering technical advances in fine-scale molecular imaging and activity-dependent neural tagging is beginning to reveal separate and quantifiable memory traces of fear versus extinction in the rodent brain. This work reveals extinction is an active process activating separate neural populations with distinct subcortical-cortical pathways between the amygdala, hippocampus, and medial PFC [7]. Leading-edge neurophysiological research confirms early theories—dating back to the time of Pavlov—that extinction is a form of new learning, and not unlearning. Invasive neurophysiology also confirms context-dependent retrieval-based accounts of extinction; that is, the balance of activity in neural populations shifts according to the context at test [7].

This proposal adapts theoretical and technical insights from rodent neurophysiology to substantively advance methodological and theoretical approaches to the cognitive neuroscience of fear and extinction in humans. A precise methodology for tagging specific neural populations in the human brain using non-invasive neuroimaging is not feasible with current technology. Therefore, I have devised a way to leverage cutting-edge advances in machine learning and multivariate neural pattern similarity analyses of high-resolution functional MRI data to localize spatially distributed patterns of activity unique to the encoding and retrieval of fear and extinction in the human brain. Once a multivariate pattern is isolated, we will use decoded neurofeedback so that participants can self-modulate patterns of neural activity to enhance safety memory retrieval and diminish fear.

The knowledge gained from this ambitious research project could provide an objective neural target to make psychiatric treatment more effective, persistent, and generalizable. Thus, there is strong potential for this work to have broad implications for mental health research, particularly disorders characterized by the inability to retrieve memories that compete against maladaptive associations. This project includes healthy adults and individuals with PTSD. I focus on PTSD because linking a multivariate signature of extinction memory retrieval to the neuropathophysiology of PTSD can have direct benefit to exposure therapy—the gold-standard treatment based on the principles of extinction. Exposure therapy for PTSD can be effective, but 50% of treated patients will not achieve sustained remission, and 1/3 will drop-out of treatment. Future transdiagnostic work will target a variety of psychiatric diseases explained under a computational psychiatry framework that emphasizes reinforcement learning and Pavlovian conditioning models.

In the sections below, I will highlight the methodological and theoretical innovation that are the core purpose of this grant. I will then discuss the innovativeness of this proposal, which touches on the ultimate scientific and mental health impact of this work. Finally, I will describe my qualifications as a candidate to lead this high-risk, high-reward line of research, which ultimately seeks to improve mental health in affected populations.

<u>Theoretical background and methodological innovations:</u> I propose to localize and causally manipulate distributed patterns of neural activity across distributed brain networks during the encoding and retrieval of fear and extinction by implementing cutting-edge multivariate pattern analyses and closed-loop fMRI decoded neurofeedback. Since the earliest laboratory research of conditioned learning, it has been widely appreciated that extinction of a learned behavior constitutes a form of new learning, and is not the unlearning or forgetting of a previous association. Learning theory proposes—and research confirms—that extinction is contextually-specific, and that extinguished behavior returns under a variety of circumstances [6]. Recent leading-edge neurophysiology research in rodents reveals collections of neural populations in the amygdala and hippocampus active at encoding and retrieval of separate behavioral states. ***Once these neural populations are isolated in the rodent brain, they can be manipulated to enhance or diminish threat expression*** [8]. However, *it is unknown whether memories of fear and extinction can be dissociated in the human brain and whether these neural representations can be* ***causally manipulated*** *to enhance or diminish fear*. Here, we seek to isolate these memories and to manipulate their neural signatures to impact behavioral expression of fear and safety, which could ultimately be used to innovate treatment for psychiatric disease.

<u>Neurophysiological evidence for separate representations of fear and extinction memory in the rodent brain:</u> Identifying quantifiable memory traces in the brain is challenging because memory representations are widely distributed within and across discrete brain regions, memories change over time, and not all experiences induce persistent changes in the brain. Fear conditioning has proved an indispensable model to answer questions on the nature of learning and memory representations in the brain; it is rapid, strong, stable, and has objective neural and behavioral correlates conserved across species [9]. Further, learning models based on the principles of conditioning provide explanatory power for characterizing a range of psychiatric disorders [4], and has been especially beneficial to understanding the neuropathophysiology of PTSD [10]. One of the most important recent discoveries on the neuroscience of associative learning has been the localization of specific cell types in rodents that appear selective for either conditioning or extinction [7,8] (**Fig. 1, left side**). Neural tagging in transgenic mice show populations of neurons in the amygdala that are active during fear memory formation are reactivated during fear memory retrieval. These "fear neurons" selectively respond to a fear conditioned stimulus (CS). Another set of neurons were identified as specific to the encoding and retrieval of fear extinction. Fear and extinction neurons can be further distinguished from one another by distinct projections to the medial prefrontal cortex and ventral hippocampus. This suggests a unique modulatory role between memory traces that mediate the balance between expression of high and low fear states. In short, fear and extinction appear to be distinctly organized and leave measurable traces in the rodent brain. *But it is unknown whether fear and extinction memories can be dissociated in the human brain (schematized in* ***Fig 1, right side****), and how interactions between critical brain regions orchestrate the balance between retrieval of a fear memory versus an extinction memory.*

<u>Existing techniques for isolating representations of fear and extinction in the human brain:</u> Functional MRI offers the ability to broadly translate neuroscience research from rodents to humans [11]. The most common method has been univariate voxel-wise subtraction methods of fMRI activity. Univariate fMRI detects voxels that show a maximal response on a given set of trials, often reflected in averaged activity within a region-of-interest that has been spatially smoothed across voxels to improve signal-to-noise. These analyses are optimal for detecting areas of the brain that show robust activity across a large collection of voxels that mostly respond in the same way. Oftentimes, activity to one condition is contrasted with (subtracted from) activity to another condition. In

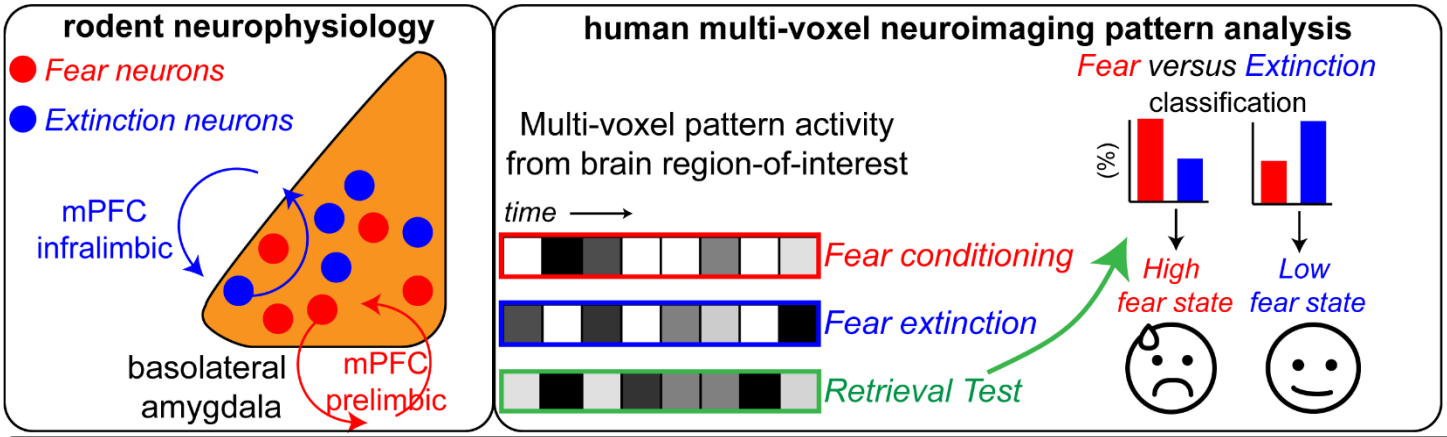

Figure 1 Cartoon of intermingled fear and extinction neurons in amygdala using rodent neurophysiology, and a possible approach to using multi-voxel pattern analysis to separate fear and extinction memories using advanced neuroimaging methods in humans

human neuroimaging of fear and extinction, this approach shows activity in a collection of brain regions that exhibit stronger activity to a conditioned stimulus (CS+) that predicts an aversive outcome as compared to a control stimulus (CS-) that is never paired with the aversive outcome. This work consistently shows stronger activity to the CS+ than CS- in the insula, anterior cingulate cortex (ACC), striatum, sensory cortex, and thalamus extending into the midbrain and brainstem. The ventromedial PFC (vmPFC) is considered homologous to the rodent infralimbic cortex, and is associated with learning, storing, and retrieving extinction memories.

But mean univariate subtraction methods ignore information coded at finer spatial resolutions and distributed across collections of voxels. This limitation is especially relevant when trying to translate animal neurobiology research of learning and memory to human neuroimaging. For instance, despite clear evidence that the amygdala is crucial for fear conditioning and the vmPFC is crucial for extinction, fMRI evidence of amygdala activity during conditioning and vmPFC activity during extinction is inconsistent at best. In fact, a recent meta-analysis of fear conditioning fMRI studies with 677 participants across 27 independent studies pointedly showed that the amygdala is *not* among the collection of areas consistently activated during conditioning [12]. Studies reporting amygdala activation often do so using fairly liberal statistical thresholds that do not control for multiple comparisons across the brain. In the PI's own fMRI research, we have observed amygdala activity using the typical contrast of CS+ greater than CS- in less than half of our studies. One explanation for a lack of observable amygdala activity during human fear conditioning comes from neurophysiology, which shows sparsely distributed neurons respond to the unpaired safety stimulus (CS-) [7]. Likewise, fMRI meta-analyses show that the vmPFC is rarely identified in human fear extinction, despite abundant neurophysiological evidence of its role in extinction.

How multivariate brain imaging techniques can be used to dissociate conditioning and extinction processes: Although neuroimaging technology does not exist to isolate individual neurons in the human brain, recent developments in high-resolution multivoxel pattern analysis (MVPA) make it possible to non-invasively measure how representational content is encoded within and across brain regions in fine detail [13]. Advances in machine learning of neuroimaging analysis has revealed not only which brain areas are engaged by a task, but also how different types of information are processed at a finer-grained spatial resolution within the same brain region. This technique is superior to the common univariate approach for detecting differences in neural activity between subtle experimental manipulations, and have led to considerable breakthroughs in cognitive neuroscience. My proposal is to push the boundaries of theoretical models and MVPA techniques honed in the cognitive neuroscience of human memory to decode memories specific to the encoding of fear versus extinction. My plan is built on two advances on computational models of memory. First, neurocognitive processes active at memory formation are reactivated during retrieval [14]. Second, information is linked to the context where it was encoded [15]. **I will apply machine learning as well as pattern similarity analysis of fMRI data to decode information content unique to competing experiences of fear and extinction in the human brain, and then manipulate these patterns using closed-loop multivoxel decoded neurofeedback.**

Contextual specificity of extinction: Extinction renders the meaning of the CS ambiguous; the same stimulus can now signal threat and the absence of threat. The brain resolves threat ambiguity based on context, a process mediated by the hippocampus that may involve feedforward inhibitory projections from the ventral hippocampus to the infralimbic cortex. Because extinction is contextually specific, fear suppressed in the extinction context is often expressed in a different context, an effect known as "renewal." Renewal can explain why clinical treatments often fail to generalize beyond the therapeutic environment. There are numerous parallels to the role context plays in fear extinction and in human episodic memory. For example, a classic finding is that memory is worse when the room changes between study and test. Importantly, "context" also includes time, mood, internal conditions, and other types of mental states. Like the physical context, similarity between the "mental context" of encoding and retrieval helps guide retrieval. Moreover, reinstating a "mental context" might counteract the deficit usually caused by a change in the *physical* context between learning and test. **I will evaluate whether mental reactivation of the extinction context prevents fear renewal by synthesizing theoretical and empirical work on contextual reactivation of *episodic* memory with theoretical and empirical work on contextual specificity of extinction.**

How to generate, track, and causally manipulate a "mental context" using fMRI: Memories are comprised of numerous details experienced at the time of encoding. Integration of these details provides a "mental context representation" that can later guide memory retrieval. A computational model of memory formalizes how information is bound to a gradually accumulating contextual representation, known as the Temporal Context Model [15]. This model was initially developed to explain a number of memory recall phenomena; for instance, why

remembering an item from a list facilities recall for neighboring items on that list [16]. Multi-voxel neuroimaging experiments can cleverly incorporate this model to decode brain activity related to the retrieval of items that had been encoded in different contexts. **I will combine theoretical principles of episodic memory recall with associative learning models to derive and causally manipulate a novel measure of competitive reactivation of either fear or safety memories in a novel context.** Evidence from my R00 project suggests that the strength of neural context reinstatement orchestrates initiations of a high or low fear state in other brain regions (**Fig. 4**), helping guide emotional responses in a novel context when threat is ambiguous. The next and highly ambitious stage of this work, proposed here, is to *causally manipulate* this neural signature using closed-loop fMRI neurofeedback. This research will involve healthy adults and patients with PTSD. Extinction-retention deficits in PTSD are well-documented and may be a factor in relapse after exposure treatment.

Real-time closed-loop multivariate decoded fMRI neurofeedback: A technique to enhance the strength and generalizability of an extinction memory? One inherent limitation to fMRI methods is that activity might be *related* to a cognitive operation, but is not *necessary* for the behavioral expression of that cognitive operation. Neurofeedback is a form of brain-computer interface that can be used to uncover the causal role of fMRI-related activations. Recent technical innovations in ***real-time MVPA*** decoding has advanced the potential for fMRI neurofeedback by allowing closed-loop training activation on distributed patterns of activity, rather than on averaged fMRI signal from a single region of interest, referred to as decoded neurofeedback or DecNef (**Fig. 2**). The goal of neurofeedback is to modify a particular behavior through self-modulation of its underlying neural substrates. Neurofeedback training requires voluntary self-regulation of a neural circuit based on feedback of its activation. It has been used to study brain-behavior relationships for cognitive processes such as visual perception and attentional control. **Here, I hypothesize that increasing multivariate signals linked to either a fear or extinction context will boost the strength and generalizability of either fear or extinction memories, respectively.** My lab's ability to decode the multivariate neural signature of a mental context has been demonstrated in an R00 funded project, and the infrastructure needed to perform MVPA neurofeedback training is in place and operational at the University of Texas at Austin. *This research is important for drawing causal inferences on the role of brain activity related to fear and safety memory retrieval.*

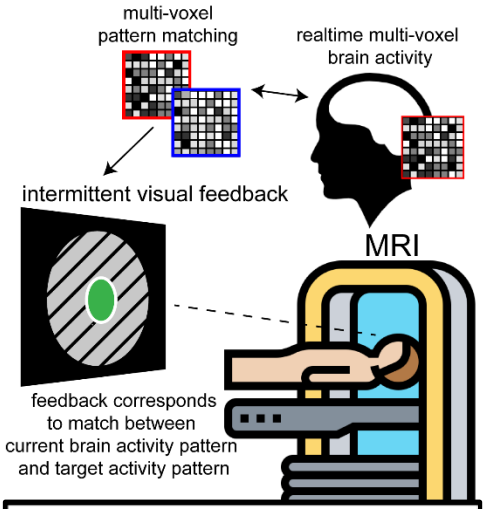

Figure 2 Example of closed-loop multi-voxel decoded neurofeedback. Neural activity is decoded based on a target pattern and presented as visual feedback.

Research strategy: Although the widely held view is that the extinction memory is encoded in a new memory trace separate from that of the fear memory, evidence for this in rodents is just emerging and evidence for this in humans is lacking entirely. There is a major issue inherent to typical experimental extinction protocols that make it challenging to validate whether behavioral expression corresponds to the specific retrieval of an extinction memory, per se. Foremost, extinction memory retrieval is typically assayed via absence of a defensive behavioral response, such as freezing in rats or sweating in humans. In experimental extinction preparations, there are a number of alternative possibilities for why a single response system (e.g., freezing) is, or is not, expressed at test that are unrelated to the retrieval of an extinction memory, per se. Likewise, activity in the vmPFC during extinction learning and extinction-recall in human fMRI has been taken as evidence of an extinction process; but this is a reasoned hypothesis based on animal neurophysiology. *Currently, there is not an unambiguous method to dissociate fear and extinction memories in humans.* My lab has developed a technique, based on theoretical and technical insights from animal neurophysiology and cognitive neuroscience of human memory, to identify and isolate separate memories of fear and extinction in the human brain. This approach rests on two fundamental features of memory: (**1**) neurocognitive processes active at memory formation are reactivated during retrieval, and (**2**) information is linked to the **context** where it was encoded.

The 'encoding specificity principle' states that memory retrieval is facilitated when external or internal conditions at the time of memory retrieval match those from the time of memory formation. Numerous findings across diverse psychological disciplines support the role of context on memory retrieval. In associative learning research, this principle explains the contextual-dependence of extinction. More recently, computational models of episodic memory have formalized how mental context retrieval guides temporal order memory [15]. These models have been leveraged on multivariate fMRI data to decode the mental context where an item was

previously encoded. **The goal of this study is to use neurofeedback to induce matching between the encoding and retrieval of an extinction memory, thereby helping promote retrieval of safety over fear.** Accordingly, I will synthesize theoretical and empirical advances on how memories are linked to the mental context where they were formed in order to derive a neural signature of emotional memory reactivation. ***I will then manipulate this mnemonic pattern using closed-loop decoded neurofeedback with the ultimate goal of enhancing retrieval of the extinction memory and diminish expression of the fear memory.***

Basic research design features: To leverage mechanistic insights on the neuroscience of conditioning and extinction from animal models, we utilize a Pavlovian conditioning framework. I am an early stage independent investigator who has already published over 40 behavioral and neuroimaging studies of human fear conditioning. In the basic protocol, subjects first learn that a neutral CS predicts an aversive unconditioned stimulus (US, shock). This CS-US association generates an anticipatory conditioned response (CR), which we measure by increases in sympathetic arousal (i.e., skin conductance responses, SCR). The amygdala is crucial for learning the CS-US association and expressing the CR [2], but it is often unobserved in fMRI studies of human fear conditioning [12]. Extinction involves repeated presentations of the CS without the US. The vmPFC is important for the learning and consolidation of extinction memories. But there is still much we do not know about the neural circuitry of extinction, especially in humans. In addition to the vmPFC, extinction appears to involve the amygdala, as well as the hippocampus for contextual modulation of extinction memory retrieval. Each study employs a discrimination, partial reinforcement, delay fear conditioning design, with a CS+ that co-terminates with a mild electrical shock to the right wrist (US), and a CS- that serves as a within-subject unpaired control stimulus that is never paired with the shock. Each trial is 6 second duration with an inter-trial interval of 10±2 seconds.

Participants and Power Analysis: In total, we plan to include a total of 150 18-40 y/o adult subjects matched for age, sex, race, and IQ. This includes 25 healthy and 25 PTSD subjects per each of the 3 conditions as detailed below. A power analysis (SPSS Sample Power 2, IBM Corp) on fMRI and behavioral data from the R00 research estimates a sample size of 24 per group to yield 80% power to detect successful fear acquisition in fear-related regions (2-sample t-test of beta values from the dorsal ACC and insula). MVPA classification accuracy in this sample size is also well above chance, as shown in **Fig. 4A**. Using data from the R00, this sample size also yields sufficient power to detect MVPA classifier evidence for neural reinstatement of the extinction context that is foundational to this proposal (**Fig. 4B**). A sample size of 25 is considered sufficiently powered for extant neurofeedback research [17].

PTSD characteristics: Further details on the PTSD population are in the Protection of Human Subjects section. In brief, the goal of this project is to characterize differences in how fear and extinction memories are formed, stored, and retrieved in the healthy brain and in disease. While PTSD is characterized by a heterogeneity of symptom profiles, a PTSD sample is a valid clinical model for understanding disordered fear and extinction learning, as well as impaired contextual modulation of extinction. We include patients who meet DSM-5 diagnostic criteria for current PTSD as assessed using the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders (SCID), and the Clinician-Administered PTSD Scale for DSM-5 (CAPS), the gold-standard diagnostic interviews for determining current PTSD diagnosis and symptom severity. Participants will be eligible if they meet criteria for current PTSD, as determined by the presence of a Criterion A event in addition to a severity score of 2 or greater on 1 symptom in clusters B and C and on 2 symptoms in clusters D and E, in addition to meeting criteria F and G. Dr. Charles Nemeroff (collaborator) is a leading expert in PTSD research. While an RDoC approach would consider a transdiagnostic sample, it is possible (perhaps even likely) that the heterogeneity of different patient groups (e.g., PTSD, OCD, PD, GAD) would make it challenging to characterize participants with different symptoms and symptom severity, among other factors. However, we do not exclude for comorbid Anxiety or Depression (but do exclude for bipolar and current SUD).

***Detailed research design features*:** My lab has developed a protocol to separately tag the context of fear learning versus fear extinction during fMRI (**Fig. 3**), and then decode the relative strength of context-dependent neural reactivation at test. Subjects undergo fear conditioning and extinction on Day 1. The CSs are two auditory tones—500 Hz and 1000 Hz—counterbalanced between subjects as CS+ and CS- and presented through MRI compatible headphones. The key feature of this paradigm is the contextual manipulation on Day 1. Throughout conditioning and extinction, subjects are presented with a series of task-irrelevant pictures of animals and tools (counterbalanced between subjects). *These category-specific pictures build the "context."* Animal-related or tool-related neural activity [18] should therefore be assimilated into a mental context representation specific to the formation of fear or extinction memory. The rationale for presenting a stream of pictures throughout the scan is

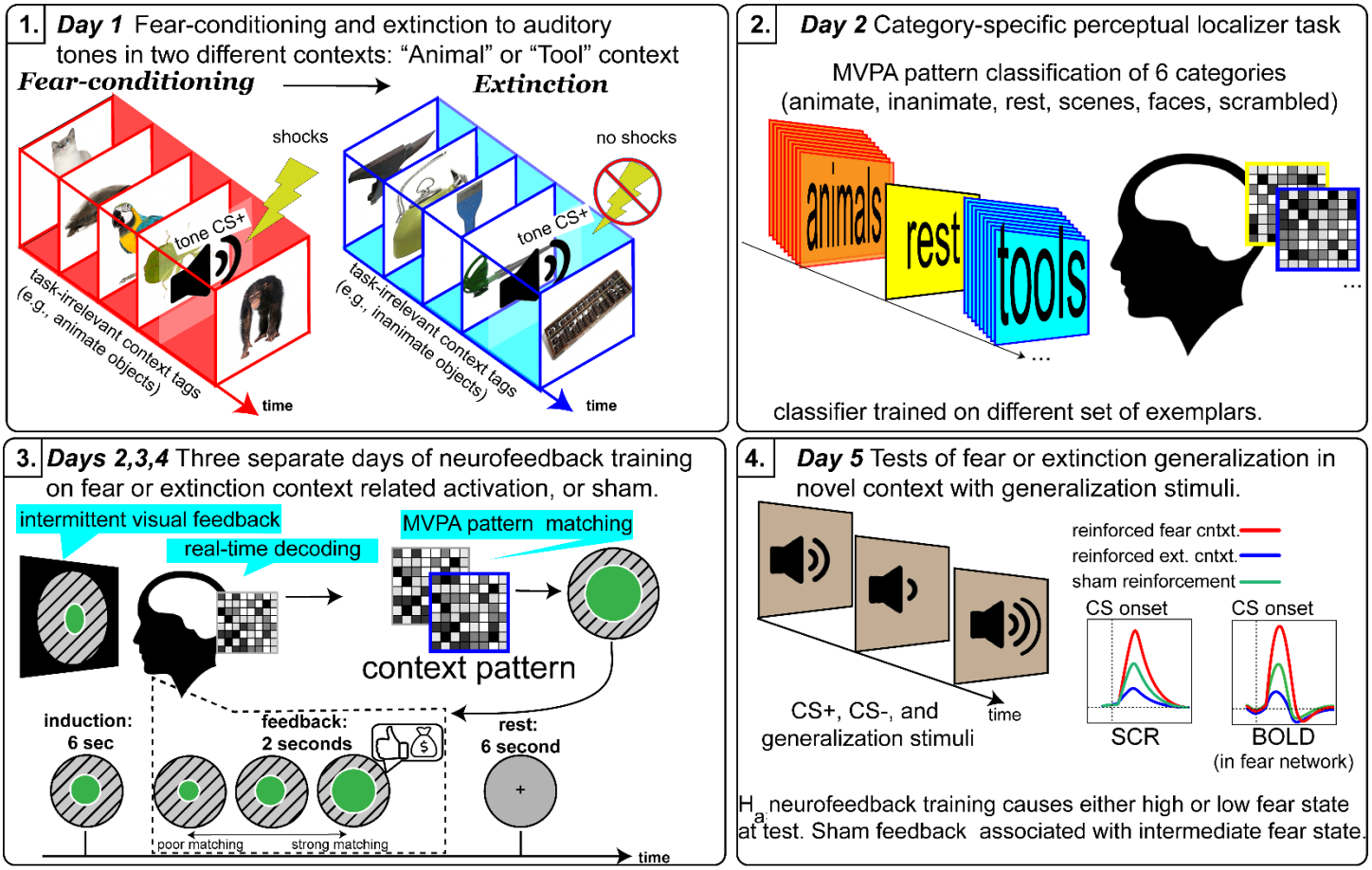

Figure 3 Subjects undergo conditioning and extinction to a tone CS. Pictures of animals and tools are used to "tag" either the conditioning or extinction context, counterbalanced. Over three sessions, real-time fMRI neurofeedback based on MVPA pattern classifiers of the fear or extinction context (or sham) are used to manipulate the strength of contextual representations. The effect of the training sessions is evaluated on the 5th day with tests of contextual renewal and stimulus generalization to novel tones.

to ensure that the context information is consistently activating visual cortex. This should produce a robust "mental context tag" that we can decode later using fMRI pattern classification for the target brain activity pattern during real-time decoded neurofeedback. Subjects will undergo a series of functional localizer scans on Day 2 in order to train pattern classifiers to recognize multivariate activity patterns unique to animate versus tool activity patterns. Additional categories (e.g., scenes, scrambled scenes, and rest) are used in the category localizer to fine tune the precision of the trained classifiers.

Real-time closed-loop multivariate decoded neurofeedback: To better understand how people can manipulate fMRI activity to strengthen memories of fear or extinction, we will use non-invasive closed-loop fMRI neurofeedback to help participants endogenously control neural activation levels **(Fig. 2, 3)**. The goal is for subjects to self-modulate neural activity patterns associated with a particular learning context (fear or extinction), thereby strengthening the likelihood of reactivating that context, and increasing the likelihood of retrieving and reinforcing the memories encoded in that context. Participants receive intermittent neurofeedback (once per trial) in the form of an expanding circle. The subjects' task is to maximize the size of the circle during a 6-sec induction period on each trial. The circumference will reflect the amount of context reinstatement for the tagged mental context (i.e., the amount of activation in visual cortex reflecting the fear or extinction context) over the preceding 6-s of brain activity (accounting for the hemodynamic lag). For two experimental groups, the size of the circle will be configured to increase when scene activation in occipitotemporal cortex is high, and subjects will receive feedback of their success that is tied to monetary reward that they will receive at the end of the session. For one of those groups, the conditioning context (e.g., decoded patterns reflecting animal-related activity) will be reinforced; for the other group, the extinction context (e.g., decoded patterns reflecting tool-related activity) will be reinforced. The third group is a control group composed of subjects for whom neurofeedback is not related to their own brain activity, but is instead yoked to a matched experimental subject's performance. **Subjects undergo three days of neurofeedback training.** Notably, subjects are not given any instructions about what to think about in order to modulate the size of the circle. Similar covert real-time neurofeedback has been used

in some recent neuroimaging tasks. The important advance in my proposal is the ability to enhance the neural representation of either fear or extinction by strengthening reactivation of the context in which fear or extinction was learned. **Thus, the critical test of neurofeedback training is a fear renewal test and stimulus generalization test on Day 5.** For this, we present a number of tone frequencies that vary in pitch between the CS+ and CS- for a test of fear generalization to stimuli that approximate a learned threat.

Interpretation and Future Directions: These data will be interpreted in the framework of emerging rodent models of fear and extinction [7], associative learning theory models of fear relapse [6,19], and models of temporal context binding from human episodic memory research [15]. If this project is successful, it will reveal for the first time in humans that contextual renewal is mediated by reactivation of the mental context at the time of learning, and this pattern can be modulated to induce high or low fear states. It also has far reaching implications for developing a neural signature of safety that could estimate the success of clinical treatments. Finally, decoded neurofeedback could be used to ameliorate symptoms of fear and anxiety in psychiatric disorders by teaching patients verifiable techniques to self-modulate retrieval of safety associations following therapy.

Overcoming potential problems: The nature of this ambitious research involves inherent methodological risk. One could argue that if we already knew how to solve these problems, then the questions might not be worth exploring to begin with. Strong relationships with expert collaborators will help defend against many of the potential methodological and interpretative challenges likely to arise in this ambitious undertaking. I highlight a few of the foreseen circumstances that might arise in the course of this project.

(1) Some studies report that as many as 1/3 of subjects are unable to control neurofeedback in fMRI [17]. We will use intermittent neurofeedback, which is more effective than continuous neurofeedback with no explicit strategy instructions. If subjects struggle to self-regulate their neural activity, then we can extend the delay period to 16-s to allow more time to sample neural activity to derive a feedback signal.

(2) There remain important theoretical questions on the psychobiological mechanisms of neurofeedback that require more attention in the field [17], but these questions do not present a hurdle to the present design.

(3) Reinforcing the fear context might on first consideration be a form of appetitive "counterconditioning" that indirectly *counteracts* fear of the CS that was conditioned in that context. This is an emprical question, but evidence in rodents suggests the opposite. And the ultimate goal of this work is to diminish fear.

(4) Fear extinction is time limited and fallible – even healthy adults can show relapse of fear behaviors. Thus a potential problem is that our subjects might have difficulty reactivating the memory of the extinction context. If this proves problematic, then we can employ strategies to strengthen extinction learning. For instance, we could double the number of extinction trials or replace the shock with a non-aversive outcome [20,21] as I am currently investigating as part of an NIMH K99/R00.

Is it fear or threat? LeDoux and colleagues recently challenged the idea that defensive behaviors in conditioning paradigms can be taken as evidence for the emotion of fear, per se [22]. We also acknowledge that the shock used in human research does not induce severe levels of intense emotional distress akin to a PTSD-like experience. However, subjective ratings collected at the end of our experiments confirm that subjects report "fear" or a synonymous construct (e.g., dread) during the task, supporting the construct validity of the conditioning design.

***Preliminary data from R00*:** We have collected data from 25 healthy adults and 25 patients with PTSD on an NIMH R00 funded fMRI task that provides the motivation and support for this ambitious proposal. In this experiment, task-irrelevant pictures of scenes were injected between each extinction trial to "tag" the mental context of extinction. The next day, subjects underwent a fear renewal test in a novel context. Altogether, our preliminary data demonstrate feasibility of key elements of this high-risk high-reward design. Specifically, this data supports my lab's ability to (**1**) tag a mental context, (**2**) use MVPA classifiers to train and then decode the reactivation of a mental context, and (**3**) link spontaneous contextual reactivation to extinction neurocircuitry in healthy and PTSD participants (**Fig. 4**). First, we verified MVPA pattern classifiers are able to successfully decode categorical representations of the task stimuli in visual cortex (**Fig 4A**). We then applied the classifiers to data from the fear renewal test and found strong evidence of neural reactivation of the extinction context from the previous day (**Fig. 4B**). That is, the classifier detected spontaneous activity in ventral visual cortex during test that was classified as belonging to the category of context tags they had encountered between trials during extinction learning. Importantly, the amount of scene reactivation at test correlated with vmPFC and hippocampal activity in healthy controls, but not PTSD. Thus, preliminary results provide important foundations for the proposed research on causal brain-behavior relationships at the core of this proposal.

Preliminary data on decoded neurofeedback: Work from collaborator Jarrod Lewis-Peacock's lab (**Fig. 4 C**) shows the ability to provide meaningful real-time MVPA decoded neurofeedback. Pattern classifiers trained on sensorimotor cortex activity was associated with making button presses with individual fingers. These classifiers were used to decode, in real-time, brain activity associated with pressing a finger. Neurofeedback was provided based only on that subject's brain activity, about whether they were pressing the correct "target" finger for that trial. Real-time decoding was well above chance. We will extend the "context tagging" procedure firmly established by the R00 project to create detectable multi-voxel patterns of brain activity. Then, we will perform decoded neurofeedback training *to reinforce the match between encoding and retrieval of the memory*.

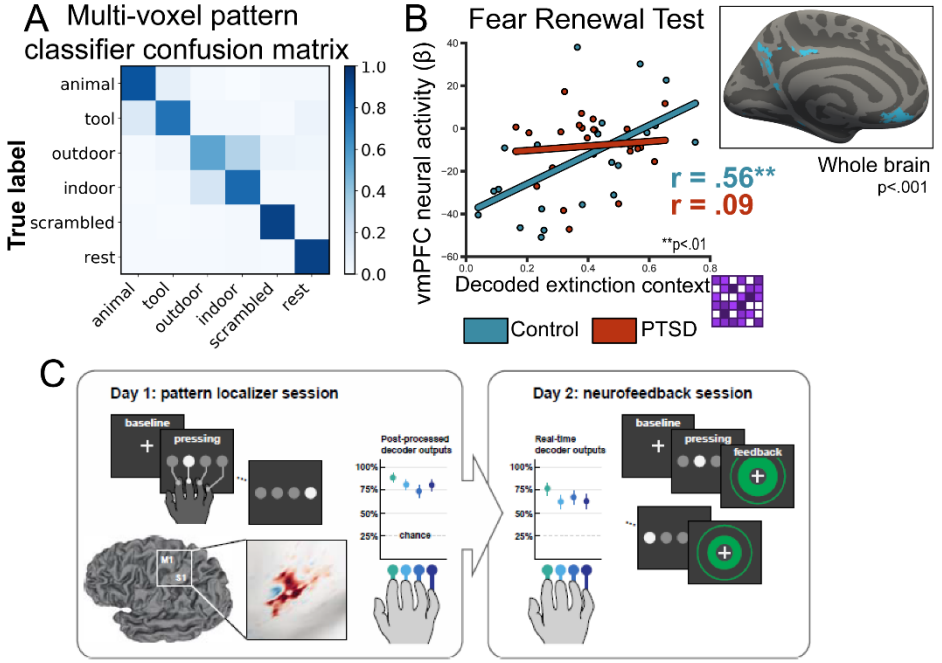

Figure 4 Preliminary data provide feasibility for this proposal. (A) From the R00 to PI Dunsmoor, a confusion matrix shows successful MVPA classification in visual cortex. (B) Classifier evidence for neural reinstatement of extinction context correlates with vmPFC activity in Healthy-Controls, but not PTSD. (C) From a separate task by collaborator Lewis-Peacock, we can decode in real-time MVPA-trained activity.

Real-time fMRI protocol: Whole brain fMRI data will be acquired using a 3T Siemens Vida MRI scanner with a 64-channel head coil at the Biomedical Imaging Center at the University of Texas at Austin. The standard method provided by the scanner manufacturer to obtain images in real time is plagued by several bottlenecks, which can serve as severe rate-limiting factors for real-time fMRI neurofeedback. Our solution overcomes all of these bottlenecks by increasing processing speed at several points along image reconstruction. The results obtained from this pipeline from our imaging facility are impressive. Using the standard methods, DICOM images were delivered 7 ± 12 sec after k-space data acquisition. The first measurements using the **new methods** are showing raw image arrival a mere *0.08 ± 0.07* seconds after k space acquisition. This dramatic reduction in real-time fMRI data transfer latency, and perhaps even more importantly, the increased reliability of that latency, is a feature we consider to be crucial to this project and to the field in general.

Research collaborators: This project is bolstered by a world leading expert on the neuropathophysiology of PTSD and two leading experts in advanced multivariate and real-time functional neuroimaging at UT Austin.

Dr. Charles Nemeroff, M.D., Ph.D. is the Acting Director of Psychiatry and Director of the Institute of Early Life Adversity at UT Austin Dell Medical School. He has over 1,000 scientific publications in basic neuroscience, pathophysiology of mood and anxiety disorders, and psychiatric treatment studies. He will provide expertise and rigorous oversight on the clinical translational aspects of this project.

Dr. Cameron Craddock, Ph.D. is the Director of the Imaging Analysis and Informatics Core and Associate Professor in the Department of Diagnostic Medicine at UT Austin Dell Medical School. He will support aspects associated with the functional neuroimaging and real-time neurofeedback.

Dr. Lewis-Peacock, Ph.D. is an Assistant Professor in the Department of Psychology. His lab combines functional neuroimaging and computational approaches to explore human learning and memory processes. He will contribute his expertise for the development, implementation, and analysis of multivariate pattern analysis and real-time decoded neurofeedback.

III. Innovativeness

Although many neuroimaging studies have sought to identify the neural circuits of fear and anxiety in humans, there are crucial gaps in our ability to identify, isolate, and manipulate neural representations of fear and safety in the human brain. This proposal is motivated by pioneering advances in rodent neurophysiology showing that separate neural populations code for fear and safety. Once identified, these neural populations can be

manipulated (e.g., via targeted optogenetic stimulation) to enhance or diminish fear behavior. Technology to isolate individual neurons in human neuroimaging does not exist. But advances in machine learning and pattern analysis of high resolution human functional neuroimaging are now able to identify neural activity in fine detail. Moreover, closed-loop neurofeedback of multivariate fMRI activity (i.e., decoded neurofeedback) now allows experimenters to modulate neural activity of these fine scale neural representations. I propose to combine these theoretical and technical advances to translate advances in rodent neurophysiology to human clinical populations, with the goal of **causally enhancing** the neural reinstatement of safety memories and diminishing the neural reactivation of fear. This work represents a novel and transformative approach that extends beyond typical human neuroimaging research of fear and anxiety.

It is increasingly clear from pioneering technical advances in neural tagging that distinct neural populations mediate the balance between expression of a fear or extinction [7]. But whether these competing experiences instantiate stable and distinguishable memory traces in the human brain, and whether humans can manipulate these neural representations through self-regulation, is unknown. I aim to detect a stable memory trace of fear and extinction by leveraging advances in multivariate neuroimaging techniques for human memory. This represents the first effort to combine these tools to localize distinct representations of fear and extinction in the human brain, as well as to compare these representations in a healthy and clinical population. **Furthermore,** determining causality in neural systems is a challenge—particularly so for human neuroimaging, which is mostly correlational by design. Thus, another critical innovation lays in the potential to causally manipulate the memory of safety to make extinction memories stronger, more generalizable, and more resilient. I will extend this novel, and potentially transformative research to PTSD, with the goal of ultimately developing a clinical intervention.

IV. Investigator Qualifications

I have long been fascinated with understanding how emotion and cognition interact to determine how humans learn about and remember emotionally meaningful events. Following college I sought postbaccalaureate research positons in neuroimaging labs investigating the cognitive neuroscience of emotion. I received the Intramural Research Training Award from NIH to work in the laboratory of Dr. Peter Bandettini. I conducted, analyzed, and published two fMRI projects on human fear conditioning as first author (Dunsmoor et al., 2007, *Behavioral Neuroscience*; 2008, *Neuroimage*). Research experience at NIMH sparked my interest in using neuroimaging to investigate how fear shapes learning and memory for my PhD.

PhD training in cognitive neuroscience: Duke University, Kevin LaBar's Lab: I was awarded an NRSA (F31) from NIMH during graduate school to learn sophisticated fMRI analyses including multi-voxel analysis, functional connectivity, and resting state analyses to investigate different routes of fear generalization in humans: the perceptual route (between stimuli that physically approximate a learned threat) and the conceptual route (between stimuli that share a categorical relationship but minimal physical overlap). I developed this novel line of research along with my mentor Dr. Kevin LaBar. When I began my PhD training, existing publications on fear generalization in humans were extremely rare. We published one of the first contemporary behavioral studies of fear generalization in humans (Dunsmoor et al., 2009, *Learning & Memory*) and the first fMRI investigation of fear generalization in humans (Dunsmoor et al., 2011, *Neuroimage*). In the past several years, the topic of fear generalization has received substantial interest in the field (reviewed in Dunsmoor & Paz, 2015, *Biological Psychiatry*; Dymond, Dunsmoor et al., 2015, *Behavior Therapy;* Dunsmoor & Murphy, 2015, *Trends in Cognitive Sciences*).To widen the scope of my research questions, I collaborated with Dr. Alex Martin at NIMH, an expert on neural systems of object representation, to examine how fear conditioning modulates cortical representations of object concepts (Dunsmoor, Martin, & LaBar 2012, *Biological Psychology* ; Dunsmoor et al., 2014, *Cerebral Cortex*). I also gained experience with computational models of associative learning (Dunsmoor & Schmajuk, 2009, *Behavioral Neuroscience*). Overall, my graduate training was productive, and I managed to publish 8 empirical papers (7 first authored) and received the APA Dissertation Research Award in addition to the NRSA.

Postdoctoral Training: New York University, Elizabeth Phelps' Lab: My ultimate research goal is to contribute to a better understanding for how humans can learn to control unwanted fear memories. I joined Dr. Elizabeth Phelps' laboratory at NYU to gain new expertise in the neurobehavioral mechanisms of fear extinction. Liz Phelps is recognized as a world leader in this domain, using innovative experimental techniques and methods to tackle important questions on the cognitive neuroscience of emotion broadly, and fear extinction specifically. I was awarded a K99/R00 Pathway to Independence Award from NIMH to investigate novel strategies to improve the control of fear. As part of my K99 mentoring team, I received training from Helen Blair Simpson (Columbia)

on clinical translational research, Joseph LeDoux on neuroscience models of extinction in rodents, and computational methods from Nathaniel Daw. In addition to the success of achieving my training goals at NYU, I was fortunate to publish a number of my research findings in high impact journals including *Nature* [23], *Neuron* [3], *Biological Psychiatry* [5,21], *Psychological Science* [24,25], *Journal of Neuroscience* [20], and *PNAS* [26,27]. Overall, my postdoc training was productive, and my postdoc work resulted in 23 empirical papers (12 first-authored).

Assistant Professor at the University of Texas at Austin: I am currently directing a lab on several projects related to understanding how emotion affects learning and memory processes in humans. In addition to starting my lab with support from an R00, I was recently honored to receive a CAREER Award from the National Science Foundation (NSF), and a Young Investigator Award (NARSAD) from the Brain & Behavior Research Foundation. This early funding has already spurred a number of productive collaborations across the departments of Psychiatry, Psychology, and Neuroscience at UT. These collaborations present excellent opportunities for interdisciplinary research. By translating basic learning and research findings from rodents, to healthy adults, and ultimately to clinical populations, we can gain insight into compromised emotional learning circuitry in pathological anxiety. This information will contribute to neurobiological models of pathological fear and anxiety and may lead to innovative neuroscience-based therapies. Thus, the line of research described in this proposal has the potential to make a significant impact. Overall, my academic career has prepared me to lead a productive independent laboratory and set a solid foundation from which to pursue the research set forth in this proposal. With my expertise in associative learning and neuroanatomical models of fear and extinction, use of sophisticated functional neuroimaging, and clinical translational experience honed in my K99/R00, I feel ideally suited and uniquely qualified to carry forward the proposed research focusing on localizing and manipulating separate memory traces of fear and extinction in the human brain.

V. Suitability for the New Innovator Award program: This ambitious research project is uniquely suited to the New Innovator Award, rather than a more "traditional" grant mechanism, because I seek to develop an entirely new line of work whereby we can isolate and manipulate a neural representation specific to a memory of fear versus safety in the human brain. The use of multivariate decoded neurofeedback to enhance retrieval of a specifically valenced emotional memory is the embodiment of a high-risk high-reward project. Pioneering advances in rodent neurophysiology lay the theoretical groundwork, and findings from my K99/R00 funded research lays a solid foundation to purse this highly innovative line of work. Self-modulation of neural patterns associated with a memory of safety has broad implication for alternative clinical treatments for a host of psychiatric disorders that can be viewed through the lens of associative learning.

VI. Statement of research effort commitment: If chosen to receive this award, I will commit a minimum of four person-months (33.33%) of my research effort to the project supported by the New Innovator Award.

VII. References

1 Etkin, A. & Wager, T. D. *American Journal of Psychiatry* **164**, 1476-1488 (2007).
2 LeDoux, J. E. *Annual Review of Neuroscience* **23**, 155-184 (2000).
3 Dunsmoor, J. E., Niv, Y., Daw, N. D. & Phelps, E. A. *Neuron* **88**, 47-63 (2015).
4 Milad, M. R. & Quirk, G. J. *Annual Review of Psychology* **63**, 129-151 (2012).
5 Dunsmoor, J. E. & Paz, R. *Biological Psychiatry* **78**, 336-343 (2015).
6 Bouton, M. E. *Learning & Memory* **11**, 485-494 (2004).
7 Tovote, P., Fadok, J. P. & Lüthi, A. *Nature Reviews Neuroscience* **16**, 317-331 (2015).
8 Josselyn, S. A., Köhler, S. & Frankland, P. W. *Nature Reviews Neuroscience* **16**, 521-534 (2015).
9 Maren, S. *Eur J Neurosci* **28**, 1661-1666 (2008).
10 Nemeroff, C. B. *et al. Journal of psychiatric research* **40**, 1-21 (2006).
11 Phelps, E. A. & LeDoux, J. E. *Neuron* **48**, 175-187 (2005).
12 Fullana, M. *et al. Molecular Psychiatry* (2015).
13 Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. *Trends in Cognitive Sciences* **10**, 424-430 (2006).
14 Rugg, M. D., Johnson, J. D., Park, H. & Uncapher, M. R. *Progress in brain research* **169**, 339-352 (2008).
15 Howard, M. W. & Kahana, M. J. *Journal of Mathematical Psychology* **46**, 269-299 (2002).
16 Kahana, M. J. *Memory & cognition* **24**, 103-109 (1996).
17 Sitaram, R. *et al. Nature Reviews Neuroscience* **18**, 86 (2017).
18 Dunsmoor, J. E., Kragel, P. A., Martin, A. & LaBar, K. S. *Cerebral Cortex* **24**, 2859-2872 (2014).
19 McConnell, B. L. & Miller, R. R. *Learning and Motivation* **46**, 1-15 (2014).
20 Dunsmoor, J. E. *et al. Journal of Neuroscience* **39**, 3264-3276 (2019).
21 Dunsmoor, J. E., Campese, V. D., Ceceli, A. O., LeDoux, J. E. & Phelps, E. A. *Biological Psychiatry* **78**, 203-209 (2015).
22 LeDoux, J. E. & Pine, D. S. *American Journal of Psychiatry* **173**, 1083-1093 (2016).
23 Dunsmoor, J. E., Murty, V. P., Davachi, L. & Phelps, E. A. *Nature* (2015).
24 FeldmanHall, O., Dunsmoor, J. E., Kroes, M. C., Lackovic, S. & Phelps, E. A. *Psychological science* **28**, 1160-1170 (2017).
25 Dunsmoor, J. E. & Murphy, G. L. *Psychological Science* **25** (2014).
26 FeldmanHall, O. *et al. Proceedings of the National Academy of Sciences* **115**, E1690-E1697 (2018).
27 Dunsmoor, J. E., Otto, A. R. & Phelps, E. A. *Proceedings of the National Academy of Sciences* **114**, 9218-9223 (2017).