A. Significance. Anxiety Disorders and Posttraumatic Stress Disorder (PTSD) have a combined prevalence estimated at nearly 22% of the US adult population. Emerging research is detailing brain regions and functional networks affected across these disorders [1-3]. This research reveals abnormalities in the amygdala, hippocampus, and ventromedial prefrontal cortex (vmPFC), regions critically involved in forming and retrieving extinction memories [4]. Accordingly, neuroimaging research has characterized abnormalities associated with deficits in extinction of conditioned fear in psychiatric populations [5]. Decades of neurobehavioral research on extinction has culminated in its inclusion as a paradigmatic tool for the Research Domain Criteria Project (RDoC) to further advance transdiagnostic understanding of disorders of Acute Threat ("Fear") in the Negative Valence Systems Matrix. Extinction also forms the basis for exposure therapy, the most widely used evidence-based treatments for pathological fear and anxiety [6,7]. Overall, experimental extinction is a valuable model to conceptualize the etiology, maintenance, and relapse of pathological anxiety [5,8-10], and advances in extinction research are significant to help innovate clinical treatment. However, despite recent advancements in the neuroscience of extinction, there remain crucial gaps in our understanding of how the human brain forms, stores, and retrieves separate memories of fear and extinction.

Rodent neurobiological research shows distinct representations of fear versus extinction memory traces coded within and between the amygdala, hippocampus, and mPFC. Pioneering technical advances in fine-scale molecular imaging and activity-dependent neural tagging reveals extinction to be an active learning process activating separate neural populations with distinct subcortical-cortical pathways [11,12]. This leading-edge research confirms theories of extinction—dating back to the time of Pavlov [13,14]—that extinction is new learning, and not unlearning. Neurobiological research also confirms retrieval-based accounts of extinction [15]; that is, neural activity in canonical extinction networks shifts according to the context the animals is in at test [16]. *Whether a similar neural organization exists in the human brain, whereby fear and extinction memories are segregated in separated neural regions and ensembles, is unknown.* Directly translating the neuroscience of extinction from animal neurophysiology to human neuroimaging has remained a challenge, because precise methodology for tagging neural populations in humans is lacking. Neuroimaging evidence of extinction memory retrieval success or failure has been inferred using univariate subtraction methods of fMRI activity [17]. But coarse univariate measures of average fMRI activity have been insufficient for identifying how competing memories of fear and extinction are learned, stored, retrieved, and expressed. Indeed, several neuroimaging meta-analyses have failed to show consistent engagement of the amygdala and vmPFC in conditioning and extinction [17-21], despite overwhelming evidence for the role of these regions in laboratory animals.

Here, we propose to isolate quantifiable and separate memory traces of fear and extinction in the human brain, to discover whether and how memory representations of extinction change over time, and compare neural signatures of extinction memory fidelity between healthy adults and patients with PTSD. *We focus on PTSD because linking a multivariate signature of fear extinction to the pathophysiology of PTSD can have direct benefit to exposure therapy—the gold-standard treatment based on the principles of extinction.* We have devised a way to leverage advances in multivariate pattern analysis (MVPA) [22,23] of fMRI data to localize spatially distributed patterns of activity unique to the encoding and retrieval of fear conditioning and extinction in humans. We developed a series of novel tasks optimized for detecting stable and separate memory traces of fear and extinction over time by combining theoretical approaches to the study of extinction from animal models with technical advances in the cognitive neuroscience of human memory. In short, *this research is significant because it combines theoretical knowledge on the neuroscience of associative learning with advances in the cognitive neuroscience of human memory to better understand how fear and extinction are separately learned, remembered, and expressed in the healthy brain and in disease.*

A.2. Evidence of separate representations of fear and extinction memories in the rodent brain. 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 [24,25]. Fear conditioning is an indispensable model to answer questions on the nature of memory representations in the brain; it is rapid, strong, stable, and has objective neural and behavioral correlates conserved across species [26]. Further, learning models based on the principles of conditioning provide explanatory power for characterizing a range of psychiatric disorders [5,8,9,27], and has been especially beneficial to understanding the pathophysiology of PTSD [2]. But it has remained challenging to investigate fear and extinction memory representations in the human brain. As detailed below, conventional approaches to translate the neuroscience of extinction to human neuroimaging have failed to identify consistent and robust engagement by key regions of interest. The research proposed here overcomes these limitations to investigate extinction

memory representations in the human brain by integrating sophisticated computational analysis approaches employed elsewhere in the field of cognitive neuroscience of human memory. ***A division of labor in the medial PFC***: Neural evidence that extinction is new learning (not unlearning) was provided by early work showing that extinction is NMDA-dependent [32]. Subsequent work showed the infralimbic (IL) cortex [33,34], considered homologous to the human vmPFC [5,35], is a critical site of extinction memory formation and retrieval; whereas the prelimbic (PL) cortex, homologous to the dACC is involved in fear expression (**Fig 1**). These areas provide top-down control over the amygdala to determine expression and suppression of conditioned fear [36]. Importantly, the balance in activity between the IL and PL is contextually driven [37]. For instance, neurons in the IL show enhanced activation when an extinguished CS is encountered in the extinction context, but reduced activation when the extinguished CS is encountered in another context [38]. Human neuroimaging has successfully translated evidence from rodents that the dACC is strongly active during fear conditioning. Indeed, meta-analyses show the dACC is, along with the insula and thalamus, the most consistently active regions during fear conditioning [17-19]. However, neuroimaging evidence of vmPFC involvement in ***extinction*** is surprisingly scant. Indeed, meta-analyses show the vmPFC is *not* among a collection of regions active during extinction learning or recall [18,21]. This discrepancy between animal neurophysiology and human neuroimaging has been a puzzle, but is likely due to methodological limitations of conventional neuroimaging analytical approaches used to investigate extinction.

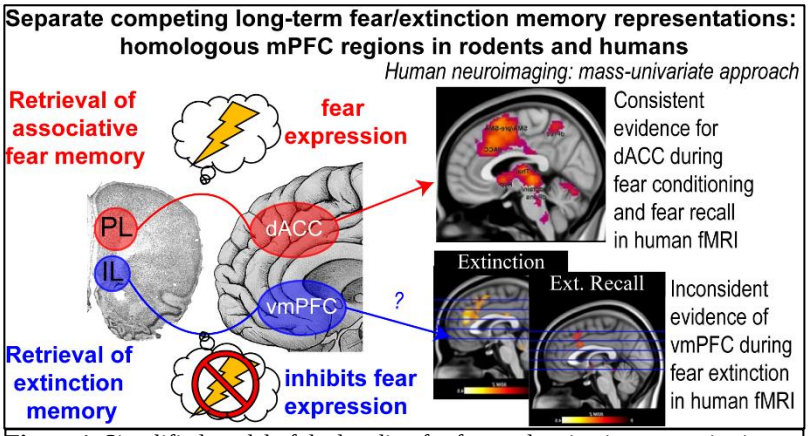

Figure 1. Simplified model of dual coding for fear and extinction memories in dissociable mPFC regions. Translation of vmPFC findings to humans is inconsistent. Data from meta-analysis (Fullana et al., 2016; 2018)

A.2. Applying multivariate pattern analysis (MVPA) to investigate conditioning and extinction in the human brain. Functional MRI offers the ability to broadly translate neurophysiological research from rodents to humans [39], but lacks the spatial resolution to detail the same level of neuronal organization. Conventional fMRI analysis uses mass univariate voxel-wise subtraction methods based on the general linear model [17,18,39]. Univariate fMRI detects voxels that show a maximal response on a given trial or 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. Oftentimes, activity to one condition is contrasted with (subtracted from) activity to another condition. This approach is optimal for detecting brain regions that show robust activity across a large collection of voxels that mostly respond in the same way. Univariate approaches are best understood as identifying *engagement* of a particular brain region in a specific task. But mean univariate subtraction methods ignore information coded at finer spatial resolutions and distributed across collections of voxels [40]. This limitation is especially relevant when trying to translate the neurobiology of animal models to human neuroimaging (**Fig 1**).

By comparison, MVPA is an important approach to investigate how *information* is processed in distributed voxels. Recent developments in MVPA make it possible to non-invasively decode information content of mental representations at small spatial scales with astounding accuracy and specificity. This advance reveals 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. We will use MVPA of fMRI data to investigate how fear and extinction memories are separately represented and contextually modulated. *This approach is important, as conventional univariate approaches have failed to consistently translate major findings from animal neurophysiology to humans*. As a strong example, despite clear evidence that the amygdala is crucial for fear conditioning [4,48], evidence of amygdala *activity* in univariate human fear conditioning fMRI studies is inconsistent at best. In fact, a recent meta-analysis of univariate 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 [17]. 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. Taking an MVPA approach to fear conditioning has revealed sparsely distributed voxels in the amygdala coding for the fear conditioned (CS+) and control (CS-) stimuli during fear conditioning [41,46]. Likewise, despite some early reports

[52,53], there is inconsistent evidence of vmPFC activity during extinction learning and extinction retrieval in humans using conventional univariate approaches, including in the PI's own recent K99R00-funded work [54]. Indeed, a recent meta-analysis of 1300 participants failed to reveal amygdala or vmPFC in human fear extinction [21]. One explanation for failures to detect vmPFC engagement is the baseline condition (CS-) already elicits enhanced vmPFC activity, as shown in the inverse contrast from CS- > CS+ during fear *conditioning*. As such, contrasts of mass-univariate activity to the CS+ versus the CS- during fear extinction renders the comparisons between the already known safety signal CS- and extinguished CS+ obsolete, as the vmPFC may be responding to both cues to a similar level. Altogether, our goal 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 and extinction (*described in more detail below*). Our plan is built on two advances on computational models of episodic memory. First, neurocognitive processes active at memory formation are reinstated during retrieval [55-58]. Second, information is often linked to the context in which it was encoded [59,60].

A.3. PTSD symptom clusters and fear extinction. PTSD is a heterogeneous disorder composed of multiple symptoms, including re-experiencing, avoidance, negative cognitions and mood, and hyperarousal. Detecting symptom clusters most associated with deficits forming and retrieving memories of safety is an important concern to clinical translational neuroscience. There is some evidence that activity during conditioning and extinction in a network of regions (e.g., amygdala, dACC, insula) correlates with avoidance and hyperarousal symptoms [61,62]. Notably, successful extinction likely targets avoidance and hyperarousal symptoms (as opposed to re-experiencing and negative mood), and thus improving extinction should presumably have corresponding effects on reducing PTSD symptoms. Whether innovative forms of extinction mitigate PTSD symptoms is unknown. This project is the first attempt to establish the neurobehavioral mechanisms underlying the effects of enhanced extinction learning in PTSD. This may ultimately inform further augmentations to exposure based therapy to yield even more effective treatment plans. We therefore plan to use symptom severity, along with other important markers of childhood trauma and potential comorbid depression severity, as covariates in our analyses, in line with emerging studies linking PTSD symptoms to fMRI measures of conditioned fear learning [61,63].

A.4. Aim 1 background: *Pavlovian conditioning and episodic memory: a hybrid approach to isolate* specific *memory representations of conditioning and extinction*. Emotional memory research tends to be approached from two academic traditions that differ substantially in methodology and intellectual tradition: Pavlovian conditioning and episodic memory. Conditioning is often characterized as an *implicit* and reflexive form of learning, as evidenced by the studies showing conditioning in lower animals and in humans in the absence of awareness. However, conditioning is not purely implicit, and there is abundant evidence on cognitive processes in Pavlovian conditioning and extinction in humans [64-66]. One reason conditioning is often construed as purely implicit is that the memory demands in a typical conditioning protocol tend to be minimal: e.g., the use of a single unimodal CS repeated numerous times. Conditioning and episodic memory have also been presented in a taxonomy of memory that splits implicit and explicit memory systems as entirely separated systems [67]. But emotional learning invokes multiple memory systems in the brain [68]. For instance, the CS-US representation is known to involve multiple elements and associative content (e.g.,

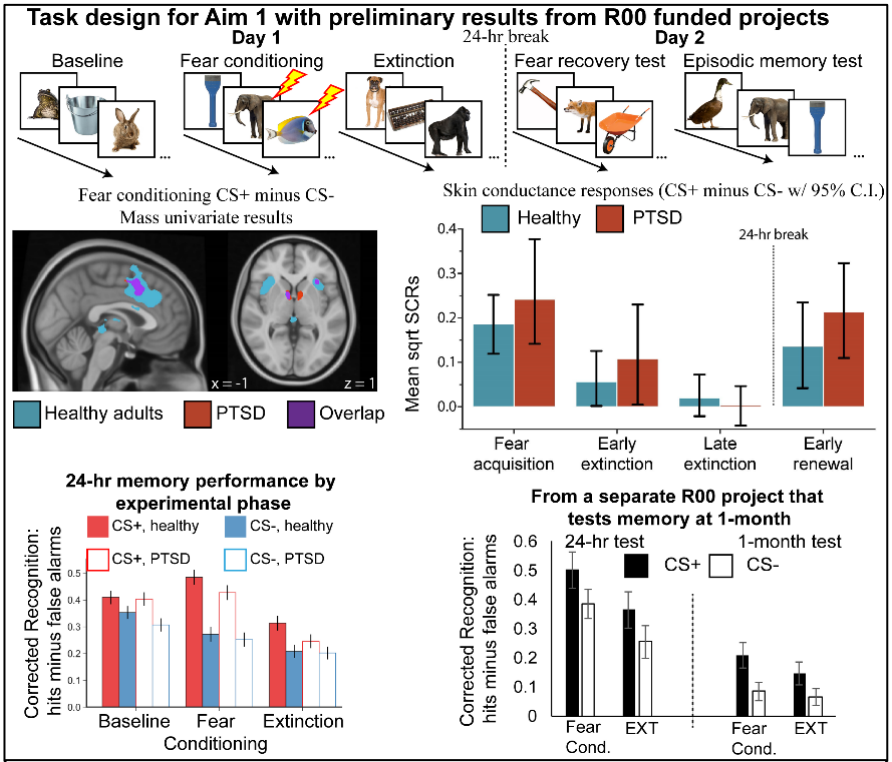

Figure 2. Prelim. results show dACC and insula activity, and selective SCRs and recognition memory for CS+ vs CS-. Memory for CS+ extends to 1 month.

emotive, temporal, contextual, sensory elements) [69,70]. As conditioning tasks are usually simplified designs translated from animal protocols, measures of *explicit* memory are rarely assessed or considered meaningful. But there is considerable research on how emotion affects episodic memory outside the conditioning domain [71,72]. Detecting the overlaps between fear conditioning and emotional episodic memory requires a subtle modification to traditional conditioning procedures.

Category-conditioning. To assess how emotional *learning* experiences are selectively prioritized in long-term episodic memory, the PI developed a hybrid conditioning/episodic memory task that incorporates trial-unique (i.e., non-repeating) basic level exemplars from distinct categories (e.g., animals and tools) as CSs during differential (CS+/CS-) fear conditioning (**Fig 2**). *We have repeatedly found that fear conditioning affects item memory by selectively enhancing recognition memory for CS+ versus CS- items [46,73-79].* We also recently found that fear *extinction* appears to segment memory traces between information encoded during conditioning and conceptually-related but extinction-specific CS+ exemplars encoded shortly thereafter [75,77]. Specifically, memory for extinction-specific CS+ exemplars are remembered at a lower rate than conditioning-specific CS+ exemplars from the same object category. That extinction produces a weaker episodic memory trace than fear conditioning has particular relevance for understanding why fear often returns over time in psychiatric disorders like PTSD. That is, extinction memories might simply be harder to retrieve than fear memories because both the associative and episodic content of an extinction memory is more transient. This also suggests that it might be possible to *enhance* extinction memory retrieval by focusing on the *episodic* content specific to extinction. A more robust extinction episodic memory trace might help outcompete expression of the fear memory trace, which likewise contains implicit and explicit elements. *Our preliminary data support this suggestion, showing that encoding-retrieval overlap in the vmPFC in PTSD patients is robust only for CS+ exemplars that were explicitly remembered.* Altogether, conditioning and extinction are often assumed to be entirely implicit forms of learning, but behavioral evidence shows selective enhancement in long-term episodic memory for fear conditioned exemplars, and relatively weaker episodic memory for extinction-specific exemplars. In sum, we propose that integrating theoretical knowledge and experimental techniques from conditioning and episodic memory can provide novel insights into the long-term stability of neural representations of fear and extinction memory.

Does neural similarity between encoding and retrieval reveal separate representations of fear and extinction specific memory? Retrieving a memory activates neurocognitive processes active at the time of encoding, sometimes referred to as transfer-appropriate processing [55]. Advances in MVPA affords the opportunity to investigate whether similarity between neural states active at encoding and retrieval predicts memory performance [58]. Specifically, representational similarity analysis (RSA) [23] can be used to evaluate the correspondence between activity at encoding and activity at retrieval (referred to as "encoding-retrieval overlap") across distributed voxels [40] (**Fig 5 below**). This approach is frequently used in human episodic memory research and shows encoding-retrieval overlap in the hippocampus and cortical areas involved in memory formation [57,80-82]. We leverage these advances in estimating encoding-retrieval similarity to address a fundamental question on the organization of *conditioning* and *extinction* memories. We focus our analysis predominately on discovering whether fear and extinction are segregated in dissociable mPFC regions and amygdala connectivity between these regions, in line with rodent models. We further compare encoding-retrieval similarity (ERS) for CS+ and CS- specific items encoded during conditioning and extinction at different retention intervals: Half of the encoded trials will be tested at 24-hours, and the other half will be tested at 1 month. If the neural representation of extinction memories are more vulnerable to decay, it might provide a mechanism to explain why spontaneous fear recovery increases over time [83]. *We therefore predict stronger degradation of extinction-specific episodic memory in PTSD versus healthy adults.* Note: We can assess ERS for each CS item as a function of when it was encoded, *regardless of whether a subject* ***remembers*** *that item*. But because the retrieval session is a recognition memory test, we have the *added* ability to gauge ERS as a function of explicit memory performance.

A.5. Aim 2 background: *Contextual specificity of extinction.* Extinction renders the meaning of the CS ambiguous—that is, the CS now signals threat or the absence of threat [84]. The brain resolves this ambiguity based on context. Specifically, fear is suppressed in the extinction context but is expressed in a different context, an effect known as "renewal." Renewal can explain why clinical treatments often fail to generalize beyond the therapeutic environment. Context-dependent extinction retrieval and contextual renewal are mediated by the hippocampus [84-86]. For example, contextual renewal is associated with increased activity in the hippocampus in rodents, in particular neurons in ventral hippocampus projecting to the IL/vmPFC [85,87]. Lesions or inactivation of the hippocampus prevents fear renewal, making extinction retrieval context independent [88-90]. Evidence for the role of the human hippocampus on fear extinction retrieval is mixed, and is not necessarily consistent with

emerging neurobiological models of renewal. For instance, early reports showed evidence of enhanced activity in the hippocampus and vmPFC during successful extinction recall in the extinction context [53,86]. Yet neurobiological models posit that hippocampal activity is associated with activity *outside* the extinction context, leading to renewal by activating the amygdala [38] and inhibitory neurons in the mPFC [85,87]. The goal of **Aim 2** is to help determine how context mediates extinction retrieval by decoding "neural context reinstatement" at test.

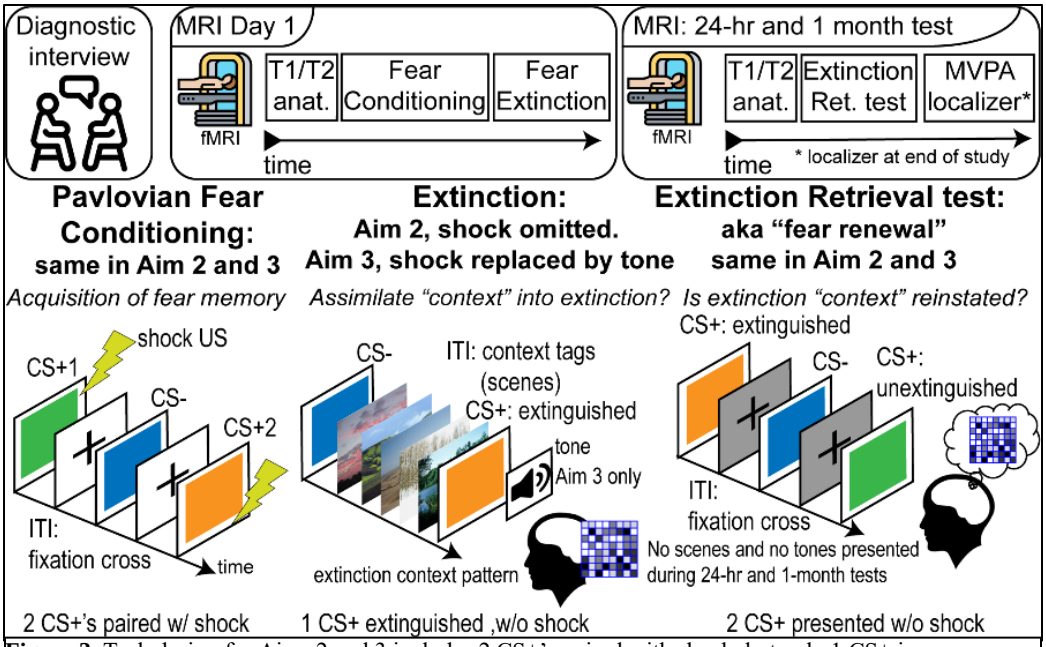

Figure 3. Task design for Aims 2 and 3 includes 2 CS+'s paired with shock, but only 1 CS+ is extinguished. During extinction, task-irrelevant pictures of scenes are presented to "tag" the context. We can use MVPA approaches to decode spontaneous neural reinstatement of the extinction context at test. In Aim 3 only, the shock is replaced by a neutral tone during extinction, consistent with our prior work on "novelty-facilitated extinction."

There are numerous parallels to the role context plays in fear extinction and in human memory more generally [91-93] that we leverage for the Approach of **Aim 2**. For example, a classic finding is that memory is worse when the room changes between study and test [94]. Importantly, "context" also includes time, mood, internal conditions, and other types of mental states [92]. Like the physical context, similarity between the "mental context" of encoding and retrieval helps guide retrieval [95]. Moreover, reinstating a "mental context" appears to counteract the deficit usually caused by a change in the physical context between learning and test [94]. *A similar process of mental context reactivation might occur in cases where extinction learning does successfully transfer to a new physical environment [91]. For instance, reminders of the extinction context can help prevent renewal in rats [96].* **Aim 2** will evaluate whether *mental context reinstatement* prevents fear renewal by synthesizing theoretical [60] and empirical [97] work on contextual reactivation of *episodic* memory with theoretical [84] and empirical [15,85] work on contextual specificity of extinction.

"Mental context." Integration of numerous details from the time of encoding provides a "mental context representation" that can later guide memory retrieval. A computational model of human memory has formalized how information is bound to a gradually accumulating contextual representation, known as the Temporal Context Model [60]. This model was 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 [98]. Neuroimaging experiments have cleverly incorporated this model to decode brain activity related to the retrieval of items that had been encoded in different contexts [97,99]. In **Aim 2**, we leverage these theoretical principles to derive a novel measure of *extinction* context reinstatement (**Fig 3 and described in Aim 2 Methods**). We predict that *mental* context reinstatement can orchestrate initiations of a low fear state in other brain regions, helping guide emotional responses in a novel context when threat is ambiguous. We compare results to healthy adults and PTSD, as extinction-retention deficits in PTSD are well-documented [2,100] and may be a factor in relapse.

A.6. Aim 3 background: *Novelty-facilitated extinction: a strategy to enhance fear extinction and prevent the return of fear.* Advances in extinction research are clinically relevant only insofar as they can be leveraged to advance treatment. A major focus in the field of extinction research is on strategies to enhance extinction to prevent the return of conditioned fear. This includes innovative *non-pharmacological* behavioral approaches built on a solid foundation of associative learning theory [10,70]. The PI has developed a new approach that involves replacing, rather than just omitting, aversive outcomes with novel neutral outcomes—a procedure we have referred to as "novelty-facilitated extinction." Our first behavioral study showed that replacing an expected shock during extinction with a neutral unexpected tone reduces spontaneous recovery tested at 24-hours in healthy adults and in rats [101]. These behavioral results have been replicated [102,103], including during fMRI [54] where we

used computational modeling to show that a learning parameter that indexes surprise strength and dynamically modulates learning rates ("associability") characterized physiological arousal and vmPFC activity during novelty-facilitated extinction. Further, subjects who showed faster within-session updating of associability during enhanced extinction also expressed less return of fear the next day. Finally, we found enhanced connectivity between the amygdala and vmPFC during recall of novelty-facilitated extinction versus standard extinction. The idea that novelty boosts learning is a core concept in associative learning models [28,104], and supported by new evidence using fiber photometry in mice showing that novel stimuli engage the midbrain dopaminergic system to boost Pavlovian conditioning [105]. Importantly, several psychiatric treatments rely on patient's ability to learn new associations to countervail negative thoughts and associations. *Thus, incorporating **novelty** into a treatment protocol might be a straightforward and inexpensive means to boost new learning during therapy*. **Aim 3** extends novelty-facilitated extinction research to a PTSD patients and integrates MVPA and computational modeling approaches to answer three important new questions: First, does novelty-facilitated extinction reduce extinction retention deficits that characterize PTSD? Second, does enhanced extinction generate robust extinction memories that persist for up to 1 month in healthy adults and in PTSD? Third, does enhanced extinction modulate multivariate neural signature of extinction memory? *This Aim brings advances in the neuroscience of extinction one step closer to clinical translation, and could innovate treatment for disorders of fear and anxiety by incorporating novelty during treatment to enhance new safety learning.*

B. Innovation. This research represents key innovations to our understanding of how the human brain separately encodes, stores, and retrieves conflicting memories of fear and extinction, and has implications for advancing treatment for comorbid PTSD for which extinction deficits are a clinical endophenotype. This research addresses a fundamentally important but as yet unexplored question in neuroimaging of fear extinction—how is an extinction memory represented in the human brain, and is the memory trace stable and separate from the fear memory? A neurobiological model is forming, based on research in rodents, whereby extinction is represented in separate neural populations and connections between the amygdala, hippocampus, and distinct subregions of the mPFC. These regions balance between expressions of fear or extinction based on the context at test. But whether this model applies to humans is largely unknown due to a substantial cross-species translational gap in our ability to quantify and track extinction memory traces over time in the human brain. This work pushes the field forward by addressing this translational gap. We address shortcomings of conventional neuroimaging techniques applied to the study of human fear extinction by using cutting-edge computational approaches that allow us to quantify reactivation of an extinction memory over time. This represents the first effort to combine these tools to localize distinct representations of fear and extinction in the human brain and compare these representations to a patient population for whom dysregulated extinction learning and contextual processing deficits are well characterized. Thus, the proposed work is innovative not only in the question it is designed to address, but in melding cutting-edge neuroimaging techniques to do so.

C. Approach. Since the earliest studies of Classical Conditioning [13], it has been widely appreciated that extinction is new learning, not the unlearning of a previous association. Learning theory proposes—and research confirms—that extinction is a weak and contextually-specific memory, and that extinguished behavior returns under a variety of circumstances [8,31,83]. Animal neuroscience reveals connections within and between the amygdala, hippocampus, and mPFC separately maintain long-term memories of fear and extinction. *The organization of fear and extinction memories are in the human brain is mostly unknown.*

C.1. The basic fear conditioning/extinction protocol. PI Dunsmoor is an early stage independent investigator who has published over 40 behavioral and neuroimaging studies of human fear conditioning in healthy and patient populations. Each study employs a discrimination, partial reinforcement, delay fear conditioning design, with CS+'s that co-terminate 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. During conditioning, the CS+ is paired with shock on half the trials. Partial reinforcement is used to delay extinction [107,108]. *Relevant to **Aim 1**, we consistently find that CS+ items unpaired with shock are remembered at the same level as paired CS+ items paired with shock [73-76,78]—in other words, enhanced memory for CS+ generalizes regardless of whether a specific item is reinforced.* In all studies the CS is 6 seconds in duration with an inter-trial interval of 10±2 seconds.

C.2. Psychophysiology and Shock: SCR will be acquired from the palm of the non-dominant hand using a BIOPAC System and analyzed using criteria established by PI Dunsmoor e.g., [46,109]. SCR gauges phasic

increases in autonomic arousal induced by sympathetic nervous system activity, and has served as a concomitant of human conditioning for nearly a century. The US is a 10 millisecond shock delivered to the right wrist using a BIOPAC device, and calibrated to each subject's tolerance level in accordance with the IRB.

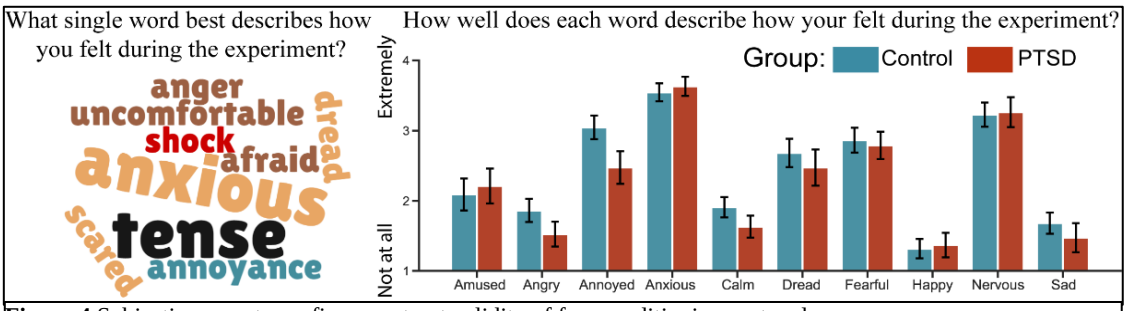

Figure 4 Subjective reports confirm construct validity of fear conditioning protocol.

C.3. Is it fear or threat? We acknowledge that shock used in human research is not intended to induce levels of intense emotional distress akin to a PTSD-like experience. However, self-reported descriptions for how subjects' felt during fear conditioning in our R00 research (**Fig 4**) confirms that every subject described the fear conditioning experience with a construct synonymous with fear (e.g., anxious, dread), supporting the **construct validity** of the conditioning design. In short, the shock in these experiments is sufficient to induce conditioned behavioral responses and brain activity. And while the sensation is not intended to be painful, the dread and anticipation induced by shock is described nearly universally with terms like "afraid" and "scared."

C.4. Participant characteristics and recruitment. *We seek to enroll an equal number of men and women in every group*. Members of the research team include experts in PTSD research and treatment trials. For all studies, a urine toxicology screen is administered to screen drug use. We plan to enroll 120 individuals with PTSD and 120 healthy adults aged 18-50. We match PTSD and healthy subjects on age, sex, and IQ. Participants with PTSD are eligible if they meet DSM-5 diagnostic criteria as assessed using the SCID-5 and the CAPS-5. This is 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. The specific form of trauma is not considered for inclusion/exclusion. We do not exclude for concurrent psychotherapy and evidence based PTSD treatment. ***Medication:*** The use of SSRIs or other antidepressants are not exclusionary, as long as subjects have been on a stable dose and regimen for at least 4 weeks. Although antidepressant medication presents certain challenges for fMRI research, there is good reason not to limit this investigation to an un-medicated sample only [100] as, for instance, antidepressant medication is common in this patient group and our results should be generalizable. Benzodiazepine use and moderate to severe cannabis use disorder is exclusionary. ***Recruitment:*** Subjects make a baseline screening and diagnostic interview, back-to-back MRI visits, and a 1-month MRI visit. *We anticipate no problem with adequate recruitment in this timeline.* Two recruitment sites for PTSD will be the outpatient Behavioral Health Care Services at Ascension Seton Shoal Creek outpatient clinic, and Integral Care Counseling & Mental Health (**see *Letters of Support***). To maintain subjects for the 1-month follow-up visit, we adopt methods from longitudinal research to minimize attrition [101], which includes periodic emails, text messages, and reminder phone calls. The PI has successfully recruited and retained > 40 patients with PTSD for fMRI experiments in less than 2 years at UT Austin (R00MH106719).

C.5. Participants and Power Analysis. All fMRI analyses apply rigorous statistical control for multiple comparisons [112]. Planned enrollment includes 40 subjects per group in each experiment described below. This proposed sample sizes were determined in order to detect statistically reliable fMRI results that survive correction for multiple comparisons [113,114], and to account for attrition of ~3-5 subjects per group over the multi-day experiments. In addition, it is sufficiently powered to detect associations between MVPA signature of context reinstatement in the PPA and fMRI-BOLD activity in canonical extinction neurocircuitry including the vmPFC, in healthy adults based on our R00 preliminary data. First, a power analysis (pwr library for R [115]) on the large effect size (d = 1.37) observed in 25 healthy fMRI participants from the preliminary data estimates 99% power to detect fear acquisition in the dACC (alpha = .05, t-test of parameter estimates from the CS+ > CS- contrast) with as few as 12 subjects. Thus, the proposed sample sizes are more than sufficient to detect successful *fear conditioning* in a canonical fear acquisition region (important to establish). Second, MVPA classification accuracy for decoding scenes in the PPA was also well above chance and yielded a large effect size (d = 2.41). Power analyses

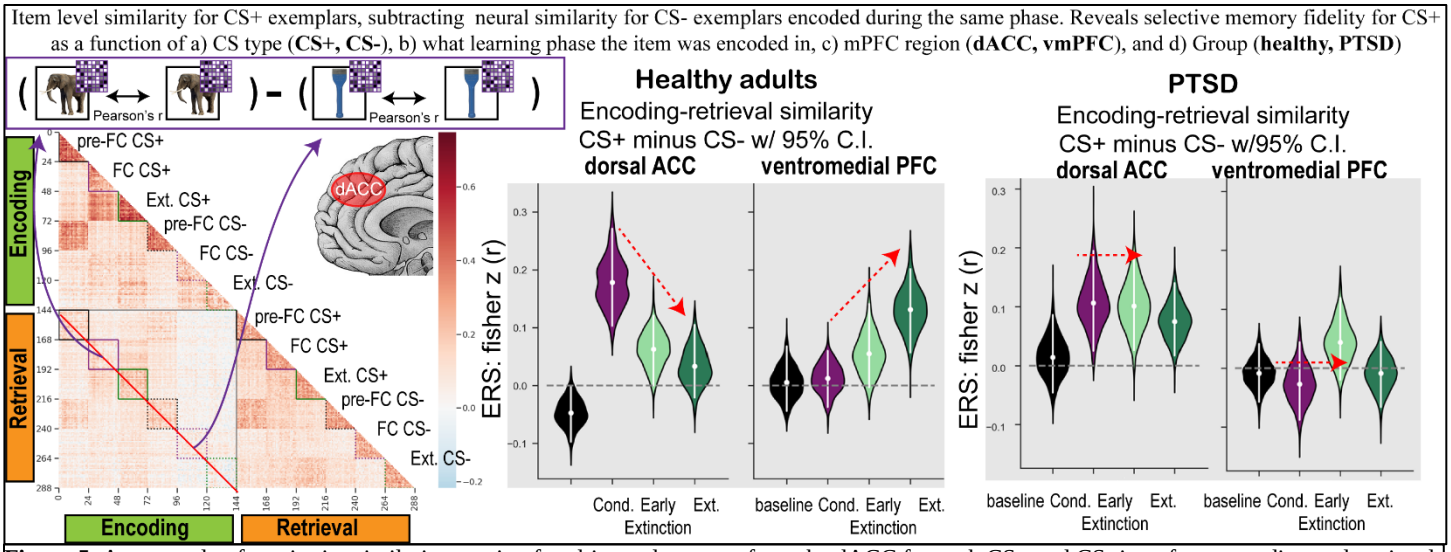

Figure 5. An example of a pairwise similarity matrix of multi-voxel patterns from the dACC for each CS+ and CS- item from encoding and retrieval, ordered by which phase items were encoded. Healthy adults show a double dissociation of selective (CS+ minus CS-) item-level encoding-retrieval similarity in dACC and vmPFC for fear conditioning and extinction. This pattern is not seen in PTSD, such that dACC is selective for CS+ from both conditioning and extinction, and no selective ERS in vmPFC for extinction. Error bars reflect 95% confidence intervals.

estimates 95% power to decode scenes using our MVPA classifier with as few as 12 subjects, highlighting the robust nature of the MVPA analysis. And finally, power analysis on the regression analysis using MVPA derived extinction context showed 35 subjects will yield a large effect size (d = .674).

C.5. Aim 1: <u>Identify stable and distinct memory traces of fear vs. extinction over time</u>. We plan to use MVPA to quantify *item-level* patterns of encoding-retrieval similarity (ERS) in separate brain regions as a function of whether items were encoded during conditioning or extinction. We plan to track neural representations ove time and compare the fidelity of the memory trace between healthy adults and patients with PTSD. ***<u>Task Design</u>.*** On Day 1, subjects encode pictures of animals and tools in a baseline phase without any shocks, followed by fear conditioning wherein pictures of animals and tools serve as CS+/CS- (category counterbalanced between subjects). Following fear conditioning, subjects see different pictures of animals and tools without shock during extinction. Note that each picture is a different exemplar (e.g., there are not two different pictures of a chimpanzee). Subjects return 24-hours (and 1-month later) for a recognition memory test comprised of half the exemplars encoded from conditioning and extinction the previous day and an equal number of novel category exemplars to account for false alarms (half at 24-hr and half at 1-month). Subjects make old/new recognition memory judgements with confidence ratings. *<u>As in our prior work</u>* [46,73,74,78,116], *<u>the use of separate object categories allows us to test memory for items related to the CS+ versus a control category (CS-) in a within-subjects design</u>*. By combining fear conditioning with episodic learning, each trial is effectively isolated as a single learning episode encoded at a specific moment in time. Therefore, at a recognition memory test, we can measure episodic memory for each trial as a function of *when* that trial was encoded. This allows us to use estimate ERS [82] using the factors of condition (CS+, CS-), phase (baseline, conditioning, extinction), retention interval (1-day, 2-week), group (healthy, PTSD) and memory performance (remembered, forgotten). ***<u>Imaging analysis plan and Item-level similarity</u>*.** We use representational similarity analysis (RSA) [23] to look for similarity patterns of each CS+ exemplar from baseline, conditioning, and extinction, and subtract similarity patterns from corresponding CS- items encoded during the same phase (**Fig 5**). Controlling for pattern similarity to the CS- is important, because it helps ensure that heightened pattern similarity to the CS+ in a given region would be a result of associative learning, and not simply because subjects are viewing the same image across days. Pattern similarity is estimated using pairwise Pearson correlations (transformed to Fisher's Z score) for each item from encoding-to-retrieval. The ROI approach focuses on regions of the highest interest from animal models of dissociable fear and extinction memory: the dACC, vmPFC, amygdala, and hippocampus. A subsequent searchlight analysis, corrected for whole-brain multiple comparisons, will be used to reveal other areas of selective CS+ > CS- item similarity using the factors of interest (e.g., encoding phase). ***<u>Preliminary data</u>.*** Data from 24 healthy adults 24 individuals with PTSD symptoms show better memory for CS+ vs. CS- encoded before, during, and after conditioning, replicating our prior behavioral findings (**Fig 2**). We also show that we can use RSA to detect separate encoding-retrieval similarity (ERS) that is selective to the CS+ items encoded during conditioning and extinction (**Fig 5**). Specifically, ERS for CS+ versus CS- items is elevated in dACC for items

encoded during conditioning, and elevated in vmPFC for items encoded during extinction. *Interestingly, these results were* ***not*** *driven by explicit memory, suggesting that selectivity in ERS for CS+ items is not determined by explicit memory processes* (**Fig 6**). The PTSD group showed another pattern of results, whereby ERS in dACC was selective for CS+ encoded during both conditioning and extinction, with no selectivity in vmPFC for CS+. Remarkably, when data was split as a function of memory (Remembered vs. forgotten), PTSD patients did show elevated ERS in the vmPFC for CS+ items encoded during extinction that were later remembered (**Fig 6**). These results show early signs of a neural organization consistent with animal models of dissociable representations of fear and extinction memory mPFC regions, and indicate that this neural organization is abnormal in PTSD. Aim 1 seeks to hone our design and investigate neural representation of fear versus extinction at remote time points of 1 month. ***Potential Problems and Alternative Strategies*****:** We first note that this investigation will yield new insight into long-term representation of fear vs. extinction specific memory in healthy adults and PTSD whether or not our specific hypotheses are correct. Also, **Aim 2** can proceed regardless of the results from **Aim 1**. If we do fail to see differentiation in representational similarity between conditioning and extinction memories, then an alternative strategy would be to induce stronger differentiation between the fear and extinction *contexts*. One way to do this is to manipulate the context of encoding by presenting CS trials on a different background image cf. [53]. Another way to enhance differentiation between memory traces is to separate encoding into two sessions separated by at least 24-hours.

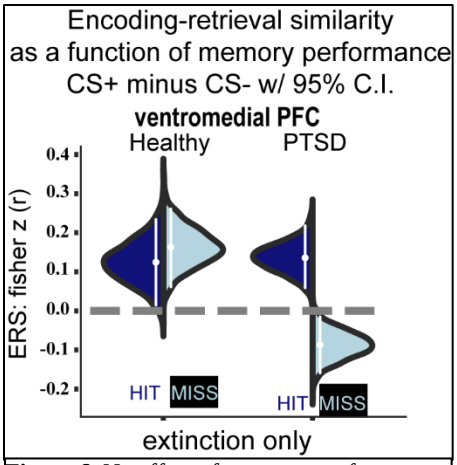

Figure 6. No effect of memory performance (remembered vs. forgotten) on ERS in vmPFC for healthy adults. But PTSD subjects do show selectively enhanced ERS for CS+ vs CS- items in vmPFC for remembered items, suggesting a role of episodic encoding on neural organization in PTSD.

C.6. Aim 2: Decode contextual reactivation of fear and extinction memory. The encoding specificity principle [59] states that memory recall is facilitated when the context of retrieval matches the context of encoding. 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 [93]. More recently, computational models of episodic memory have formalized how mental context retrieval guides temporal order memory [60]. These models have been leveraged on multivariate fMRI data to decode the mental context where an item was previously encoded [97,99]. **Aim 2** synthesizes theoretical and empirical advances on how memories are linked to the mental context where they were formed in order to derive a novel neural signature of fear and extinction memory reactivation. Based on extensive work on the contextual-specificity of extinction recall [117], we predict that the balance of neural reactivation of the fear or extinction context controls the behavioral expression of fear upon test. ***Task.*** We have developed a protocol to separately tag the context of conditioning and extinction during fMRI in healthy adults (**Fig. 3,7,8**), and then decode the relative strength of spontaneous neural reactivation of these contexts at test. Subjects undergo conditioning and extinction on Day 1, followed by a fear renewal test 24-hours later. The CSs are three discriminable color square—green, blue, and yellow—counterbalanced between subjects as two CS+'s and a CS-. Both CS+'s are paired with shock during fear conditioning, but only one CS+ is extinguished. Both CS+'s are then represented at the 24-hour and 1-month "fear renewal" test. ***Neural context tag:*** The key feature of the task design is how we define "context." Throughout extinction, between each CS trial (i.e., during the intertrial interval) subjects are presented with a stream of ***task-irrelevant*** pictures of scenes. This series of category-specific pictures make up the "context." Each task-irrelevant scene picture is presented for 800 ms and separated by 200 ms blank screen. The rationale

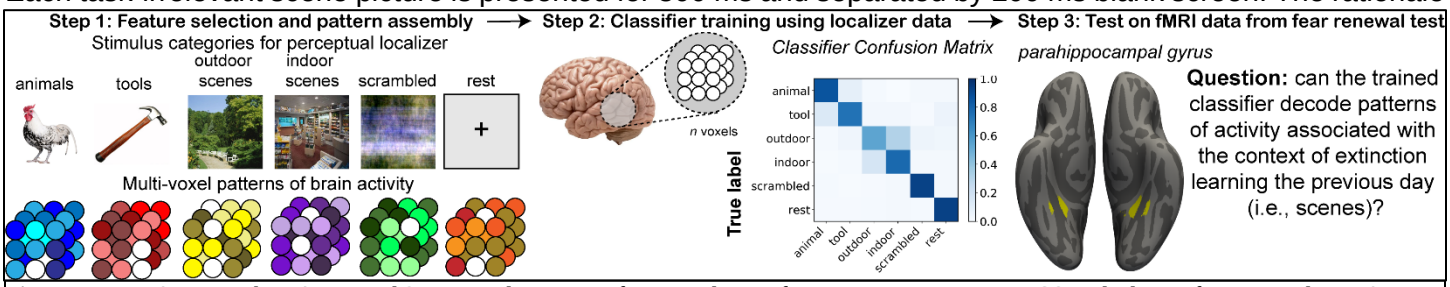

Figure 7. MVPA approach to Aim 2 and 3 to covertly measure fMRI evidence of context reinstatement on CS trials during fear renewal test. A perceptual localizer is used to train pattern classifiers, and used to track mental context reinstatement (in this case, scene patterns in PPA) at 24-hour fear renewal (i.e., extinction memory retrieval) test in the absence of any scene presentations.

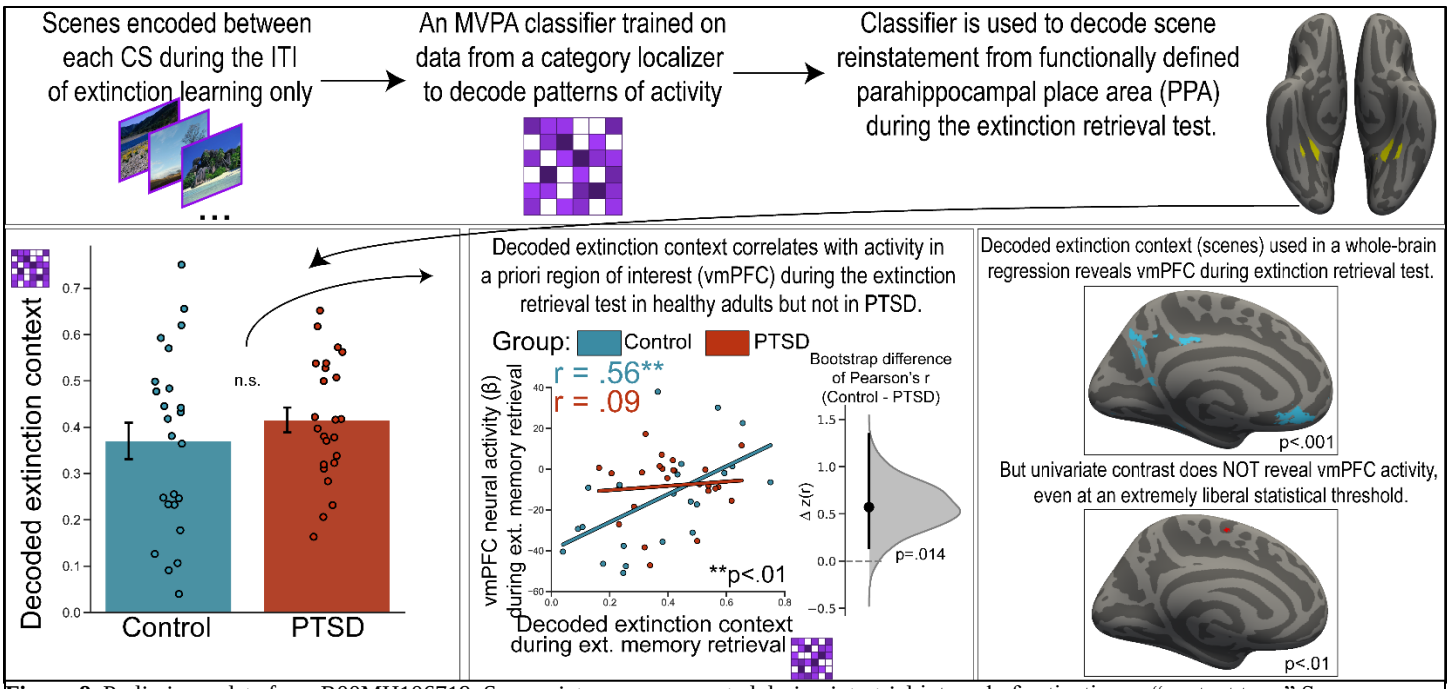

Figure 8. Preliminary data from R00MH106719. Scene pictures are presented during intertrial interval of extinction as "context tags." Scene evidence at test decoded from PPA used as covert measure of neural context reinstatement. The degree of scene reinstatement correlates with activity in vmPFC in healthy adults, but not PTSD. Aim 2 hones this task design in important ways, and tests reinstatement at remote time point.

for presenting a stream of different scene pictures throughout the scan, rather than presenting a single static background image, is to avoid habituation and ensure that the context information is consistently activating category-selective visual cortex. This should produce a robust "mental context tag" that we can decode later using fMRI pattern classification if and when that context is reactivated. Subjects will undergo a series of functional localizer scans at the end of the experiment in order to train pattern classifiers to recognize multivariate activity patterns unique to the context tag category (scenes). Additional categories (e.g., animals, tools, scrambled scenes, and rest) are used in the category localizer to fine tune the precision of the trained classifiers. *This design is informed by innovative work on neural context tagging in human memory* [97,99,118]. The critical test of this design is the 24-hour and 1-month fear renewal test which occurs in a novel context that does not match either conditioning or extinction. In this test, the subjects will see all the CSs again, but the ITI is blank and there are no scenes ever presented. This is analogous to an ABC contextual renewal design in Pavlovian conditioning [119]. Because extinction was tagged with pictures of scenes, we can use the trained fMRI classifiers to quantify the neural reinstatement of the extinction context by classifying activity in the *parahippocampal place area (PPA)* associated with processing images from that stimulus category [22]. This allows us to relate neural context reinstatement to behavioral expression. As such, we simultaneously collect SCRs and explicit ratings of threat expectations to link the index of neural context reinstatement to behavioral variables. ***fMRI Analysis: Using MVPA to detect neural context reinstatement:*** To classify the mental context tags for conditioning and extinction, we will conduct two 6-minute localizer scans after the conditioning/extinction session. MVPA of localizer data from the whole brain is used to train category-level classifiers for each subject. After verifying successful classification using cross-validation testing within the localizer data, data from all runs of the localizer task will be combined to train category-level classifiers, per subject, and applied to data from the Day 2 extinction-

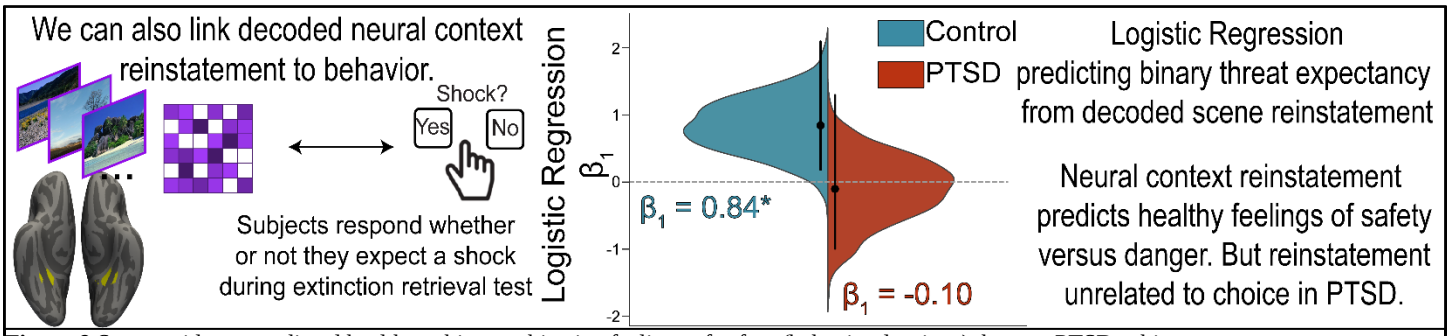

Figure 9 Scene evidence predicted healthy subjects subjective feelings of safety (behavioral ratings), but not PTSD subjects.

recall test. ***Preliminary Data* (Fig 8)*:*** Scene evidence in PPA during extinction-recall correlated with vmPFC and hippocampal activity in healthy controls, but not PTSD. The amount of decoded scene activity predicted subjects' feelings of safety versus danger in healthy controls, but not PTSD (**Fig 9**). Thus, preliminary results provide important foundations for the proposed research to investigate the effects of alcohol use on fear extinction processes. ***fMRI Predictions*:** Based on strong preliminary results, we predict MVPA classifiers will show neural context reinstatement of multivoxel patterns in the PPA associated with the extinction context (scene-related activity patterns) during a 24-hour fear renewal session in healthy adults that correlates with activity in the vmPFC. We also predict an inverse relationship between the *strength* of the extinction-context classifier throughout the fear renewal session and the *strength* of autonomic arousal on CS+ (extinguished) trials versus CS+ (unextinguished) trials. That is, subjects who successfully reinstate the memory for the extinction context will be those who show the best behavioral evidence of diminished arousal and shock expectancy at test. Further, we expect that at 1-month that neural reinstatement of the extinction context will diminish in healthy adults, consistent with the idea that extinction memories are transient and the fear memory recovers over time even in the healthy brain. ***Potential Problems and Alternative Strategies*:** There are important differences between immediate versus delayed extinction [124]. An "immediate extinction deficit" would propose that neural reactivation of the extinction context is weaker when extinction immediately follows conditioning. A future goal is to characterize how extinction memories are organized when extinction occurs at different time intervals.

C.7. Aim 3: Does strengthening extinction learning enhance integrity of extinction memory representations? Extinction is a weak and impermanent form of new inhibitory learning, and a number of factors contribute to relapse of the original fear behavior, even in healthy adults. The PI has investigated new behavioral strategies to enhance extinction by replacing aversive outcomes with novel events, referred to as novelty-facilitated extinction [54,101]. In **Aim 3**, we adapt MVPA approaches to investigate whether enhanced extinction modulates multi-voxel patterns associated with extinction memory retrieval over time. **Aim 3** builds on K99R00 funded research and the design of Aim 2. We include a healthy and PTSD group who undergo identical fear conditioning procedures as in Aim 2, but then undergo an augmented form of fear extinction in which the shock is replaced (rather than merely omitted). Regardless of our specific hypotheses for Aim 2, Aim 3 represents the first between-groups multi-day neuroimaging investigation in PTSD to compare standard versus enhanced extinction. We also evaluate long-term extinction memory at 1 month, which is extremely rare in human research, but consistent with diagnostic criteria for assessing PTSD (Criterion F: symptoms persist for > 1 month), and thus furthers the bridge to translational relevance. ***Task*.** The task design is similar to Aim 2, and includes a fear conditioning phase with 2 CS+'s and 1 CS-. Extinction includes one CS+, but whereas in Aim 2 the CS+ was presented without shock, in Aim 3 the shock is omitted and **replaced** by a neutral, low-volume, 500 millisecond pure tone (440 Hz) presented through MRI compatible headphones. We have found that replacing shocks with a neutral tone accelerates extinction learning and helps prevent the return of fear in humans and rodents, and this finding has been replicated in other labs as well [102,103]. Extinction on Day 1 also include context tags (scene pictures) presented throughout the intertrial interval during extinction only. The 24-hour and 1-month extinction recall/fear renewal test is identical to that described in Aim 2, and includes both the extinguished and unextinguished CS+ and CS- and no tones and no scene pictures during the ITI. ***Computational modeling of extinction learning*.** Associative learning models describe extinction as new learning generated by the surprising omission of the US [31]. A popular computational learning model, the Pearce-Hall model [28], proposes that omission of the US governs the rate and effectiveness of extinction by modulating a property of the CS known as *associability*. As it pertains to extinction, the core feature of associability is to determine how well the CS forms a new association whereby the CS no longer signals the US ("CS-no US" association). Associability is dynamically (trial-by-trial) determined by the unsigned (absolute value) prediction error on the previous trial. We have recently applied this modeling approach (Dunsmoor et al., 2019, *Journal of Neuroscience*) to assess the neurocomputational mechanisms of extinction in healthy adults during fMRI. This same computational model has been used in other human neuroimaging fear conditioning studies in PTSD [62] and healthy volunteers [125 126] in the context of fear reversal. *In Aim 2 and 3, we will use computational modeling to test the hypothesis that novelty facilitates associability-modulated within-session fear extinction in the vmPFC, as well as prediction error related activity in mesolimbic dopaminergic systems.* ***Preliminary Data*:** We have shown that replacing shocks with a surprising and neutral stimulus can accelerate extinction and prevent the return of fear. Novelty-facilitated extinction evoked heightened vmPFC on CS+ trials during extinction and 24-hour extinction recall as compared to standard extinction (**Fig 10**). ***Neural questions and predictions:*** It is well established that the vmPFC is functionally critical to fear-extinction in rodents. But the precise role of the vmPFC on long-term extinction memory retrieval is unclear. Prior human neuroimaging research shows vmPFC and hippocampal activity is

correlated with successful extinction retrieval after one day [52,53]; but evidence at much longer time points is lacking. Recent evidence in rats has challenged earlier suggestions that the vmPFC is the site of long-term extinction memory storage [127]. Importantly, projections between the vmPFC and amygdala orchestrate the balance between expressions of threat and extinction memories over time [128]. One possibility is that extinction memories continue to undergo plasticity and migrate beyond the vmPFC over time, perhaps relying on basolateral amygdala or thalamic networks for long-term extinction memory retrieval. Interestingly, human episodic memory research shows that the hippocampus is critical for retrieval of newly formed memories, but retrieval gradually depends on the medial PFC over time [129]. How extinction memories in humans are maintained and retrieved by these brain regions over long periods of time, and whether this network can be modulated by strengthening extinction learning, is unknown. We predict enhanced extinction will extend to 1 month in PTSD, as evidenced by diminished autonomic physiological arousal (i.e., sweating) to the CS+ in the enhanced versus standard-extinction group. Based on animal models suggesting neural reorganization between vmPFC-amygdala networks after extinction [127], we predict that extinction memories will be maintained through relatively stronger activation in the amygdala, and inhibitory connections between the basolateral amygdala and the vmPFC, over time. This prediction can be extrapolated from our findings in healthy adults at 24-hour testing. ***Potential problems and alternative strategies*:** It is possible that this technique is not robust in PTSD and effects will not persist to 1 month. If this is the case, it provides motivation for a more intense intervention to strengthen extinction, perhaps through a combination of learning theory based strategies designed to increase inhibitory regulation [130]. Importantly, because data on remote extinction-retrieval in human neuroimaging is lacking, *results provide important insights on neural mechanisms associated with success or failure of extinction-retention over long time intervals regardless of our hypotheses*. ***Significance*:** Extinction strategies are therapeutically relevant only insofar as the effects persist well after the extinction session. The effects of standard extinction tend to be time limited, and extinction-retention deficits in PTSD are well-documented [2]. Abnormalities in extinction neurocircuitry are associated with severe deficits in extinction-retention in PTSD, and may be a factor in relapse after clinical treatment. Enhanced extinction might compensate for extinction-retention deficits in PTSD, leading to persistent diminution of maladaptive defensive behavior long after the extinction session. We predict that strengthening extinction learning by maximizing surprise [101] will extend the extinction-retention window in PTSD beyond 24-hours. This research is significant because it harnesses emerging behavioral, neuroscience, and theoretical knowledge on learning and memory that may ultimately contribute to innovative clinical treatments. ***Interpretation and Future Directions*:** If this procedure helps prevent the return of fear in PTSD, it has far-reaching implications for how to override maladaptive associative learning, and provides a simple technique that could be straightforwardly adapted and implemented as a therapeutic tool in the future. If we fail to replicate prior findings (Dunsmoor et al., *J Neuro*, 2019), we will assess whether any parametric changes in the task design may have contributed.

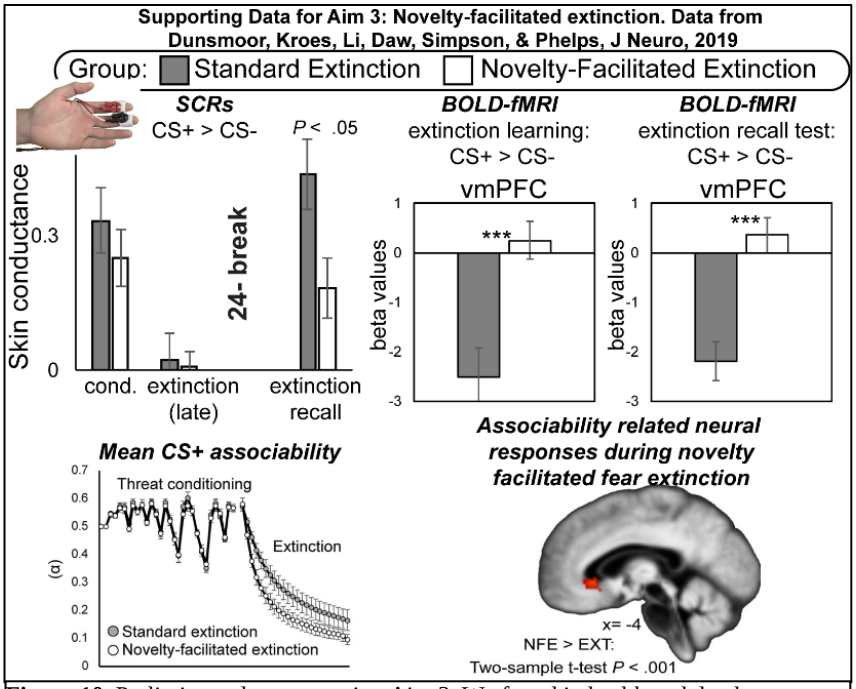

Figure 10. Preliminary data supporting Aim 3. We found in healthy adults that replacing a shock with a neutral tone facilitated extinction learning and helped prevent return of fear, compared to simply omitting shock entirely. Computational modeling showed that vmPFC activity tracked a parameter indexing surprise strength (associability). Aim 3 hones this design, extends the task to PTSD, combines it with MVPA approach from Aim 2, and tests strength of novelty-facilitate extinction at 1 month.

GENERAL IMAGING PROTOCOL AND ANALYSIS*:* Whole brain fMRI data will be acquired using a 3T Siemens Vida MRI scanner with a 64-channel head coil at UT Austin. Multivariate analyses (MVPA, RSA) will be conducted using a combination of tools including the Princeton MVPA toolbox in MATLAB, the scikit-learn toolbox in Python, and custom in-house software.