Title: Remember to forget what you fear

Subtitle: A behavioral and neurocognitive search for interventions that persistently attenuate threat memories

Contents

Chapter 1 - General Introduction

To thrive in dynamic environments, we need to adapt our emotional responses to our circumstances. When we encounter a threatening situation, we are quick to learn to fear similar situations. But it is equally important to update our memory when these events no longer pose a threat. If we fail to learn that a previously threatening situation is now safe, the threat memory becomes maladaptive and could disrupt daily functioning. Consider the COVID-19 pandemic. During the first lockdown, we learned to fear crowded environments. This threat response was adaptive, as it helped us to avoid the danger of contracting COVID-19. However, once the virus has become endemic and we have built up sufficient levels of immunity, it will be adaptive to learn that crowded places were safe again. If the virus no longer poses a serious threat and regular on-site work and activities can be resumed, fear of crowded places is no longer adaptive and may disrupt our daily functioning. For example, if a regular work-day, consisting of a commute on public transport, a work-day in an open office plan and a lunch with colleagues, feels like a series of threats while the risks are in fact negligible, the threat responses may have become maladaptive. Yet, while we quickly acquire threatening associations, unlearning them can be more challenging. For over a century, research has studied how we can change threat responses to prior threats and has given considerable insight into the optimal ways to reduce these. Although we understand how we can learn that previously threatening situations are now safe, our ability to retain this new safety memory appears to be weaker than our ability to retain the original threat memory, often resulting in relapse of threat responses. To prevent relapse, we may want to attenuate the original threat memory in a more persistent manner. While this could be an effective solution, it may feel like an undesirable treatment to change our memories, as they are closely intertwined with our identity. If you are suffering from relapses of a fear for crowded places years from now because of a COVID-19 pandemic that lies in the past, would you want to take a drug that alters your memory about that pandemic?

The standard treatment for maladaptive threat responses, i.e., exposure therapy, consists of controlled forms of (often imaginary) exposure to the feared situation in a safe, therapeutic setting. This allows fearful associations to subside, while novel safety learning occurs, resulting in an attenuation of threat responses. Unfortunately, the original threat memory often resurfaces, leading to a relapse of threat responses. **The general aim of this thesis is to investigate whether we can identify novel forms of safety learning that can prevent the recovery of threat responses after initial safety learning.** To this end, we created threat associations in a controlled setting by pairing specific events with uncomfortable electric shocks and explored whether we could alter an experimental model for classic exposure therapy in ways that prevented relapse of threat responses. While eventually we would like to be able to attenuate maladaptive threat memories, we also ask whether

the public has a positive opinion towards treatments that alter the original memory. Before I describe these studies in more detail, I will provide a neurocognitive background of emotional memory and threat responses, and outline experimental models used to study their mechanisms. Thereafter, I will discuss different intervention opportunities that may allow us to persistently attenuate threat responses. At the end of this introduction, I will introduce the next chapters of this thesis.

Memory for threatening situations

Emotional enhancement of memory

Memory allows us to store past experiences to adaptively guide our future behaviour. But not everything we experience is equally important to remember, and this is where emotions come into play. Memory systems prioritize the retention of emotionally salient events (LaBar & Cabeza, 2006). In threatening, stressful situations, memory is strengthened for aspects of a situation that are directly related to threats and for events that occurred in the same context immediately preceding the threat, enabling us to recognize future threats in advance so that we can avoid dangerous situations (Schields et al., 2017). To further characterize the effect of emotion on memory, it is helpful to dissect emotions into two orthogonal components: valence, the dimension that varies from pleasant (positive) to unpleasant (negative) with neutral as intermediate value, and arousal, that varies from calm to excitement. Specifically, irrespectively of valence, emotionally arousing events show an emotional advantage in memory that increases over the course of days or weeks (Kleinsmith & Kaplan, 1963; LaBar & Phelps, 1998; Sharot & Phelps, 2004). Emotional memory research also typically makes a distinction between two forms of emotional memory: explicit (declarative) and implicit (non-declarative) memory (Squire & Zola, 1996). Explicit memories are memories to which we have direct conscious access and can be further sub-divided into episodic memories and semantic memories. Episodic memory refers to our ability to remember experiences and provide contextualized recollections of events, including knowledge of time, place or other details, while semantic memory consists of factual world knowledge (Tulving, 2002). Implicit memories, on the other hand, are a type of memories that we do not directly recollect consciously, and include conditioned memories, referring to our ability to associate neutral stimuli with (emotionally salient) outcomes or behaviours (Squire & Zola, 1996). Returning to the example of fear for crowded places after COVID-19, both conditioned, implicit emotional memories and explicit episodic memories could play a role. The introduction of heavy fines for gatherings of more than three people could have created implicit associations between gatherings and financial threats, while we may also have vivid, episodic recollections of that one stressful supermarket visit at peak hours where it was impossible to keep enough distance.

Memory consolidation

As we experience the world around us, we briefly hold sensory information in our sensory memory. Depending on what part of the information we attend to, some of this information enters into short-term memory, where it may be held for seconds to hours. These new memories are initially sensitive to disruption, but can develop into more stable, long-term memories over time in a process known as consolidation (Squire & Alvarez, 1995). While short-term memories are thought to be mediated by fast but short-lived synaptic plasticity that only requires modification of existing synaptic proteins, consolidation into long-term memories is thought to require gene transcription and the synthesis of new proteins, which may take place over the course of hours (Johansen et al., 2011). The slow consolidation of memories allows experience to retroactively strengthen memory, for instance strengthening our memory for a series of events that was initially neutral but turned out to lead up to threat (McGaugh, 2000). In case of emotionally arousing experiences, the stress hormones (nor)epinephrine and cortisol are released and can enhance memory for the experience, either immediately during encoding or retroactively during the consolidation window (Gold & Van Buskirk, 1975; Lupien & McEwen, 1997; Sandi & Rose, 1994).

Experimental paradigms for fearful memories

The most commonly used paradigm to study aversive emotional memories models implicit, conditioned memories and is known as fear conditioning (LeDoux, 2009). In one of the earliest fear conditioning experiments in humans, an 11-month old infant named Albert was conditioned to fear a rat after simultaneous presentation of the rat together with a loud noise that frightened him (Watson & Rayner, 1920). Thus, during the acquisition phase of a fear conditioning experiment, a neutral stimulus (e.g. a rat), termed the conditioned stimulus (CS), is paired with an aversive stimulus (e.g. loud noise), termed the unconditioned stimulus (US). As the CS-US association is learned, presentation of the CS by itself will evoke a conditioned response (CR) that is similar to the threat response originally evoked by the US. Given that most work on “fear” conditioning in fact investigates the detection of and response to threat, and it does not measure subjective feelings of fear, it may be more precise to refer to it as threat conditioning instead of fear conditioning (LeDoux, 2014).

The classic form of Pavlovian threat conditioning is known more specifically as cued threat conditioning, as the CS is a discrete cue in the environment, such as a sound or a specific picture. While cued threat conditioning is used to model implicit conditioned memories, variations of the classic threat conditioning paradigm can also be used to study different types of emotional memories. Category conditioning is a variation on cued threat conditioning in which the CS spans a semantic category (e.g. pictures of different animals) and is thought to rely on explicit semantic memory to mediate the association between items of a category and the US (Dunsmoor et al., 2014). While cued

threat conditioning is thought to rely exclusively on implicit memory, category conditioning additionally allows us to study how emotional associations affect episodic memory (Dunsmoor & Kroes, 2019). After the acquisition phase of a category conditioning experiment, item recognition can be measured for the individual (non-repeating) category exemplars from the conditioned category (CS+, partially reinforced with shocks) and the unconditioned category (CS-, never reinforced). After controlling for memory specificity using novel exemplars, enhanced recognition of the (unreinforced) CS+ exemplars presented during acquisition indicates emotional enhancement of memory for the conditioned category (Dunsmoor et al., 2014).

Another variation of the classic threat conditioning paradigm is contextual threat conditioning. In contextual conditioning the onset of the US cannot be predicted based on an isolated cue in the environment, as is the case in cued and category conditioning. Instead, during contextual conditioning, the US is predicted by the context, such as a specific room or location, that consists of a specific configuration of multiple elements. In contrast to cued threat conditioning, that appears to be a direct form of associative learning that associates the sensory qualities of the CS and US, contextual threat conditioning seems to be a two-step process, that first requires encoding of the context that then facilitates associative learning between the context and the US (Maren et al., 2013).

Once conditioned threat responses have been acquired, presentation of the CS alone evokes a CR. However, upon repeated presentation of the CS in absence of the US, the CRs may diminish and disappear in a phenomenon known as extinction. Extinction learning can be used as an experimental model for safety learning. However, according to the mainstream view, extinction is thought not to modify the original threat memory, but is instead considered a form of new learning (Bouton, 2004; Vervliet et al., 2013). As a result, threat responses are sensitive to relapse that can be demonstrated through four phenomena (Bouton, 2002, 2004). First, extinction learning is context-specific, and a change of context after extinction can increase threat responses, known as *renewal*. Second, *spontaneous recovery* of threat responses can be observed with the passage of time after extinction. Third, threat responses to the CS can be *reinstated* by un-paired presentations of the US. Finally, *rapid reacquisition* shows that new CS-US pairings introduced after extinction result in faster acquisition of CRs compared to the initial acquisition. These four phenomena, renewal, spontaneous recovery, reinstatement and rapid reacquisition, can be used in experimental settings to assess to what extent different forms of extinction learning can persistently supress threat responses.

Physiological correlates of fear

When we are confronted with a threat, we experience a subjective feeling of fear. We may feel concerned, worried, scared, frightened and so forth. But we also show a set of defensive responses,

including defensive behaviours, autonomic and endocrine responses and startle reflex potentiation (LeDoux, 2009). These threat responses can be measured and used as behavioural or physiological correlates of fear. In turn, although this is a simplification, the strength of the underlying memory trace can reasonably be inferred from the magnitude of the threat responses (Bouton & Moody, 2004). In a Pavlovian conditioning experiment, for example, delivery of a painful shock (the US) paired with an auditory tone (the CS) leads to the formation of a CS-US memory. If presentation of the CS alone evokes a conditioned threat response, we can conclude that there is a learned threat association between the US and the CS, and we infer the strength of this memory trace from the magnitude of the CR.

Defence behaviours are species specific, and differ depending on the proximity of the threat (Fanselow, 1994). In rodents, the most commonly measured response to threat is freezing (Blanchard & Blanchard, 1969; Blanchard et al., 1968). Freezing is an innate defensive response to danger observed in many species that could serve to avoid detection, while optimizing perceptual processes and preparing for fight-or-flight responses (Roelofs, 2017). In addition to measuring behavioural read-outs of threat detection, such as freezing, we can also assess threat responses more directly by measuring physiological output. Upon confrontation with a threat, immediate threat responses are coordinated by the hypothalamus, resulting in a fast activation of the autonomic nervous system (Cannon, 1929). The autonomic nervous system consists of the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). Both the SNS and PNS are activated upon confrontation with threat, and the physiological results depend on which system is dominant (Iwata et al., 1987). The PNS is classically known as the “rest and digest” system and the SNS as the “fight or flight” system, although in contrast to what these names suggest, both systems can be dominantly active and drive behaviour under threat. In response to distal threat, phasic dominance of the PNS drives freezing and heart rate deceleration, which may allow for action preparation and information processing (Livermore et al., 2021; Roelofs, 2017). Imminent threat, on the other hand, drives SNS dominance, activating the sympatho-adrenomedullary system, resulting in release of norepinephrine from a brain region named the locus coeruleus, release of epinephrine from the adrenal medulla, heart rate acceleration and increased blood flow to the muscles (De Kloet et al., 2005; Livermore et al., 2021). Activation of the locus coeruleus is thought to mediate arousal and can be measured indirectly through pupil dilation (Aston-Jones & Cohen, 2005), where greater pupil dilation responses (PDRs) reflect greater arousal (Bradley et al., 2008). Activation of the SNS can also be inferred from skin conductance responses (SCRs) driven by sympathetic activation of sweat glands (Critchley, 2002; Lang et al., 1993; Steckle, 1933). In addition to activation of the autonomic nervous system, detection of threat also leads to startle reflex potentiation (Ledoux, 1997). Startle responses are a defensive reflex in response to sudden sensory events, and are increased in threatening compared to safe contexts (Davis &

Astrachan, 1978), a phenomenon known as fear-potentiated startle (FPS). In humans, fear-potentiated startle responses can be measured in eyeblink responses using electromyography at the orbicularis oculi muscle that closes the eyelid during the startle blink (Lang et al., 1990).

Neural correlates of fear and extinction

From early lesion studies, it has become evident that parts of the temporal lobe are required for the expression of fearful behaviour (Brown & Shafer, 1888). A similar behavioural pattern including an apparent loss of fear, was termed the Klüver-Bucy syndrome, and this was later suggested to result from damage to the amygdala, an almond-shaped structure located in the medial temporal lobe (Klüver & Bucy, 1937). Since these initial findings, classical Pavlovian threat conditioning has been used extensively to study the neural mechanisms underlying the acquisition and extinction of conditioned threat responses. From these experiments, it has become clear that the amygdala plays an essential role in the acquisition and expression of conditioned threat responses (Bechara et al., 1995; Blanchard & Blanchard, 1972; Gentile et al., 1986; Hitchcock & Davis, 1986; Klumpers et al., 2014). The amygdala consists of several interconnected subnuclei, that play distinct roles in threat processing. The basolateral amygdala receives sensory inputs, such as auditory inputs from a tone used as CS and the nociceptive inputs from e.g. a shock used as US (Ledoux, 2003). During threat conditioning, the two sensory inputs (CS- and US-related) converge on synapses in the basolateral amygdala, triggering a plastic change in synaptic strength that likely mediates the threat memory (Ledoux, 2003). In response to CS or threat detection, the basolateral amygdala activates the central nucleus of the amygdala through intra-amygdala connections. In turn, the central nucleus of the amygdala projects to areas of the brainstem to orchestrate the expression of conditioned threat responses, including activation of the autonomic nervous system, hypothalamic-pituitary axis and behavioural defence responses (LeDoux, 2009). The amygdala also appears to play a critical role in the emotional enhancement effect on memory, as activation of β-adrenergic receptors in the amygdala is necessary for the enhancement of memory through arousal-evoked epinephrine and glucocorticoids (Liang et al., 1986; McGaugh, 2000; Roozendaal & McGaugh, 1996). Different sub-types of conditioning may additionally require the involvement of other brain regions. Specifically, while the amygdala is involved in both cued and contextual threat conditioning, conditioning to a context also requires another structure in the medial temporal lobe known as the hippocampus (Phillips & LeDoux, 1992). The amygdala can process simple stimuli, and context conditioning likely requires the hippocampus for identification of the more complex combination of stimuli as a specific context. The hippocampus in turn activates the amygdala to coordinate a threat response (LeDoux, 2009). Studies investigating the neural signature of threat processing in humans have further identified a neural 'fear network' that is consistently and robustly activated during threat conditioning paradigms, and most notably includes the anterior insular cortex,

dorsal anterior cingulate cortex, and ventral striatum (Fullana et al., 2016). Surprisingly, while damage to the amygdalar complex has been shown to be associated with weaker threat conditioning in humans (Klumpers, Morgan, et al., 2015), human neuroimaging studies generally do not reveal threat-related activation of amygdala, possibly due to limited spatial resolution of the imaging methods used in humans (Fullana et al., 2016; Visser et al., 2021). During the acquisition phase, a separate 'safety network' shows increased activation for the safe cues (CS-, unreinforced CS) compared to the cues that signal threat (CS+, reinforced CS) in, among other areas, the ventromedial prefrontal cortex (vmPFC), lateral orbitofrontal cortex, posterior cingulate cortex, and the hippocampus (Fullana et al., 2016). During extinction learning, the CS+ evokes similar activity in e.g. the anterior cingulate cortex, medial prefrontal cortex and insular cortex (Fullana et al., 2018). Work in rodents has further indicated that the vmPFC is required for extinction learning (Morgan et al., 1993). In humans, the vmPFC appears to be deactivated during conditioning, whereas its activation increases during the course of extinction learning (Milad & Quirk, 2012). Based on work in the presumed homologous region of the vmPFC in rodents, the infralimbic prefrontal cortex, it is thought that vmPFC activation during extinction may inhibit activation of the amygdala (Milad & Quirk, 2012).

Intervention opportunities: reconsolidation and extinction

Two distinct strategies for persistent attenuation of conditioned threat memories can be identified in recent research. The first strategy, disrupting reconsolidation, targets the original threat memory directly through the application of amnestic treatments while the memory is in a labile state (see Figure 1.1). The second strategy, enhancing extinction, aims to enhance the consolidation of the novel extinction memory, to provide a stronger and longer lasting inhibition of the threat memory. Below, I will discuss these interventions in more detail.

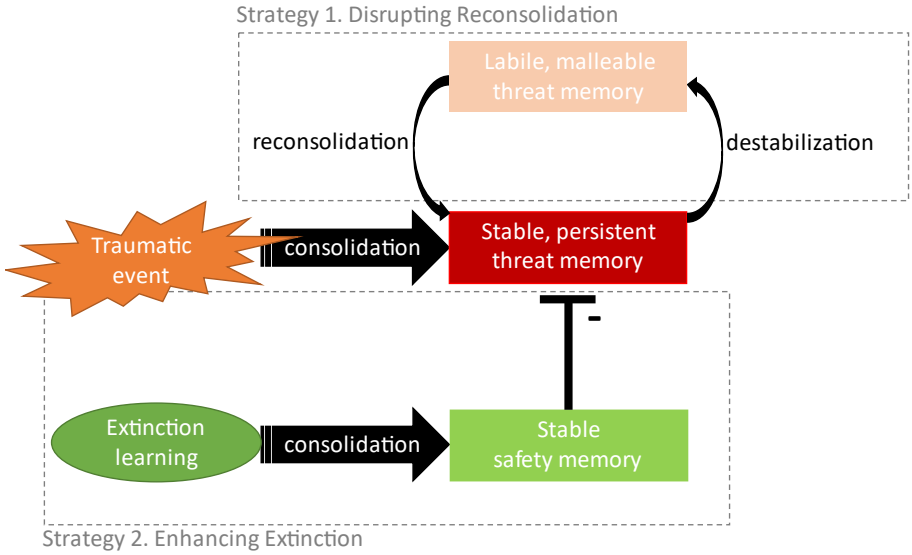


Figure 1.1 Two types of strategies for persistent attention of conditioned threat memories.

1. Reconsolidation

Classic work on threat conditioning suggested that threat memories are sensitive to disruption while they are undergoing consolidation. Amnestic interventions such as electroconvulsive shock (Duncan, 1949), protein synthesis inhibitors (Flexner et al., 1965) or new learning (Gordon & Spear, 1973) during the consolidation window can disrupt the memory. After several hours (~6) however, once the memory is consolidated, these interventions are no longer effective. This work suggests that memories initially exist in a labile state, but enter a stable state once consolidated. Amnestic interventions applied during this window were found to disrupt the formation of long-term memory, while short-term memory remained intact (Schafe & LeDoux, 2000). Once consolidation is complete, these interventions, such as infusion of anisomycin to block protein synthesis, no longer affect long-term memory.

Surprisingly, a seminal study by Nader et al. (2000) demonstrated that after reactivation (i.e., recall of the memory after a reminder), consolidated memories once again require novel protein synthesis to persist in long-term memory. In a cued threat conditioning experiment, rats acquired CRs to an auditory CS after pairings of the CS and aversive electric foot shocks (US). The next day, the consolidated memory was reactivated by a single, unreinforced presentation of the CS. Rats that received an infusion of anisomycin in the amygdala did not show any freezing in response to the CS 24 hours later, suggesting a disrupted threat memory. Based on these and similar, but neglected earlier findings (Misanin et al., 1968; Przybyslawski et al., 1999; Przybyslawski & Sara, 1997), it was suggested that reactivated memories re-enter a labile state and again require a consolidation-like process now referred to as reconsolidation (Nader et al., 2000). By itself, the reconsolidation process appears to

enhance memory and allow for memory updating, while amnestic interventions during the reconsolidation window can persistently disrupt the memory (Tronson et al., 2006). A direct replication of the effects of anisomycin in humans is not possible due to severe side-effects, but comparable reconsolidation interference effects have been demonstrated in humans using other amnestic treatments. Reconsolidation interference was first demonstrated in humans in a study that showed that administration of β-adrenergic receptor antagonists after reactivation disrupts conditioned threat responses (Kindt et al., 2009), and has been replicated using other interventions (Kroes et al., 2014; Vallejo et al., 2019). These findings suggest that amnestic interventions following threat memory reactivation can persistently attenuate threat memories.

Based on the finding that threat memories re-enter a labile state after a reminder, Monfils and colleagues (2009) hypothesized that extinction training during the reconsolidation window might also lead to a similar, persistent attenuation of threat memories. Indeed, while regular extinction training left rats sensitive to the return of threat responses after spontaneous recovery, renewal or reinstatement, extinction preceded by a single reminder led to a persistent attenuation of threat responses (Monfils et al., 2009). Shortly after, the finding that Post-Retrieval Extinction (PRE), also known as the reminder-extinction procedure, prevents recovery of threat responses was replicated in humans by Schiller et al. (2010). Given that behavioural interventions are inherently safer and more accessible than pharmacological interventions, PRE holds great clinical potential. While classic extinction training is focussed on creating a novel, competing safety memory, extinction during the reconsolidation window might attenuate the original threat memory and thereby prevent relapse (Phelps & Hofmann, 2019). However, whereas PRE has been shown to be able to prevent the recovery of threat responses after cued threat conditioning in humans (Monfils et al., 2009; Schiller & Delgado, 2010), it may not be equally effective for other types of threat memories, such as context-conditioned threat memories, given that these depend on distinct neural substrates (Phillips & LeDoux, 1992). Work in rodents, however, suggests that a reminder before extinction can strengthen attenuation of context-conditioned threat memories compared to regular extinction as well (Flavell et al., 2011; Liu et al., 2014; Monti et al., 2017; Piñeyro et al., 2014; Rao-Ruiz et al., 2011). The ultimate goal is to understand whether PRE can facilitate persistent attenuation of context-conditioned threat memoires in humans, in line with findings in rodents. Yet this translation of results for context-conditioning to humans is complicated by practical boundaries. While it is already challenging to create two or more highly controlled, distinct environments in the laboratory in which participants can freely navigate (i.e. requiring multiple connected and distinct rooms), we would also need wearable devices to measure the subtle behavioural (Hagenaars et al., 2014) or physiological indices of fear in humans. However, the introduction of Virtual Reality (VR) allows for the creation of such highly controlled, ecologically

valid environments, and have been shown to facilitate an immersive experience in which human participants feel present in the environment (Sanchez-Vives & Slater, 2005). In **chapter 2**, we use a virtual reality paradigm to test whether the reminder-extinction procedure can persistently attenuate contextual threat responses in humans.

Yet despite the initial promise of PRE, follow-up studies have yielded mixed results (for a meta-analysis, see Kredlow et al., 2015). Failed attempts to verify or replicate the original reports of the efficacy of PRE and its translation in humans suggests that the original effect sizes may have been inflated (Chalkia et al., 2020; Luyten & Beckers, 2017), but it could also be that our understanding of the destabilization and subsequent disruption during the reconsolidation window is currently too limited to consistently reproduce the effects. It has been suggested that PRE may be subject to boundary conditions, i.e. experimental parameters that can block reconsolidation from occurring (for a review, see Zuccolo & Hunziker, 2019). For example, it has been suggested that the likelihood of memories undergoing reconsolidation depends on the age of the memory (Milekic & Alberini, 2002; Suzuki, 2004), memory strength (Suzuki, 2004; Wang et al., 2009), and the extent to which the reminder generates a prediction error (Exton-McGuinness et al., 2015; Pedreira, 2004). The proposed boundary conditions may reflect properties of the reconsolidation process. Given that strong threat memories can be encoded differently compared to weak fear memories, their susceptibility to reconsolidation interference may differ (Haubrich et al., 2020). In **chapter** 3, we ask whether increasing the intensity of the aversive US during threat conditioning reduces the susceptibility of threat memories to the reminder-extinction procedure.

2. Enhancing extinction

A second strategy that could be used to attenuate maladaptive threat responses may be to enhance extinction or safety learning. Guided by an understanding of how regular extinction learning is consolidated and acts to supress threat responses, we may be able to enhance extinction learning by targeting the underlying neural processes. The dominant view on extinction is based on Pavlov's early formulations of extinction as an internal inhibition of conditioned responses, and views extinction as novel learning that inhibits the expression of the conditioned response but does not eliminate it. In addition, it has frequently been demonstrated that extinction learning is relatively weak compared to threat conditioning, as it typically lacks emotional potentiation. This phenomenon has also been described as "adaptive conservatism", as the better-safe-than-sorry approach can be highly adaptive in a world where danger signals are rarely actually followed by imminent threats, but rapid defence responses are still necessary when threat occurs (Dunsmoor, Niv, et al., 2015). Traditionally, evidence that memory for threat conditioning is not erased by extinction and that the original threat memory is more persistent than the extinction memory, comes from the phenomena of spontaneous recovery,

reinstatement, renewal and rapid recovery of threat responses (Bouton, 2002). The relative strength of episodic memories generated by threat and extinction learning can also be compared using the category conditioning paradigm in which semantic categories (e.g., pictures of animals) are either paired with a US (CS+ category) or not (CS- category). While threat conditioning selectively enhances recognition memory for CS+ category exemplars compared to CS- exemplars, recognition of CS+ items presented during within-session extinction drops compared to CS+ items presented during the acquisition phase (Dunsmoor, Campese, et al., 2015). Thus, extinction learning is thought to create a safety memory that is relatively weak compared to conditioned threat memories, suggesting that enhancing or strengthening extinction learning could pave the way to a more persistent attenuation of threat responses.

Early models of extinction learning viewed extinction as a process of "unlearning" that decreased the associative value between a CS and US. According to the influential Rescorla and Wagner model, learning occurs through prediction errors that arise when the predicted outcome and the actual outcome do not match (Rescorla & Wagner, 1972). During acquisition of conditioned threat responses, unexpected presentation of the US generates a positive prediction error and increases the associative strength between CS and US. In contrast, during extinction the unexpected absence of a predicted US generates a negative prediction error and decreases the associative strength between the CS and US. However, the Rescorla and Wagner model fails to provide a complete account of extinction learning, as it cannot explain recovery effects after extinction (Miller et al., 1995). Nevertheless, the concept that extinction learning is driven by prediction error has received increasing support over the years (Dunsmoor, Niv, et al., 2015). Specifically, phasic firing of dopamine neurons in the midbrain was shown to correlate with prediction errors modelled by the Rescorla-Wagner model (Barto, 2018; Montague et al., 1996; Schultz et al., 1997), and their activity appears to support extinction learning (Esser et al., 2021; Raczka et al., 2011; Thiele et al., 2021). Hence, while the Rescorla-Wagner has been discarded as model of extinction because unlearning fails to account for recovery of fear, prediction errors may drive extinction learning.

A second influential model of extinction learning was proposed by Pearce and Hall, and states that extinction learning involves learning of a novel association, a CS-no US association that competes for expression with the CS-US association and thereby inhibits the conditioned response (Pearce & Hall, 1980). In a further attempt to refine this model, the 'latent cause model' states that learning processes during conditioning assume unobservable (i.e. latent) causes that link the CS and US using statistical probability (Courville et al., 2005; Dunsmoor, Niv, et al., 2015). For each presentation of a CS, the presence of the CS, US and other contextual factors are taken into account to determine how likely it is that the CS presentation is related to a specific latent cause (Gershman et al., 2010). Specifically, in

the case of extinction, the unexpected omission of the US is a mismatch with the latent cause inferred during acquisition, so the extinction trials are assigned to a new latent cause where the CS is not associated with the US. A key difference between latent cause models and previous models is that latent cause models allow extinction to have two mechanisms and bridge the gap between associative and inhibitory models of extinction (Dunsmoor, Niv, et al., 2015). First, the original threat memory could be updated when the same latent cause is assumed as is the case during threat conditioning, or second, a new and inhibitory safety memory could be established when a novel latent cause is assumed.

One approach to enhance extinction learning may be to be combine extinction training with the administration of pharmacological interventions that strengthen extinction learning or retention. For example, pharmacological stimulation of the glucocorticoid or endocannabinoid systems during extinction learning has been shown to prevent the recovery of fear (for a review see De Bitencourt et al., 2013). Yet besides pharmacological interventions, behavioural modifications to the extinction procedure that classically consist of a few unreinforced CS presentations, have also been shown to reduce the susceptibility of extinguished threat memories to recovery. For example, building directly on the latent cause model, gradual extinction slowly "weans off" US presentations as a part of the extinction procedure to increase the likelihood that the extinction trials are attributed to the latent cause established during acquisition of threat responses, and prevents spontaneous recovery and reinstatement (Gershman et al., 2013; Gershman & Hartley, 2015). Similarly, massive extinction that continues for many trials after conditioned responding has stopped, was shown to prevent spontaneous recovery and could bind the latent cause associated with extinction to a more generalized context (Denniston et al., 2003; Dunsmoor, Niv, et al., 2015). Another approach to strengthen extinction that is aimed at increasing prediction errors through novelty, is Novelty-Facilitated Extinction (NFE), during which the US is replaced by a surprising, neutral outcome (Dunsmoor, Campese, et al., 2015). At a neural level, NFE enhances the recruitment of the vmPFC during extinction trials, thus increasing engagement of areas classically involved in extinction learning, while simultaneously decreasing activation in areas involved in threat processing, including the insula, dorsal anterior cingulate cortex, and thalamus. However, while these examples of behavioural variations on classic extinction do successfully reduce recovery of threat responses, evidence from the category-conditioning paradigm suggests that episodic memory for episodes of enhanced extinction is still weaker compared to memory for the acquisition of conditioned threat responses (Dunsmoor et al., 2018).

Counterconditioning

In the original study involving "little Albert" the authors suggested that in order to "remove the conditioned emotional responses", one might try to "recondition" by "feeding the subject candy or other food just as the animal is shown" (Watson & Rayner, 1920). This idea evolved into the concept of counterconditioning (CC), that involves pairing a CS with a biologically salient US of the opposite valence during CC as compared to the valence used during the acquisition of a conditioned association. Applied to the end of enhancing extinction, aversive-to-appetitive CC implies pairing a previously threat-conditioned stimulus to an appetitive or rewarding stimulus. Early studies on CC in animals showed mixed results (for a review, see Keller et al., 2020), and from the perspective of latent cause models, it seems likely that introduction of rewards may promote the inference of a novel latent cause that could render the original threat memory sensitive to relapse. However, recent work has indicated that CC can reduce spontaneous recovery of threat responses, and creates episodic memories of comparable strength as threat conditioned memories (Keller & Dunsmoor, 2020). Thus, unlike previously discussed variations on classic extinction, CC may lead to the formation and consolidation of a positive memory that provides stronger competition against retrieval of the threat memory compared to regular or enhanced extinction. In **chapter 3**, we investigate the neural mechanisms underlying CC, and ask to what extent they resemble activity patterns during regular extinction.

Public demand for reconsolidation-based interventions?

In currently available treatments for anxiety or trauma- and stressor-related disorders, patients typically actively work together with a therapist to establish novel safety beliefs through, for example exposure therapy, cognitive behavioural therapy and/or eye-movement desensitization and reprocessing (Ougrin, 2011). These psychological therapies do not affect the original traumatic memory, but rather establish new safety memories that, when treatment is successful, can win the competition for retrieval against traumatic associations (Brewin, 2006). It is clear that this course of treatment requires active involvement from the patient and will often be a struggle, but it has also been thought to allow for a process of posttraumatic growth that can result, for example, in an increased appreciation for life, a nourished sense of personal strength and a change in priorities (Tedeschi et al., 2016). Unfortunately, a substantial percentage of patients (ranging from 0 to 50%, mean 15.6%) drops out of treatment prematurely (Loerinc et al., 2015), and others, although completing treatment, may nevertheless experience relapse (Bouton, 2002). Thus, in the search for effective and tolerable treatments for maladaptive threat responses, we may need to turn to novel Memory Modification Techniques (MMTs) that directly modify original traumatic memories or artificially strengthen extinction learning.

Compared to existing forms of psychotherapy, it would be a minor change in treatment to artificially strengthen extinction learning by, for instance, administering glucocorticoids during exposure therapy (Schelling et al., 2004; Surís et al., 2010; Y. L. Yang et al., 2006), as both these treatments leave the patients' memories intact and aim to establish novel memories to inhibit maladaptive symptoms from traumatic memories. Reconsolidation-based treatments, on the other hand, are fundamentally different as they aim to remove specific memories or their emotional associations. The development of reconsolidation-based interventions has sparked a debate about the potential ethical, legal and social implications. Much of this debate can be traced back to the publication of a report by the US President's Council on Bioethics (2003) that, in short, argued that it is undesirable to use drugs to blunt or erase traumatic events from memory. They provided four main lines of argument. First, if we attenuate painful memories, we may risk blunting all of our experiences, and become numb to both the most painful and the most joyful memories. Secondly, as a community, we may have the moral obligation to remember certain terrible events truthfully (e.g. the Holocaust), because this allows us to feel compassion for those that suffered in the events. Third, in order to be held "morally responsible" for a terrible act, it is necessary that you remember that you carried out the act, even when it is painful. Fourth, given that our memories constitute a core part of our identity, we may lose part of our identity when we remove specific memories, leading to inauthentic lives: lives that are easy to live but not true to ourselves. The issuing of this report led to a number of publications on ethical concerns regarding MMTs, both from bioethicists (e.g. Erler, 2011; Henry et al., 2007; Liao & Sandberg, 2008; Liao & Wasserman, 2007; Parens, 2010) and neuroscientists in the field (e.g. Elsey & Kindt, 2016; Kroes & Liivoja, 2018). But while the development of MMTs has introduced a new vocabulary about altering specific memories that sparked ethical and legal debate, the general public may not think MMTs are very different from other currently available forms of psychotherapy (Elsey & Kindt, 2016).

In a previous survey of public opinion on the use of pharmacological interventions to weaken traumatic memories, participants were asked whether they would want to take a memory dampening drug immediately after experiencing a traumatic event (Newman et al., 2011). By applying a pharmacological intervention, such as the administration of hydrocortisone or the β-adrenergic receptor antagonist propranolol, immediately after or during trauma exposure, the chance of developing Post-Traumatic Stress Disorder (PTSD) could be reduced (see e.g. Pitman et al., 2002; Schelling et al., 2004). Generally, participants were negatively disposed against taking the drug, although attitudes differed slightly between contexts and countries (Newman et al., 2011). The authors suggested that this negative attitude could be largely due to the prophylactic nature of the treatment, as people may be reluctant to undergo pharmaceutic treatments when they believe their chance of developing PTSD after a traumatic event is low in the first place. The introduction of MMTs would allow

us to overcome the limitations of prophylactic treatment, as MMTs may alter previously consolidated memories and could be used specifically to attenuate traumatic memories after the development of PTSD. Thus, it seems likely that public attitudes towards MMTS could differ substantially from reported attitudes towards prophylactic attenuation of traumatic memories. Given that reconsolidation based MMTs could become viable treatment options in the near future, it is vital to understand whether there is a public demand for these MMTs, or whether the general public mirrors the hesitancy expressed by experts in the literature. In **chapter 5**, we describe the results of an international online survey probing public attitudes towards MMTs and try to understand the factors that shape them.

Chapter Introductions

The overall aim of this thesis is to investigate whether we can enhance the attenuation of threat memories by engaging different neurocognitive mechanisms.

In **chapter 2**, we test the hypothesis that presenting an isolated reminder before extinction enhances the attenuation of contextual threat memories. To carry out a close translation of previous work in rodents, we used virtual reality to carry out context conditioning in a highly controlled, immersive virtual environment. Participants underwent a differential threat conditioning task in virtual reality, where they received electrical shocks in one of two virtual rooms. In a between-groups design, participants either received a reminder followed by extinction (PRE) or extinction only. To assess whether the PRE persistently attenuated differential threat responses, we assessed spontaneous recovery and reinstatement (FPS, SCRs) the following day.

Given the number of studies that failed to find an effect of an isolated reminder before extinction, it has been proposed that the reconsolidation process is subject to boundary conditions. One of the proposed boundary conditions is memory strength. Yet due to ethical limitations, it is not possible to create strong experimental threat memories in humans. In **chapter 3**, we investigated in rodents whether the efficacy of the PRE may be dependent on the threat intensity. Rats underwent differential threat conditioning at different intensities of the aversive unconditioned stimulus. We measured spontaneous recovery and reinstatement of conditioned freezing responses to assess whether the PRE resulted in enhanced attenuation of threat responses compared to extinction for weak, but not for strong conditioned threat memories.

Considering the null findings regarding the PRE in **chapter 2** and **3**, we turned to alternative interventions that may attenuate threat responses more effectively than extinction. In **chapter 4**, we examined whether aversive-to-appetitive CC is more effective than regular extinction and explored whether the two processes engage distinct neural mechanisms. Participants acquired differential threat responses to categories of images (objects or animals), and subsequently underwent either CC

or extinction. We used functional magnetic resonance imaging (fMRI) to compare the neural activity during CC and extinction. Spontaneous recovery and reinstatement (differential conditioned PDRs and SCRs) were tested 24 hours later. We also measured recognition memory for the individual items presented during the acquisition and CC/extinction phases of the experiment.

In **chapter 5**, we asked a sample of participants from the public whether they found MMTs morally acceptable. We hypothesized that attitudes towards MMTs are modulated by the information available to participants and may depend on the conditions under which they are applied. In addition, we hypothesized that groups of participants with similar moral convictions may have comparable attitudes towards MMTs. In a between-subjects design, participants either read a brief or extensive introduction to MMTs. Subsequently, they responded to general statements and specific scenarios about MMTs. In the scenarios, we investigated whether the professional background, involvement in crime, presence of mental health disorders in the actor as well as the collective interest in memory retention affected public approval of MMTs. We also investigated whether attitudes towards MMTs were associated with moral intuitions or demographic variables.

In **chapter 6,** the main findings in the previous chapters are summarized and integrated with the existing literature. Here, I also discuss limitations and make suggestions for future research.

Chapter 2. Reconsolidation-extinction in humans: Investigating the efficacy of the reminder-extinction procedure to disrupt contextual threat memories in humans using immersive Virtual Reality

Maxime C. Houtekamer, Marloes J.A.G. Henckens, Wayne E. Mackey, Joseph E. Dunsmoor, Judith R. Homberg, Marijn C.W. Kroes

Abstract

Upon reactivation, consolidated memories can enter a temporary labile state and require restabilisation, known as reconsolidation. Interventions during this reconsolidation period can disrupt the reactivated memory. However, it is unclear whether different kinds of memory that depend on distinct brain regions all undergo reconsolidation. Evidence for reconsolidation originates from studies assessing amygdala-dependent memories using cue-conditioning paradigms in rodents, which were subsequently replicated in humans. Whilst studies providing evidence for reconsolidation of hippocampus-dependent memories in rodents have predominantly used context conditioning paradigms, studies in humans have used completely different paradigms such as tests for wordlists or stories. Here our objective was to bridge this paradigm gap between rodent and human studies probing reconsolidation of hippocampus-dependent memories. We modified a recently developed immersive Virtual Reality paradigm to test in humans whether contextual threat-conditioned memories can be disrupted by a reminder-extinction procedure that putatively targets reconsolidation. In contrast to our hypothesis, we found comparable recovery of contextual conditioned threat responses, and comparable retention of subjective measures of threat memory, episodic memory and exploration behaviour between the reminder-extinction and standard extinction groups. Our result provide no supportive evidence for reconsolidation of context conditioned threat memories in humans and suggest limited efficacy of the reminder-extinction procedure in preventing the return of threat memories.

Introduction

A brief reminder can return consolidated memories to a labile state, requiring re-stabilization processes to maintain the memory, a process referred to as reconsolidation (Nader et al., 2000; Nader & Hardt, 2009, but see Gisquet-Verrier et al., 2015; Lattal & Abel, 2004; Lewis, 1979) for alternative accounts. Interventions that target reconsolidation can modify long-term memories (Nader et al., 2000). This discovery has led to suggestions that reconsolidation-targeting interventions might be used to modify maladaptive memories as a treatment for stress- and anxiety-related disorders (Beckers & Kindt, 2017; Gamache et al., 2012; Kroes et al., 2016; Milton & Everitt, 2010; Schwabe & Wolf, 2009). Yet another important realization of memory research is that there are distinct kinds of memory that rely on different brain regions and are expressed in different forms of behaviour (Henke, 2010; Squire, 1992; Tulving, 1972). People with stress- and anxiety-disorders generally experience several different forms of maladaptive memory expression, such as excessive threat responses, subjective negative feelings, emotional episodic memories, and avoidance behaviours (Foa et al., 1999; Reynolds & Brewin, 1999; Vieweg et al., 2006; Williams, 2016). Critically, to date it is still unclear whether all kinds of memory undergo reconsolidation and whether these are equally sensitive to reconsolidation-targeting interventions. Here, we test whether contextual threat conditioned memories are sensitive to disruption through a behavioural reconsolidation-targeting intervention.

Studies using simple cue conditioning paradigms, in which e.g. a single tone predicts a shock, have provided evidence for reconsolidation in both rodents (Nader et al., 2000) and humans (Kindt et al., 2009, for a review, see Kroes et al., 2016; Nader & Hardt, 2009). In Pavlovian cue conditioning, the formation and storage of the mnemonic association between the conditioned stimulus (CS, e.g. a tone) and the unconditioned stimulus (US, e.g. a shock) is amygdala-dependent (for a review, see Ledoux, 2003). Unlike the highly controlled cued threat-memory paradigms used in laboratory settings, real-life emotional memories can also include information about the spatiotemporal context of a threatening experience, and are more hippocampus-dependent (Alvarez et al., 2008; Bouton, 2002; Eichenbaum et al., 2007; Marschner et al., 2008; Phillips & LeDoux, 1992; Voogd et al., 2019). Therefore, to understand the implications and limitations of reconsolidation interventions for the potential treatment of stress- and anxiety-disorders, it is imperative to also know the impact of interventions targeting the putative reconsolidation process on forms of threat memories that are primarily hippocampus-dependent.

In rodents, evidence for reconsolidation of hippocampus-dependent memories has been obtained using Pavlovian contextual threat conditioning paradigms, where animals learn an association between a particular contextual environment (i.e. the conditioning chamber) and an aversive outcome (i.e. a shock) in the absence of a discrete cue signalling the outcome (Boccia et al., 2004; Debiec et al., 2002;

Flavell et al., 2011; Lee, 2008; Taubenfeld et al., 2001). In humans, however, supportive evidence for reconsolidation of hippocampus-dependent memories is scarce and generally stems from outside the Pavlovian threat conditioning domain, relying on memory paradigms originating from the episodic memory domain, such as word-lists (Hupbach et al., 2007; Kroes et al., 2010) or stories (Kredlow et al., 2016; Galarza Vallejo et al., 2019; Kroes et al., 2014; Schwabe & Wolf, 2009). A recent study trying to unite approaches from the Pavlovian threat conditioning and episodic memory fields, using a category threat conditioning procedure in humans, indicated that episodic memory for items that were part of a Pavlovian threat conditioning experience can undergo reconsolidation, but that the efficacy of reconsolidation-interventions may decrease when episodic memory demands increase (Kroes, Dunsmoor, Lin, et al., 2017), consistent with suggestions from studies with rodents (Alberini, 2005) (see Kroes et al., 2016 for a review). Although category conditioning involves hippocampal processing (de Voogd et al., 2016a; Dunsmoor et al., 2014), the paradigm is still quite different from reconsolidation studies with rodents that have been able to directly interfere with hippocampal processing using contextual conditioning. To further close the gap between rodent and human studies it would be useful to test for the reconsolidation of contextual threat conditioned memories in humans.

Recently, an immersive Virtual Reality (iVR) contextual threat conditioning paradigm for humans was developed, which provides people with a sense of immersion in a virtual environment and allows people to learn an association between a particular contextual environment and an aversive outcome in the absence of a discrete cue signalling the outcome. This iVR context conditioning paradigm was shown to result in the acquisition of contextual threat-conditioned defensive responses, subjective feelings of threat, and episodic memory for details of the threatening spatiotemporal context (Kroes, Dunsmoor, Mackey, et al., 2017). As similar VR contextual threat conditioning paradigms have been shown to involve hippocampal processing in humans (Andreatta et al., 2015 and see Kroes, Dunsmoor, Mackey, et al., 2017 for a review), this iVR contextual threat conditioning paradigm provides an opportunity to investigate reconsolidation of contextual threat conditioned memories that are likely hippocampus-dependent and comparable to memories in studies using contextual threat conditioning procedures in rodents.

To interfere with reconsolidation, the majority of studies have used pharmacological interventions (for a review, see e.g. Kroes et al., 2016; Nader & Hardt, 2009). Yet behavioural interventions, such as the reminder-extinction procedure, may also be able to influence reconsolidation of memory (Monfils et al., 2009). This is an exciting discovery as behavioural interventions can be considered preferable as they are inherently more accessible and safe compared to pharmacological interventions (Holmes et al., 2009; Kroes & Liivoja, 2018). In the behavioural reminder-extinction procedure an isolated

reminder of a threat memory is presented to return the memory to a labile state and next (typically after 10 minutes) standard extinction training is performed. The reminder-extinction procedure has been found to persistently attenuate cued threat memories in both rodents and humans (Agren et al., 2012; Björkstrand et al., 2015; Clem & Huganir, 2010; Schiller et al., 2013; Schiller et al., 2010, yet for non-replications see Kindt & Soeter, 2013; Shiban et al., 2015 and see Kredlow et al., 2016 for a meta-analysis). Studies showing that the reminder-extinction procedure can prevent the return of threat responses have postulated that the reminder triggers a reconsolidation process, and extinction training during this reconsolidation window can overwrite the original threat memory (Monfils et al., 2009). Whether indeed the reminder-extinction depends specifically on disruption of the original memory remains to be determined (Cahill & Milton, 2019), and alternative explanations for the efficacy of the reminder-extinction procedures include active memory integration accounts (Gisquet-Verrier et al., 2015) and the enhanced-extinction account (Cahill et al., 2019a).

Presenting an isolated reminder before extinction has previously shown to enhance attenuation of contextual conditioned threat responses in mice (Rao-Ruiz et al., 2011) and rats (Flavell et al., 2011, yet for a non-replication, see Chan, 2014). In humans, however, an indirect translation of the contextual conditioning paradigm, using compound stimuli consisting of fear-relevant cues presented in different frames to represent different contexts, suggested that the reminder-extinction paradigm does not prevent spontaneous recovery of threat responses (Meir Drexler et al., 2014). In addition, the above mentioned category threat conditioning study (Kroes, Dunsmoor, Lin, et al., 2017) indicated that the efficacy of the reminder-extinction procedure might be limited when episodic memory demands increased. We therefore wondered whether we could reproduce previous findings in rodents (Flavell et al., 2011; Rao-Ruiz et al., 2011), showing attenuation of contextual threat responses after presentation of a reminder before extinction, in humans using a direct translation of rodent contextual threat conditioning paradigms.

Therefore, the objective of this preregistered study (https://osf.io/b2854/) was to investigate the efficacy of the reminder-extinction procedure to prevent the return of contextual threat conditioned memories in humans. To achieve this, participants (N=60) - in a between-subjects design - navigated through an immersive Virtual Reality (iVR) environment where they received aversive electrical shocks to create, modify, and test contextual threat-conditioned memory in a controlled laboratory setting (see Figure 2.1A-E, a modification of the tasks used in Kroes, Dunsmoor, Mackey, et al., 2017). In brief, on day 1 participants were differentially conditioned to a context signalling threat (CTX+) of receiving a transcutaneous electrical shock (US) and a safe context (CTX-). On day 2, participants in one group (reminder-extinction group) were presented with an isolated reminder of the conditioned context (CTX+), while the other group was not (extinction group). After a ten-minute break, both groups

underwent extinction training. On day 3, we tested for the return of context-conditioned threat responses, subjective threat memory, contextual avoidance, and episodic memory in both groups. We hypothesized that the reminder-extinction group would show attenuated recovery of contextual threat conditioned responses, may exhibit reduced avoidance of the threatening context, and potentially altered episodic memory. We followed our preregistered design and analyses with a few minor exceptions, which we clearly indicate below. In contrast to our hypotheses, yet in line with previous non-replications for cue conditioned threat memories (Fricchione et al., 2016; Golkar et al., 2012; Kindt & Soeter, 2013; Klucken et al., 2016; Meir Drexler et al., 2014; Shiban et al., 2015; Soeter & Kindt, 2011), contextual conditioned threat memories (Meir Drexler et al., 2014) and category threat conditioned memories (Kroes, Dunsmoor, Lin, et al., 2017), we found comparable recovery of context conditioned threat responses in the extinction group and the reminder-extinction group, and no group differences on either avoidance behaviour or the other memory tests, suggesting that the reminder-extinction procedure did not modify contextual threat memories in humans.

Methods

Participants

Sixty healthy volunteers (40 female, 20 male; 18–30 years [22.21±0.40]) completed the study. Twenty-six additional participants signed up but did not complete the study: 13 failed to attend all three experimental sessions and 13 had to be discarded due to apparatus failure. Among subjects excluded due to apparatus failure, 9 participants were excluded because the VR equipment disconnected which interrupted the task so that participant no longer sees any images through the glasses. Two participants were excluded due to malfunction of the shock equipment where they did not receive any shocks and two were excluded due to a broken fibre optic cable due to which the physiological data did not contain any event markers. Participants enrolled in the study through a local online psychology research website (SONA) and were fluent in Dutch. Exclusion criteria were: current or lifetime history of psychiatric, neurological, or endocrine illness, abnormal hearing or (uncorrected) vision, average use of more than 3 alcoholic beverages daily, current treatment with any medication that affects central nervous system or endocrine systems, average use of recreational drugs weekly or more, predominant left-handedness (to prevent potential differences in threat responses between left and right handed participants) and proneness to motion sickness. All participants provided written informed consent and received 35 Euro monetary compensation for their participation. As an additional incentive, participants could receive an additional monetary compensation of 5 Euros if they correctly answered 70% of questions on a spatial memory test at the end of the study. The study was approved by the local ethical review board (CMO region Arnhem-Nijmegen). All participants provided written informed consent. All methods were carried out in accordance with the Declaration of Helsinki.

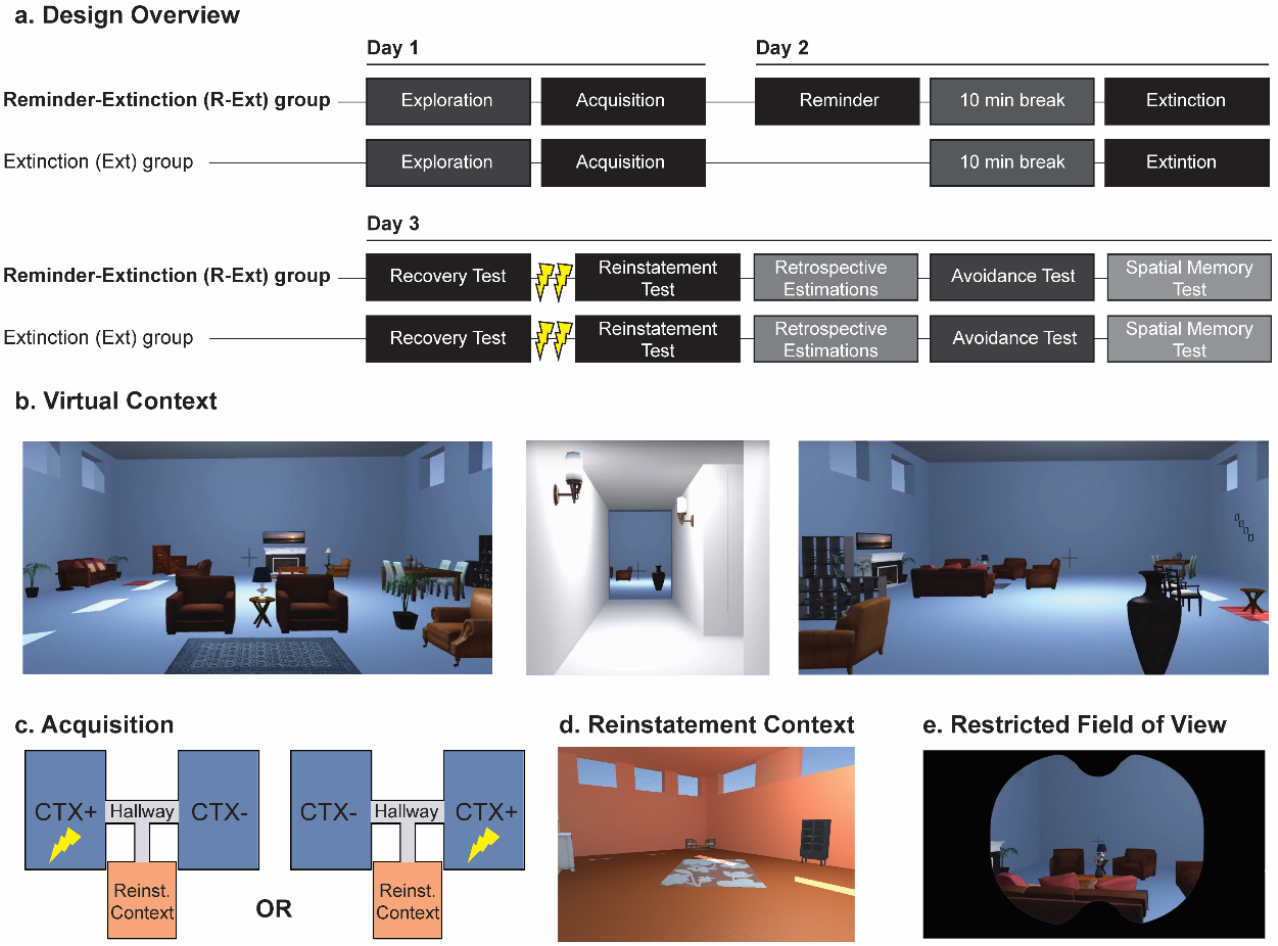


Figure 2.1. Three-day between-subjects iVR contextual threat conditioning study design. (A) Time-line of the experimental design, displayed seperately for the Reminder-Extinction (R-Ext) group and the Extinction (Ext) group. Task in the darkest hues are carried out in iVR (exploration, acquisition, reminder, extinction, recovery test and reinstatement test). (B) 2D depiction of the two blue rooms and the connecting hallway in the iVR environment. (C) Schematic depiction of the iVR context, as seen from above. The two blue rooms are the conditioned context (CTX+, coupled with shocks) and safe context (CTX-, never coupled with shocks), counterbalanced between participants. The hallway (displayed in light grey) connects the two blue rooms, and a third, orange room was used for reinstatement (Reinst. Context). (D) 2D depiction of the orange room in which participants received shocks for reinstatement on day 3. (E) To minimize potential motion sickness, the field of view was dynamically restricted: On straight paths, participants had a wide field of view, and on sharp turns, the field of view was restricted as displayed in the picture.

Immersive Virtual Reality Environment

A Virtual Reality environment was designed in Unity 5 (Unity Technologies, www.unity3d.com), based on a previously used paradigm (Kroes, Dunsmoor, Mackey, et al., 2017). The environment consisted of three virtual living rooms connected via a hallway. Two rooms had blue flooring and walls, and contained identical items, and a third room had orange walls and flooring and contained different living room decorations (see Figure 2.1B). Note, we made a critical modification to the previous version of this iVR context conditioning paradigm (Kroes, Dunsmoor, Mackey, et al., 2017) to theoretically increase the necessity for hippocampal processing, where now the CTX+ and CTX- rooms had the same colour and contained the same furniture items and could only be distinguished based on the

arrangement of the furniture items relative to each other, and their location in space relative to the orange room. To minimize potential discomfort or nausea due to the movement in iVR, a static fixation cross was presented in the middle of the screen, and the field of view was dynamically controlled to be minimal during sharp turns and maximal on straight paths (see Figure 2.1E). In addition, an opaque red line projected just above the floor displayed the path ahead.

Contextual threat conditioning task

During all contextual threat conditioning iVR tasks, participants passively navigated through two blue rooms and the hallway on pre-defined paths. During the threat acquisition task, visits to one blue room were paired with shocks (CTX+) but not in the other blue room (CTX−) or the hallway (see Figure 2.1C). The electrical shock was a 2 millisecond pulse to the distal phalanges of the second and third digit of the right hand using gelled electrodes connected to a constant current stimulator (Digitimer, DS7A; Hertforshire, United Kingdom). To measure conditioned learning, threat potentiation of the eye-blink amplitude was measured in response to loud startle probes presented throughout the task (Kroes, Dunsmoor, Mackey, et al., 2017). Startle probes and US occurred pseudo-randomly from 5–25 seconds after entering a room with the limitation that there had to be 5 seconds between each event, i.e. between each occurrence of shocks and startle probes. Noise probes in the hallway occurred 5–10 second after entry. During the task of approximately 15 minutes, each blue room was visited ten times, each visit lasting approximately 30 seconds. Visits to the CTX+ and CTX- were separated by a 15-second transition through the hallway connecting the two contexts. Six out of ten visits to the conditioned context (CTX+) were paired with one or two shocks (60% reinforcement rate), amounting to a total of eight shocks. The reminder task consisted of a single (30 second) visit to the CTX+ under extinction conditions (i.e., no shock was administered), starting and ending in the hallway, with a total duration of approximately one minute. We opted for a single reminder trial of 30 second because single reminder trials have been showed to labilize memory whilst more trials trigger extinction learning mechanisms instead (Merlo et al., 2014; Sevenster et al., 2014a). We elected for the reminder trial to be as long as an acquisition trial (30s), as is standard in reconsolidation-targeting cue-conditioning paradigms (Kindt et al., 2009; Monfils et al., 2009; Nader et al., 2000; Schiller et al., 2010a). The duration of our reminder trial was therefore slightly shorter than the 90 second to 5 minute reminders of studies in rodents showing diminishment of contextual conditioned threat responses following a reminder-intervention strategy (Cassini et al., 2017; Lee et al., 2019; Rao-Ruiz et al., 2011; Akinobu Suzuki et al., 2004) but not as long as long as the 30 minute exposure that induce extinction in these studies (Cassini et al., 2017; Rao-Ruiz et al., 2011; Suzuki et al., 2004). As our result indicate (see below) our single 30 second reminder was long enough to reactivate memory whilst brief enough not to result in extinction learning. The extinction task was of equal duration and set-up as the acquisition task. Extinction consisted of ten visits to the CTX+ and ten visits to the CTX- of

approximately 30 seconds interleaved by twenty 15-second visits to the hallway while no shocks were administered throughout the extinction task, with a total duration of approximately 15 minutes. It should be noted that we accidently did not adapt the number of CTX+ visits during extinction training for the R-Ext group to compensate for the reminder visit. Thus, including the reminder, the R-Ext group was exposed to one additional CTX+ visit under extinction conditions. To test for spontaneous recovery of the threat response, a shorter version of the task was used with three 30-second visits to each blue room interleaved with six 15-second hallway visits, with a total duration of approximately 5 minutes. No shocks were administered. To test for the reinstatement of conditioned fear responses, participants were passively guided through the third, orange room (see Figure 2.1D) where they received two un-signalled shocks. Afterwards, they were again guided through the two blue rooms, for three 30-second visits to each blue room interleaved with six 15-second visits to the hallway, totalling to approximately 6 minutes.

Physiology collection and Data Analysis

Eye-blink startle

Startle responses were measured using electromyography (EMG) of the right orbicularis muscle and evoked using startle probes (binaural bursts of 100 dB white noise presented for 50 ms). Data were collected using a BrainAmp system, recorded with the BrainVision recorder software (Brain Products GmbH, Munich, Germany) and analysed by means of an in-house analysis program written in Matlab (the MathWorks) that uses the FieldTrip toolbox (Oostenveld et al., 2011, as before in Kroes, Dunsmoor, Mackey, et al., 2017). Responses to the startle probe were found to be consistently delayed compared to latencies in previous studies (Klumpers, Morgan, et al., 2015; Kroes, Dunsmoor, Mackey, et al., 2017; Van Well et al., 2012), due to a 120 ms delay in tone presentation within the iVR task in our current set-up. Therefore, we deviated slightly from our preregistration and, based on the observed mean latency of startle responses across all conditions and participants, determined responses to the startle probe as maximum EMG response between 140 ms and 240 ms relative to our trial onset marker. A baseline measure of the mean EMG magnitude in a 500ms window prior to trial onset was subtracted from the maximum EMG response. In line with previous studies, startle responses for each trial were transformed to T-scores (z-score*10 + 50) for each participant and task separately (Klumpers, Kroes, et al., 2015; Kroes, Dunsmoor, Mackey, et al., 2017; Van Well et al., 2012).

Skin Conductance

Electrodermal activity (EDA) was assed using two Ag/AgCl electrodes attached to the distal phalanges of the second and third digit of the left hand. Data were collected using a BrainAmp system and recorded using BrainVision recorder software (Brain Products GmbH, Munich, Germany) and analyzed using an in-house analysis program written in Matlab (the MathWorks) using FieldTrip (Oostenveld et al., 2011 as before in Kroes, Dunsmoor, Mackey, et al., 2017). Skin conductance responses (SCRs) were

measured after startle bursts and during transitions from the neutral hallway to the conditioned contexts (blue rooms). Responses were defined as the through-to-peak amplitude difference in skin conductance of the largest deflection in the latency window from 0–4.9 s after event onset to ensure that responses could not be contaminated by other events (shocks or following startle probes). The raw skin conductance responses were square root transformed, in line with previous studies (Dunsmoor et al., 2012; Klumpers, Kroes, et al., 2015; Milad et al., 2007).

Heart rate

Raw pulse data were measured using a pulse oximeter and collected using a BrainAmp system and recorded using BrainVision recorder software (Brain Products GmbH, Munich, Germany). Pulse data were processed offline using in-house software to detect R-peaks automatically, following previous literature (Klumpers et al., 2017). All R-peak time-courses were visually inspected and faulty peak locations were manually corrected. Interbeat intervals, the time between two R-peaks, were calculated, converted to beats per minute (BPM), and down-sampled to 2 Hz. Heart-rate responses were defined as time-series from 0-4 s after event onset expressed as change in BPM with respect to a mean baseline during the 1 s before event onset. Average heart rate (HR) responses were calculated for each stimulus (CTX+, CTX-) per phase of each task (early: first half of trials, late: second half of trials) for each participant.

Valence and arousal Ratings

Valence and arousal ratings were obtained using self-assessment manikin scales. The valence scale ranged from 1 (=extremely negative) to 10 (=extremely positive). The arousal scale ranged from 1 (=extremely calm) to 10 (=extremely excited).

Retrospective shock estimation and contingency awareness questionnaire

The retrospective shock estimation and contingency awareness questionnaire asked participants to estimate the number of shocks they thought they had received and estimate the percentage of times that they had received a shock in each of the blue rooms for each experimental task (Kroes et al., 2016; Kroes, Dunsmoor, Mackey, et al., 2017).

Avoidance test

In the avoidance task, participants freely navigated through the two blue rooms and the hallway in search of a hidden coin reflecting a monetary reward for two minutes. Their location was continuously monitored. In reality no coins were present anywhere and the task was stopped after two minutes, allowing investigation of an equal amount of exploration time and avoidance for each participant. Behaviour was scored as the first room that was visited (CTX+ or CTX-) and the time spent in the CTX+ and the CTX-.

Spatial memory test

The spatial memory test consisted of 8 questions probing the position of furniture items in each of the two blue rooms. Participants placed images of furniture items that had been present in the rooms on a spatial grid representation of each room.

iVR experience questionnaire

The iVR experience questionnaire assessed on a 5-item scale how participants had felt during the virtual reality tasks (“I felt no discomfort”, “I was a tiny bit uncomfortable, but not too bad”, “I was slightly uncomfortable”, “I was moderately uncomfortable and slightly nauseous”, “I was very uncomfortable and very nauseous”), and whether they had experience using Virtual Reality technology (‘’No experience’’ , ‘’Once, a couple of minutes’’, ‘’Once for a while’’, ‘’For a while on several occasions’’, ‘’regularly’’) and playing video games in general (‘’No experience’’, ‘’Very limited experience, I hardly ever play video games’’, ‘’Nowadays I rarely play video games, but I used to play video games often’’ , ‘’Regularly’’, ‘’Often’’), as previously described (Kroes, Dunsmoor, Mackey, et al., 2017).

Inventories and anxiety questionnaires

Participants completed the State-Trait Anxiety Inventory (Spielberger, 1983), Intolerance of Uncertainty Scale (Carleton et al., 2007), Berkman-Syme Social Network Index (Berkman & Syme, 1994), and Childhood Trauma Questionnaire - short form (Bernstein et al., 2003).

Procedures

The design of this study is illustrated in Figure 2.1. Participants were pseudo randomly assigned to either the Reminder-Extinction (R-Ext) or Extinction (Ext) group. The study was conducted over three consecutive days. On the first day of the experiment, shocks were calibrated using an ascending staircase procedure starting with a low voltage setting near a perceptible threshold and increasing to a level deemed “maximally uncomfortable but not painful” by the participant, in keeping with prior threat conditioning protocols (Dunsmoor, Murty, et al., 2015; Kroes, Dunsmoor, Mackey, et al., 2017; LaBar et al., 1998).

Participants wore the consumer version of the Oculus Rift headset as previously described (Kroes, Dunsmoor, Mackey, et al., 2017). The headphone component of the Oculus rift was removed and replaced by Sennheiser HD 202 (Wedemark, Germany) headphones. Before the acquisition of contextual fear, participants were asked to freely explore the rooms and hallway for 2 minutes to encourage pre-exposure to the contexts prior to conditioning. After the exploration, valence and arousal ratings were obtained for the different rooms. Next, participants were given a surprise memory test and asked to locate three items in both rooms. After having completed the test, participants were told that they would be asked to complete a similar test on the third day of the experiment, and they

were able to receive an additional monetary compensation if they correctly answered at least 10 out of the 16 questions on that test. This spatial memory test and these instructions were added to ensure that participants would pay attention to, and remember, the differences in spatial layout between the two rooms.

Next, participants were equipped with measurement devices for startle response, skin conductance and heart rate. We explained that loud noises would be presented during the next virtual reality task, but that we would start with a brief task to allow the participants to habituate to the sound. Participants listened to 9 startle probes while viewing a blank grey screen (without the VR headset) to allow startle responses to habituate.

After habituation, participants were prepared for the acquisition task, and instructed that they would be visiting the two blue rooms and the hallway, and told to pay attention to the fact that a relationship existed between the two blue rooms and the shocks. The participants were told that they could not receive shocks in the hallway between the rooms. During the threat acquisition task, shocks were administered in one blue room (CTX+) but not in the other blue room (CTX−) or the hallway (see Figure 2.1C). After this task, the VR headset was removed and valence and arousal ratings were obtained for the different rooms. Next, recording equipment was removed and participants were thanked for their effort during the session.

Participants returned to the lab the following day, and were immediately equipped with recording devices for startle response, skin conductance and heart rate. Participants that had been assigned to the reminder-extinction group were told that they would again be visiting the different rooms and may receive shocks, and that the task would continue as before. To reactivate the contextual threat conditioned memory, they were guided through the CTX+ once. In line with previous studies, the reminder was followed by a 10 minute break, so that the following extinction task would fall within the putative reconsolidation window. During this break, all participants (both the R-Ext as Ext groups) watched 10 minutes of landscape scenes from BBC Planet Earth (2006 TV series). Participants were explicitly told that they would not receive any shocks during this break, and the shock equipment was visibly turned off for the duration of the break. All participants were then told that the task would continue as before, that they would again hear sounds and might receive shocks. This procedure is in line with previous reports (Kroes, Dunsmoor, Lin, et al., 2017; Schiller et al., 2010; Steinfurth et al., 2014). Participants where then subjected to the extinction task. After the task, the VR headset was removed and valence and arousal ratings were obtained for the different rooms. Recording equipment was removed and participants were thanked for their efforts.

Participants returned to the lab again the following day for a third session. Startle response, skin conductance and heart rate recording equipment was attached, and participants were instructed that the tasks would continue as before, except for the fact that there would be two shorter tasks immediately following each other. They were told that they would visit the different rooms and could receive shocks. Participants completed the spontaneous recovery task. After the spontaneous recovery task, the reinstatement task was started immediately. To test for the reinstatement of contextual threat conditioned responses, participants were passively guided to the third, orange room (see Figure 2.1D), and received two un-signalled shocks while moving through the orange room. Afterwards, the participants were again guided through the two blue rooms and responses to startle probes were measured.

Throughout all tasks in iVR, the participants were attached to the shock electrodes, the shock stimulator was set to the 'On' position and they were instructed that they could receive shocks.

After the end of the reinstatement task, valence and arousal ratings were obtained for the different rooms. In addition, participants were asked to estimate the number of shocks they thought they had received and the percentage of times that they received a shock in each of the blue rooms for each experimental task. Next, the participants completed the spatial memory test at their own pace.

Once they completed the memory task, participants were instructed that a coin was hidden in one of the two blue rooms, and that they had two minutes to find the coin. They were told that the task would end automatically if they walked to the coin in iVR. Using the oculus touch controllers, the participants navigated freely for two minutes, after which the task was stopped and the participants were debriefed about the nature of the task.

Participants completed the State-Trait Anxiety Inventory, Intolerance of Uncertainty Scale, Berkman-Syme Social Network Index, and Childhood Trauma Questionnaire - short form at their own pace. We debriefed participants about the purpose of the study and provided information about reimbursement. Finally, participants were given the opportunity to ask questions.

Statistics

Statistical analyses were performed in SPSS (IBM SPSS Statistics Inc.). Dependent measures were submitted to repeated measure ANOVAs and statistics were Greenhouse-Geisser or Huyn-Feldt corrected for non-sphericity when appropriate (i.e, if sphericity assumptions were violated and epsilon was smaller or greater than 0.75, respectively). Significant findings from ANOVAs were followed-up by paired- and independent samples t-tests. We report partial eta-square as measure of effect size. Means ± s.e.m are provided where relevant unless otherwise indicated.

Results

Participants

We first assessed whether there were any group differences in age, sex, motion sickness during the iVR tasks, experience with iVR game experience and time spent playing games. Exploratory t-test did not reveal any group differences (All P's > 0.075). The median response across groups to the iVR question about how participants experienced the iVR was "I was a tiny bit uncomfortable, but not too bad", and no participants indicated to have felt "[…] very uncomfortable and very nauseous".

Immediately following the acquisition of contextual threat conditioning, fifty-seven out of sixty participants could explicitly state the relationship between the conditioned contexts and shocks, indicating that they learned the conditioned association. We therefore opted to include as many people as possible for our different dependent measures whilst adhering to our preregistered inclusion criteria. We describe our inclusion criteria and number of included participants for each measure below.

Fear-Potentiated Startle

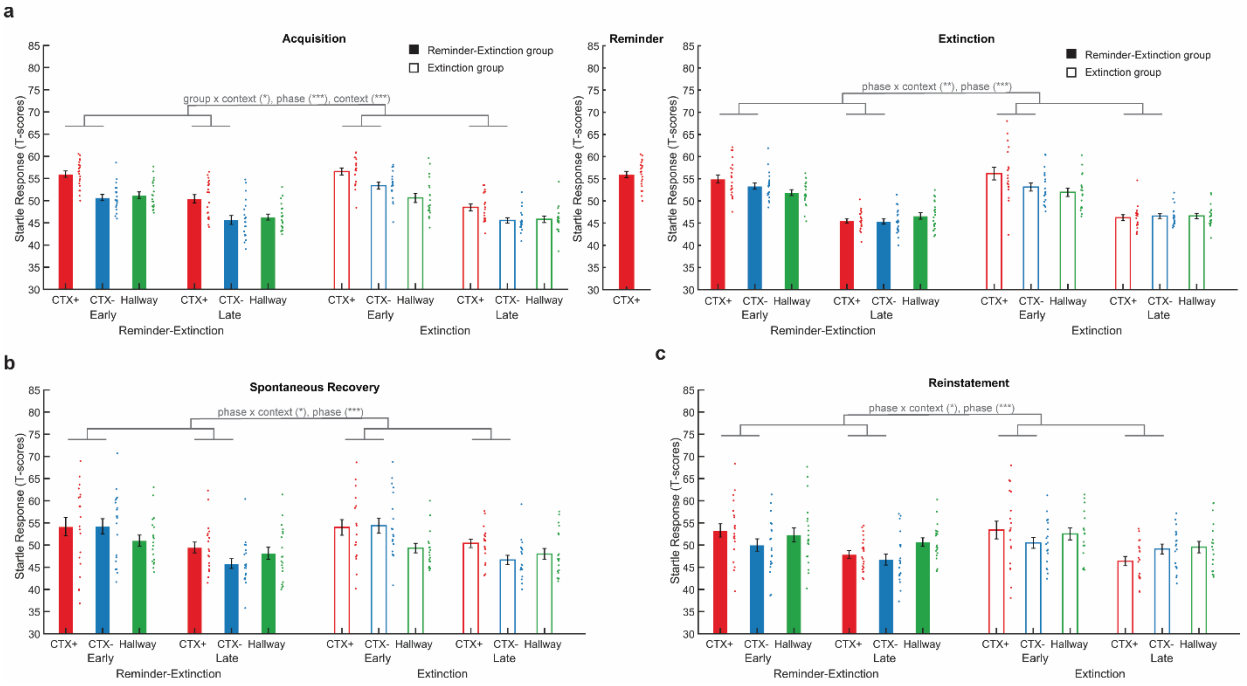


*Figure 2.2. Results of fear-potentiated startle (FPS) response. The contextual threat conditioning procedure resulted in acquisition, retention and extinction of threat-related FPS responses, but an isolated reminder before extinction did not prevent the return of FPS responses on the following day. Bars reflect mean t-scored startle responses during the early (first half of trials) and late phase (second half of trials) of each task for the the threat (CTX+, red) and safe context (CTX-, blue) and neutral hallway (green) for the Reminder-Extinction group (solid bars) and Extinction group (open bars). Error bars = s.e.m., adjacent dots represent jittered individual data-points. *p < 0.05 , **p < 0.01 , ***p < 0.001 (A) Both groups acquired comparable differential contextual threat conditioned FPS responses. On day 2, both groups showed comparable extinction of FPS responses, where differential FPS responses were fully extinguished at the end of the task. (B) Both groups showed comparable generalized spontaneous recovery of FPS responses to both the CTX+ and CTX- during the early phase of the sponteneous recovery test. Although the initial recovery is generalized, FPS reponses in the CTX+ showed slower re-extinction, indicating differential retention of the conditioned FPS responses. (C) Following two unsignaled shocks, FPS reponses showed evidence for reinstatement as differential responses to the CTX+ and CTX- are greater during the early phase as compared to the late phase.*

Fear-potentiated startle (FPS) served as our main index of contextual threat acquisition, reactivation, extinction, and recovery (Figure 2.2). As determined in our pre-registration, we only included participants who showed successful conditioning during the acquisition phases, as measured by a numerically greater startle response in the threat (CTX+) compared to the safe context (CTX-). Twenty-one participants (out of twenty-seven) showed numerically greater startle responses in the CTX+ as compared to the CTX- for the R-Ext group, and nineteen (out of thirty-three) participants for the Ext group. A complete description of results used to verify comparable acquisition, extinction and retention of contextual conditioned startle responses is included in the supplementary information. Here, for readability, complete statistics are only provided for the critical tests on the return of threat.

During the late phase of the acquisition phase on day 1, we observed comparable discriminatory contextual threat conditioned startle responses between both groups (Figure 2.2a). Unexpectedly, over the entire acquisition phase, we observed greater differential FPS responses for the R-Ext than Ext group ($t(38)=2.244$, $p=0.031$, R-Ext: 5.0±0.70, Ext: 3.0±0.50). However, separate repeated measures ANOVAs (rmANOVAs) for the early and late phase of acquisition showed that at the start of extinction, there was a trend interaction effect of group x context ($F_{1,38}=53.311$, $p=0.077$, $\eta2=0.080$) but this trend did not persist during the late phase of acquisition ($F_{1,38}=1.643$, $p=0.208$, $\eta2=0.041$). Thus, critically, in the late phase of acquisition, both groups showed comparable differences between startle responses in the CTX+ and CTX-, indicating comparable acquisition of contextual conditioned threat responses. In the R-Ext group, the reminder resulted in reactivation of the contextual threat conditioned memory, shown by greater startle responses in the CTX+ than hallway ($t(20)=3.114$, $p=0.005$, CTX+: 61.8±2.8=, hallway: 49.26 ± 1.8, as participants did not traverse the CTX- during the reminder, a comparison between FSP in the CTX+ and CTX- was not possible). In addition, the reminder trial did not trigger extinction learning, indicated by the absence of a reduction in freezing scores from the reminder trial to the first CTX+ trial during extinction ($p=0.775$). Afterwards, both groups underwent succesful extinction of contextual threat conditioned FPS, which was preceded by an isolated reminder for the R-Ext group. A group (R-Ext, Ext) x phase (early, late) x context (CTX+, CTX-) rmANOVA on FPS responses during the extinction task revealed an interaction of phase x context ($F_{1,37}=8.552$, $p=0.006$, $\eta2=0.188$) and a main effect of phase ($F_{1,37}=217.726$, $p<0.001$, $\eta2=0.855$), with no other main effects or interactions. During the late phase of the extinction task, both groups show comparable and succesful extinction, indicated by an absence of differential FPS in the late phase of extinction ($t(39)=-0.393$, $p=0.696$, CTX+: 45.6±0.32, CTX-: 45.8±0.37) that was not significantly different across groups (All Ps > 0.16).

On day three, spontaneous recovery of FPS was tested under extinction conditions. We observed comparable spontaneous recovery of FPS responses in both groups (Figure 2.2b). A group (R-Ext, Ext)

x phase (early, late) x context (CTX+, CTX-) rmANOVA revealed a significant interaction effect of phase x context ($F_{1,38}$=4.452, p=0.041, η2=0.105), and a main effect of phase ($F_{1,38}$=27.113, p<0.001, η2=0.416). Importantly, there was no significant group x context x phase interaction (p=0.905) or other interaction with or main effects of group (all P's > 0.5), indicating that spontaneous recovery was not affected by the presentation of an isolated reminder before extinction the previous day. Follow up paired t-tests revealed greater differential responses (CTX+ - CTX-) in the late compared to the early phase (t(39)=-2.134, p=0.039, early: 0.24±1.8, late: 3.7±1.1), which was driven by greater responses in the CTX+ than the CTX- in the late phase (t(39)=3.264, p=0.002, CTX+: 49.9±0.74, CTX-: 46.2±0.72) but not in the early phase (t(39)=-0.133, p=0.895, CTX+: 54.1±1.4, CTX-: 54.3±1.2). These findings seem to indicate participants across both groups initially showed generalized recovery of threat responses in both the CTX+ and CTX- and over the course of the spontaneous recovery test were slower to extinguish FPS responses in the CTX+ compared to the CTX-, indicating retention of the differential conditioned contextual threat response. To further test for the presence of spontaneous threat recovery, we assessed the change in FPS from the end of extinction to the beginning of sponteneous recovery. A group (R-Ext, Ext) x task (late extinction, early spontaneous recovery) x context (CTX+, CTX-) rmANOVA revealed a main effect of task ($F_{1,37}$=79.681, p<0.001, η2=0.683) and no significant main effects or interactions of group and or context (all Ps > 0.6). As we found no effect of group or context and an overall change in FPS from one task to another is logically expected for within day T-transformed data we did not follow up on this finding any further.

Next we tested for reinstatement of contextual threat conditioned FPS responses. We found comparable reinstatement of differential FPS responses for both groups (Figure 2.2c). A group (R-Ext, Ext) x phase (early, late) x context (CTX+, CTX-) rmANOVA revealed an interaction effect of phase x context ($F_{1,38}$=6.077, p=0.018, η2=0.138) and a significant main effect of phase ($F_{1,38}$= 17.334, p<0.001, η2=0.313), but no significant interactions with or main effect of group (all P's > 0.2). A follow up t-test revealed greater differential responses to the CTX+ and CTX- in the early compared to the late phase of reinstatement (t(39)=2.405, p=0.021, early: 3.1±1.4, late: -0.68±1.1). Specifically, we observed greater startle responses in the CTX+ as compared to the CTX- in the early phase (t(39)=2.142, p=0.039, CTX+: 53.3±1.2, CTX-: 50.2±0.89), but not in the late phase (t(39)=-0.628, p=0.534, CTX+: 47.2±0.64, CTX-: 47.9±0.82). To test for the increase in responses, mean startle responses were subjected to a task (late recovery test, early reinstatement test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There was a significant main effect of context ($F_{1,38}$=12.088, p=0.001, η2=0.241) and task ($F_{1,38}$=16.426, p<0.001, η2=0.302) and no significant main effect of group or interactions with group (all P's > 0.5). Follow up t-tests revealed that startle responses were higher during the reinstatement test than during the spontaneous recovery test (t(39)=4.122, p<0.001, late spontaneous recovery: 48.1±0.47, early

reinstatement: 52.8±0.75), and FPS responses were greater in the CTX+ than in the CTX- (t(39)=3.53, p=0.001, CTX+: 51.6±0.76, CTX-: 48.2±0.51). As reinstated responses often extinguish rapidly, we also submitted reinstatement index scores (first trial of reinstatement test - last trial of recovery test) to a context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no significant main or interaction effects of group and context (all P's > 0.3). This suggests that initial reinstatement generalizes to both the CTX+ and CTX-, and is not affected by a reminder.

Valence and Arousal

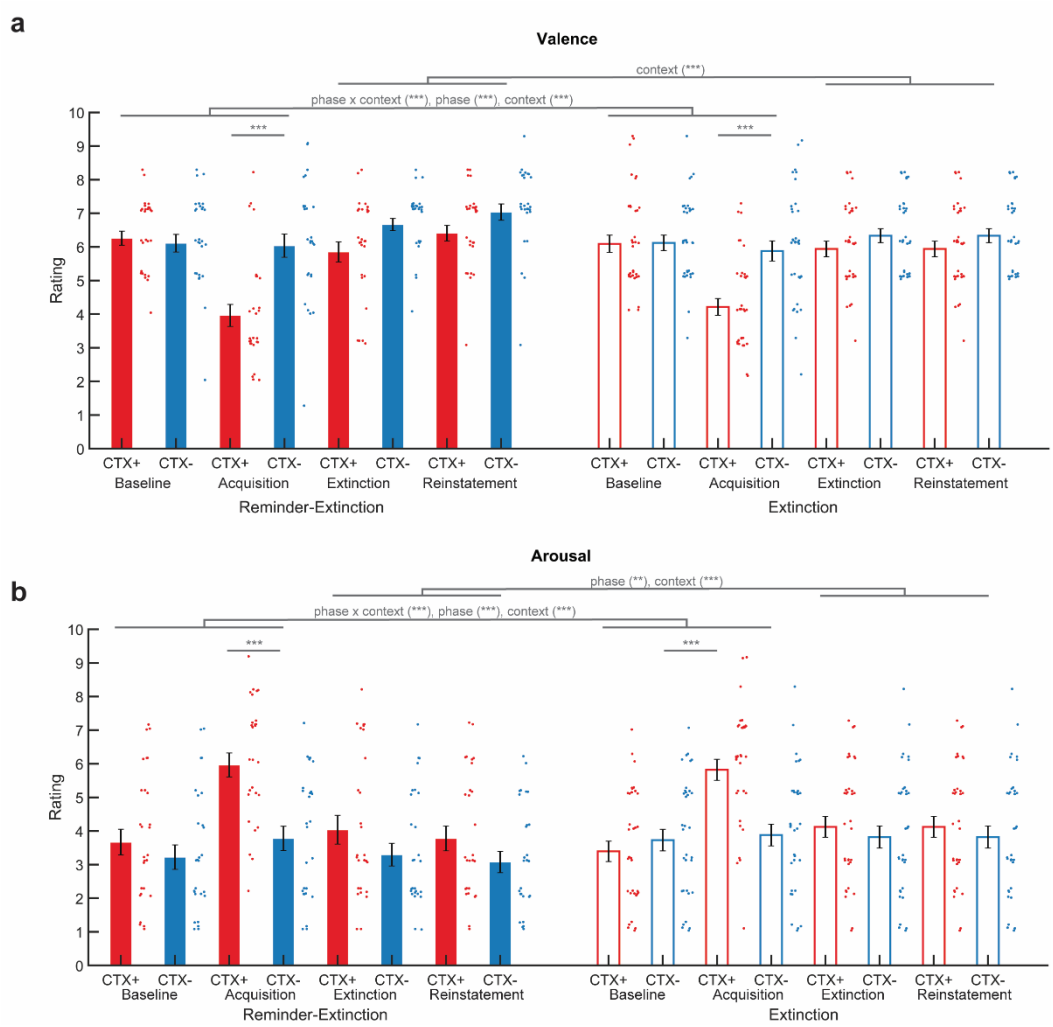


*Figure 2.3. An isolated reminder before extinction did not influence retention of valence and arousal ratings. Context conditioning resulted in acquisition of subjective threat, which was subsequently extinguished, and re-extinguished after the reinstatement test under extinction conditions. Bar plots reflecting mean valence and arousal ratings before acquistion, after acquisition, after extinction and after reinstatement of context conditioning for the threat (CTX+, red) and safe context (CTX-, blue) in the Reminder-Extinction (solid bars) and Extinction (open bars) groups. Context conditioning resulted in (A) lower valence ratings and (B) higher arousal ratings. Error bars = s.e.m., adjacent dots represent jittered individual data-points. **p<0.01, ***p < 0.001.*

Valence and arousal ratings were obtained before and after the contextual threat conditioning task, after extinction, and after the reinstatement test. The valence and arousal ratings showed successful acquisition of differential context conditioned threat memories, and this effect decreased but

persisted after extinction and reinstatement (Figure 2.3). Yet we found no evidence for an effect of the reminder-extinction procedure on these subjective measures of contextual threat conditioned memory.

Valence ratings and arousal ratings were subjected to a phase (baseline, after acquisition) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA to explore whether both groups showed similar acquisition of contextual threat conditioned memories. For valence ratings, this revealed an interaction of phase and context ($F_{1,58}$=35.790, p<0.001, η2=0.382), and a significant main effect of phase ($F_{1,58}$=52.368, p<0.001, η2=0.474) and context ($F_{1,58}$=33.563, p<0.001, η2 = 0.367). For arousal ratings, we found an interaction of phase and context ($F_{1,58}$=46.287, p<0.001, η2=0.444), and main effects of phase ($F_{(1,58)}$=40.566, p<0.001, η2=0.412) and context ($F_{1,58}$=36.430, p<0.001, η2=0.386). Neither valence nor arousal ratings showed effects of group (all P's > 0.15). Differential ratings (CTX+ - CTX-) increased after acquisition for both valence (t(59)=5.829, p<0.001, baseline: 0.05±0.14, after acquisition; 1.9±0.28) and arousal (t(59)=6.939, p<0.001, baseline: 0.02±0.16, after acquisition: 2.1±0.28). Baseline ratings were similar for the CTX+ and CTX- for valence (p>0.7, CTX+: 6.17±0.17, CTX-:6.12±0.17) and arousal (p>0.9, CTX+: 3.5±0.24, CTX-: 3.5±0.24), while after the acquisition task, valence ratings were lower for the CTX+ than the CTX- (t(59)=-6.513, p<0.001, CTX+: 4.1±0.2), CTX-: 6.0±0.22) and arousal ratings were higher for the CTX+ than the CTX- (t(59)=7.252, p<0.001, CTX+: 5.9±0.23, CTX-: 3.8±0.24). Valence ratings for the CTX+ decreased after acquisition (t(59)=-9.791, p<0.001, baseline: 6.17±0.17, after acquisition 4.10±0.20), while valence ratings for the CTX- did not change (p=0.48). Arousal ratings for the CTX- did not change from baseline to after acquisition (p>0.16), but arousal ratings for the CTX+ increased after acquisition (t(59)=8.667, p<0.001, baseline: 3.5±0.24, after acquisition: 5.9±0.23). Thus, both groups show a similar acquisition of a differential conditioned threat response in valence and arousal ratings.

To test whether the reminder-extinction procedure affected the recovery of valence and arousal ratings after completion of the reinstatement test, valence and arousal ratings were subjected to a time (after extinction, after reinstatement) x context (CTX+, CTX−) x group (R-Ext, Ext) rmANOVA. For valence, there was a main effect of context ($F_{1,58}$=18.076, p<0.001, η2=0.238) and phase ($F_{1,58}$=4.825 p=0.032, η2=0.077), but no interactions with group (all P's > 0.07). Similarly, for arousal we found main effects of phase ($F_{1,58}$=4.793, p=0.033, η2=0.076) and of context ($F_{1,58}$=12.071, p=0.001, η2=0.172), but no interactions with group (all P's >0.2). Follow up t-tests of mean ratings after the extinction and reinstatement sessions showed that valence ratings remained were lower for the CTX+ than the CTX- (t(59)=-4.048, p<.001, CTX+: 6.1±0.16, CTX-: 6.6±0.14) and arousal ratings remained higher for the CTX+ than the CTX- (t(59)=3.345, p<0.001, CTX+: 3.9±0.23, CTX-: 3.4±0.21). Differential ratings (CTX+ - CTX-) did not change from after extinction to after the reinstatement test (all Ps > 0.3). As our

reinstatement test was carried out under extinction conditions, and the arousal ratings were taken after the end of this task, it is not surprising that we do not see any effect of reinstatement on differential arousal ratings measured after the reinstatement test.

Retrospective shock estimation

To test the effect of a reminder on retrospective shock estimation and awareness at the end of the study, shock estimates and contingency awareness for the acquisition task of day 1 were subjected to a context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. For the estimated reinforcement rate, there was a main effect of context ($F_{1,58}$=120.363, p=0.000, η2=0.638), but no effect of group (p=0.711), nor interaction (p=0.948) (Figure 2.4a). As a Kolmogorov-Smirnov test indicated that estimates for the reinforcement rate of the CTX- did not follow a normal distribution (D(60)=0.473, p<0.001), we deviated from the pre-registered test and used a Wilcoxon signed-rank test as non-parametric alternative to the paired t-test. A follow-up Wilcoxon signed-rank test revealed that across both groups, the estimated reinforcement rate was higher for the CTX+ than the CTX- (Z=-6.241, p<0.001, CTX+: 52.5±3.5%, CTX-: 7.8±2.5%). For the number of shocks participants estimated to have received, there was also a main effect of context ($F_{1,58}$=145.004, p<0.001, η2=0.714), but no effect of group (p=0.803) or interaction (p=0.418) (Figure 2.4b). As a Kolmogorov-Smirnov test also indicated that estimates for the reinforcement rate of the CTX- did not follow a normal distribution (D(60)=0.482, p<0.001), we again deviated from the pre-registered tests and followed up with a Wilcoxon signed-rank test. In both groups, the estimated number of shocks was higher for the CTX+ than the CTX- (Z=-6.540, p<0.001, CTX+: 6.4±0.41, CTX-: 0.58 ± 0.18).

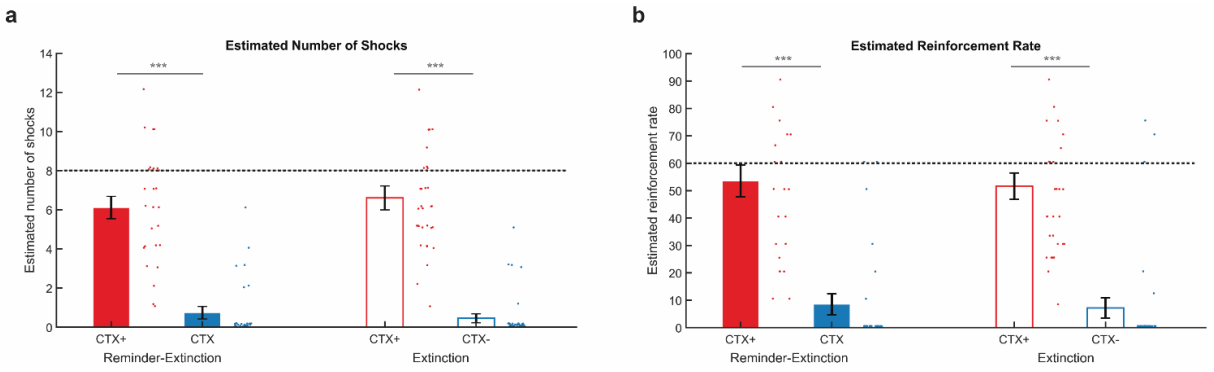


Figure 2.4. An isolated reminder before extinction did not affect explicit threat memory after context conditioning. Bar plots reflecting (A) the mean estimated number of shocks received during the acquisition task and (B) the estimated reinforcement rate during the acquisition task for the threat (CTX+, red) and safe context (CTX-, blue) in the Reminder-Extinction (solid bars) and Extinction (open bars), tested at the end of the experiment. Dashed line indicates (A) the actual number of shocks (8 in CTX+ only) and the actual reinforcement rate (60% in CTX+ only). Error bars = s.e.m., adjacent dots represent jittered individual data-points.

Spatial memory test

To test how contextual threat conditioning and the reminder-extinction procedure would affect the participants' ability to remember the spatial location of items in the contexts, we asked participants to indicate on a grid-map representation of the rooms where specific furniture items had been located.

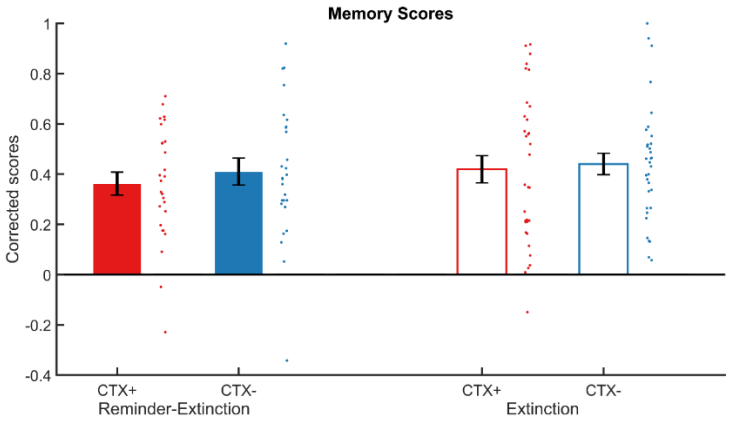


Figure 2.5. An isolated reminder before extinction did not affect memory of the location of items in each context. Bar plots reflect mean scores on the item location memory test for the threat (CTX+, red) and safe context (CTX-, blue) in the Reminder-Extinction (solid bars) and Extinction (open bars), tested at the end of the experiment. Participants remembered items from both contexts above chance level and there were no differences in location memory between contexts. A score of 0 indicates chance level. Error bars = s.e.m., adjacent dots represent jittered individual data-points.

Subjecting item-location memory scores (see Figure 2.5) were subjected to a group (R-Ext, Ext) x context (CTX+, CTX-) rmANOVA revealed no interaction or main effects of group (all P's > 0.5) and context (p>0.17). To explore whether memory scores were above chance level, mean memory scores across groups and context were subject to a one-sample t-test. Mean memory scores were above chance level of 0, a statistically significant difference of 0.41 (95% CI, 0.34 to 0.47), t(59) = -2.381, p<0.001. Hence, participants remembered the location of the items in the rooms but contextual conditioning nor the reminder-extinction procedure affected memory.

Avoidance

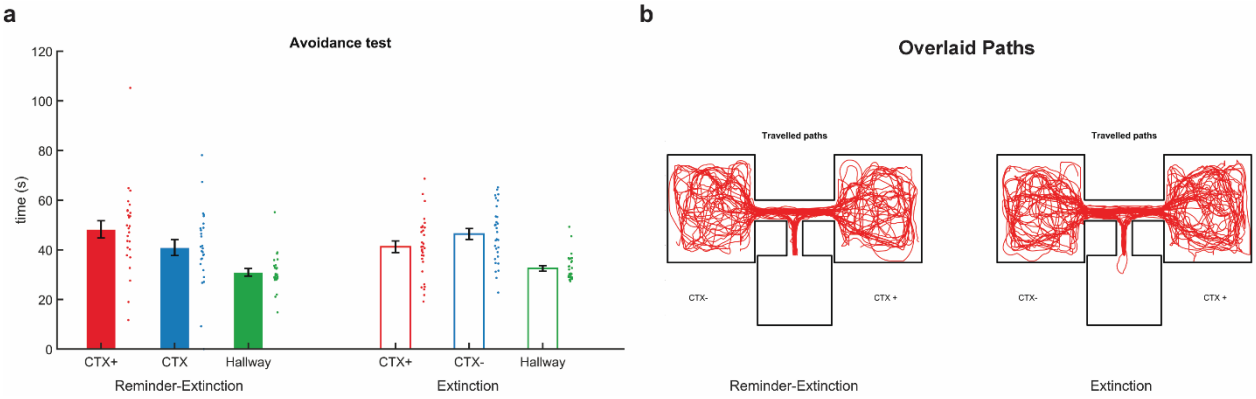


Figure 2.6. An isolated reminder before extinction did not affect free exploration behavior after re-exintinction of contextual threat conditioned responses. (A) Bar plots reflect mean time spent in the threat (CTX+, red), safe context (CTX-, blue), and hallway (green) for the Reminder-Extinction (solid bars) and Extinction groups (open bars) tested at the end of the experiment. Participants spent similar amounts of time in the CTX+ and CTX-. Error bars = s.e.m., adjacent dots represent jittered individual

data-points. (B) Travelled paths are similar in the Reminder-Extinction and Extinction group, and similar for the CTX+ and CTX-. Individual travelled paths are mirrored for a subset of particpants to display the CTX+ on the left for all participants.

After the reinstatement test under extinction conditions and the spatial memory test, participants freely navigated through the contexts while performing a cover task that required exploration of the CTX+ and CTX-. To test whether return of contextual threat memory was associated with avoidance behaviour, the time spent in each context was compared. A group (R-Ext, Ext) x context (CTX+, CTX-) rmANOVA revealed no interaction or main effects of group or context, indicating that participants did not avoid the threat conditioned context (all P's > 0.08) (Figure 2.6). To further investigate whether a reminder before extinction might reduce avoidance of the threat conditioned context, we carried out a Chi-square test to check whether there was a difference between the first room that was entered (CTX+ or CTX-) between the two groups (R-Ext, Ext). In our pre-registration, we had planned to carry out a Fisher's exact test, but given the fact that Fisher's exact test is only used when at least one of the four cells of a 2x2 table contains less than five observations, and all of our cells had at least 12 observations, we decided a Chi-square test was more appropriate. The Chi-square test revealed no differences between groups in the first room that was visited ($X^2(2)> = 0.606$, $p = 0.604$) (Table 2.1). Thus, participants explored both rooms equally and neither contextual conditioning nor the reminder-extinction procedure affected exploration behaviour.

			First Room Visited		**Total**
			CTX-	CTX+	
	Reminder-Extinction	Count	12	15	**27**
		Expected Count	13.3	13,5	**27**
	Extinction	Count	18	15	**33**
		Expected Count	16.5	16.5	**33**
Total		**Count**	**30**	**30**	**60**
		Expected Count	**30**	**30**	**60**

Table 2.1. An isolated reminder before extinction did not affect the likelihood of avoiding the CTX+ on first entry. Crosstabulation of the first room visited for the Reminder-Extinction and Extinction group during free exploration at the end of the experiment.

Skin conductance and heart-rate responses

As secondary measures of physiological responses we also obtained skin conductance and heart-rate responses. However, as these showed weak contextual threat conditioned responses at best we were unable to assess the influence of the reminder-extinction procedure on these measures. For completeness we have included the results of these measures in the Supplementary Information.

Deviations of pre-registered design

In our pre-registered design, we planned to include an additional control group to test whether the effect of a reminder on the return of contextual threat conditioned memory was time-dependent. As the reminder-extinction procedure is thought to modify memory through a reconsolidation-update

mechanism, we wanted to include an immediate memory test to gather evidence that the effect of a reminder was due to an interference with a reconsolidation process rather than immediate learning processes. This group would have been subjected to the spontaneous recovery and reinstatement tests, and all other tests planned for day 3, immediately after extinction on day 2. However, as we did not find an effect of the reminder-extinction procedure on the return of contextual threat conditioned memory, we did not test this additional control group.

Discussion

We investigated the efficacy of the reminder-extinction procedure to prevent the return of contextual threat conditioned memory in humans. On day 1, participants in both the reminder-extinction and extinction group acquired comparable discriminatory contextual threat conditioned FPS responses. Both groups exhibited initial retention during extinction on day 2 and full extinction over the course of the task. In contrast to our hypothesis, both the reminder-extinction and the extinction group showed comparable spontaneous recovery and reinstatement of FPS responses. We also found no effects of the reminder-extinction procedure on context conditioned valence and arousal ratings or explicit memory for the received shocks. Thus, we found no evidence that the reminder-extinction procedure is a more effective procedure to modify contextual threat conditioned memories in humans as compared to regular extinction.

There are several potential explanations as to why we found no effect of an isolated reminder before extinction on the return of threat responses. In line with a previous studies that observed no effect of the reminder-extinction procedure on the prevention of category conditioned threat responses (Kroes, Dunsmoor, Lin, et al., 2017) and no effect on cues presented within a contextual frame (Meir Drexler et al., 2014), the current contextual conditioning paradigm may place greater demands on hippocampal memory mechanisms than cue-conditioning, rendering the threat memory less sensitive to attenuation by the reminder-extinction procedure. This would suggest that hippocampal-dependent memories are less sensitive or even insensitive to reconsolidation-based interventions (Alberini, 2011; Kroes et al., 2016; Kroes & Fernández, 2012). In support of this hypothesis, it has been suggested that the use of expectancy ratings negatively modulates the effect of the reminder-extinction procedure on the return of fear in humans, potentially by increasing declarative awareness of contingencies and thereby increasing hippocampal-dependence (Kredlow et al., 2016). However, this explanation stands in contrast to previous studies in rodents that shown enhanced efficacy of the reminder-extinction procedure as compared to regular extinction for the attenuation of contextual fear memories in mice (Rao-Ruiz et al., 2011) and rats (Flavell et al., 2011).

Alternatively, we suggest that our findings are in line with the growing literature that is unable to replicate the reminder-extinction effects on cue-conditioned threat memories in humans (Fricchione

et al., 2016; Golkar et al., 2012; Kindt & Soeter, 2013; Klucken et al., 2016; Meir Drexler et al., 2014; Shiban et al., 2015). This raises the possibility that the reminder-extinction procedure is generally ineffective in preventing the return of conditioned threat responses, or, at best, highly dependent on potential boundary conditions (for reviews, see Auber et al., 2013; Schroyens et al., 2017; Zuccolo & Hunziker, 2019). Also, note that our main dependent measure threat potentiated startle responses differs from most previous reminder-extinction studies which have used skin conductance responses as the main dependent measure. It may be interesting for future studies to prospectively test if FSP and SCR are differentially sensitive to reminder-extinction interventions.

The limited replicability may not be limited the reminder-extinction procedure, but also seems to generalize to reconsolidation-based interventions (Chalkia et al., 2019; Schroyens, Alfei, et al., 2019). It would therefore be worthwhile to explore if other interventions such as beta-blockers (Dębiec & Ledoux, 2004; Kindt et al., 2009), propofol (Galarza Vallejo et al., 2019), electrical brain stimulation (Kroes et al., 2014), or other behavioural interventions (James et al., 2015) are capable of permanently attenuating contextual threat memories in humans. By doing so, we will hopefully reach a more mechanistic understanding of how post-retrieval interventions can impact memories.

Alternatively, even though we observe reactivation of contextual threat memory as indexed by threat-potentiated startle that, critically, did not trigger extinction learning, our reminder procedure may have failed to reactivate memory in such a way that it resulted in destabilization of the memory. If indeed the efficacy of the reminder-extinction procedure depends memory destabilization and disruption of a reconsolidation process, generation of a prediction error during the reminder may be critical for successful destabilization (Exton-McGuinness et al., 2014; Pedreira, 2004, for a review, see Exton-McGuinness et al., 2015; Sinclair & Barense, 2019). It would be of interest for future prospective studies to investigate the conditions that result in the destabilization of contextual threat memories in humans. Another explanation is that because we observed an unexpected difference between groups at the start of acquisition, the reminder-extinction group may have conditioned more strongly and the lack of a difference between groups could potentially reflect a diminishment of threat recovery in the reminder-extinction group after all. However, such an explanation does not fit with our a priori hypotheses. Moreover, considering that we randomly assigned participants to either group it is surprising to observe group differences during the initial phase of acquisition. This group difference was driven by a difference in responses in the CTX- room, not CTX+, during the early phase of acquisition. At the end of acquisition and at the start of extinction both groups show comparable differential contextual threat conditioned responses. To us this suggests that both groups acquired, consolidated, and retained comparable differential contextual threat conditioned responses, rendering this alternative explanation unlikely.

A potential limitation of the current study is that both groups underwent an equal number of visits to the conditioned context during extinction training, and the reminder visit, also carried out under extinction conditions, thus constitutes additional exposure in the R-Ext group. However, we found that the final trials of extinction training seem to have a negligible (i.e. non-significant) contribution to extinction learning. In addition, if the additional exposure to the CTX+ under extinction conditions would have had an effect, we would expect to find attenuated spontaneous recovery and reinstatement in the R-Ext group, which we did not observe. A further limitation may be that the sample size of the current study is similar to previous studies investigating the effect of the reminder-extinction procedure on cue-conditioning in humans (Schiller et al., 2010; Soeter & Kindt, 2011). For studies in humans, a meta-analysis by Kredlow et al. has reported a significant, small-to-moderate effect of the reminder-extinction procedure for further reducing the return of fear in humans as compared to standard extinction (Kredlow et al., 2016). Note that synthetic upsampling of our data to N=30 per group did not reveal any differences between groups in the return of contextual threat responses (not reported), limiting the likelihood that the lack of group differences stem from limited power. Nevertheless, for future studies, using increased sample sizes would contribute to a more convincing (non-) replication of the original findings (Brandt et al., 2014).

For the translation from laboratory research on reconsolidation to clinical applications, it is relevant to keep in mind that symptoms in stress- and anxiety-related disorders are not limited to maladaptive threat responses but also include subjective feelings, episodic memories, and avoidance behaviours. In the current study we found no evidence that conditioned arousal and valence ratings were diminished after the presentation of a brief reminder before extinction. We also probed the influence of the reminder-extinction procedure on episodic memory, and found that participants in both groups were equally able to retroactively estimate the number of shocks and the reinforcement rate they had experienced. We also did not find an effect of an isolated reminder before extinction on item-location spatial memory. These findings are in contrast to human and rodent studies which indicate that reconsolidation-based interventions can impair episodic (Hupbach et al., 2007; Kroes et al., 2010, 2014; Meir Drexler et al., 2014; Schwabe & Wolf, 2009; Vallejo et al., 2019) and spatial memories (Kim et al., 2011; Morris et al., 2006). They are also in contrast to previous studies in rodents and humans that suggest that a reminder in the absence of reconsolidation-interventions, or when interventions fail, can strengthen aversive episodic and spatial memories (Inda et al., 2011; Kroes et al., 2014; Kroes, Dunsmoor, Lin, et al., 2017). Yet, our findings may be in line with previous studies showing that reconsolidation-based interventions leave explicit knowledge about contingencies intact (Kindt et al., 2009). Hence, our results suggest that the reminder-extinction procedure fails to attenuate subjective feelings and episodic memories related to an aversive context.

At the start of the experiment, we explicitly instructed participants that their memory for item location would be tested. As a result, item-location memory may be strongly encoded and less sensitive to reconsolidation(-interventions) than incidentally encoded memories (Kroes, Dunsmoor, Lin, et al., 2017, but see Kroes et al., 2014) for instructed memory test albeit with weak memory performance). In addition, we found no emotional enhancement effect (Cahill & McGaugh, 1998; Christianson & Loftus, 1987) of contextual threat conditioning on spatial item-location memory, akin to previous work on context conditioning in humans (Kroes, Dunsmoor, Mackey, et al., 2017). This might be because the location of individual items carries little predictive value in the contextual threat learning experience, which may require a conjunctive representation of the whole space (O'Reilly & Rudy, 2001), highlighting the complicated interaction between anticipation, attention and arousal on memory (Dunsmoor, Kroes, Murty, et al., 2019). Therefore, the lack of an effect of the reminder-extinction procedure on item-location memory may alternatively be explained by the suggestion that interventions targeting reconsolidation may only reduce the emotional enhancement of episodic memories (Kroes et al., 2014; Kroes & Fernández, 2012).

Given that avoidance behaviour can diminish before explicit threat expectancies have changed (Soeter & Kindt, 2015), we also investigated whether a reminder before extinction could reduce avoidance of the threat-conditioned context by tracking participants' free exploration of the contexts after spontaneous recovery and reinstatement of the conditioned threat response. We did not find evidence for avoidance of the threat-conditioned context, as participants in both the reminder-extinction and extinction group spent comparable amounts of time in both the threat-conditioned and the safe context and were equally likely to visit either context first. Given the lack of avoidance behaviour, we are unable to say whether a reminder before extinction could affect avoidance behaviour. However, as the avoidance test was conducted after spontaneous recovery and reinstatement tests that were carried out under extinction conditions, it may not be surprising that our test did not trigger avoidance. Regardless, we think such avoidance test is an interesting new tool for the emotional memory field when tested immediately after contextual threat conditioning. Especially in light of the recent finding that a beta-adrenergic reconsolidation-intervention allowed people with spider phobia to overcome avoidance behaviours, upon which their subjective feelings of threat also diminished (Soeter & Kindt, 2015), highlighting the interaction between threat-related defensive responses, avoidance behaviours, and cognitive representations of fear (LeDoux & Pine, 2016).

In conclusion, we did not find evidence for the prevention of the return of contextual threat memories using the reminder-extinction paradigm in humans. At present, it is unclear whether this could be because the reminder-extinction procedure is ineffective in modifying hippocampus-dependent contextual threat memories specifically, or threat memories more generally. It would therefore be

worthwhile to explore if other interventions such as beta-blockers (Dębiec & Ledoux, 2004; Kindt et al., 2009), propofol (Galarza Vallejo et al., 2019), electrical brain stimulation (Kroes et al., 2014), or other behavioural interventions (James et al., 2015) are capable of permanently attenuating contextual threat memories in humans. Alternatively, even though we observed reactivation of contextual threat memory as indexed by threat-potentiated startle, our reminder procedure may simply have failed to reactivate memory in such a way that it resulted in destabilization of the memory. Many explanations and boundary conditions to the (non-)destabilization of memory can be proposed post-hoc, yet the exact conditions that allow the reactivation and destabilization of memories are rarely experimentally and prospectively investigated (Cristea & Naudet, 2019; Hardwicke et al., 2016; Vallejo et al., 2019), particularly in humans. We therefore strongly encourage future studies to investigate the mechanisms underlying the destabilization of memories. Regardless of such future investigations, our current results indicate that the reminder-extinction procedure may have limited translational value for the treatment of stress- and anxiety-related disorders.

Supplementary Information – Chapter 2

Acquisition and extinction of fear potentiated startle responses

On day one, we observed comparable acquisition of discriminatory contextual threat conditioned startle responses between both groups (Figure 2.2a). A group (R-Ext, Ext) x phase (early, late) x context (CTX+, CTX-) rmANOVA revealed an interaction effect of group x context ($F_{1,38}$=5.037, p=0.031, η2=0.117), a main effect of phase ($F_{1,38}$=81.418, p<0.001, η2 = 0.682), and a main effect of context ($F_{1,38}$=87.074, p<0.001, η2=0.696), with no other main effects or interactions. A follow-up independent t-test on the difference between startle responses in the CTX+ versus CTX- revealed unexpected greater differential responses in the R-Ext than Ext group (t(38)=2.244, p=0.031, R-Ext: 5.0±0.70, Ext: 3.0±0.50). Another follow-up independent samples t-test revealed no group differences in FPS responses in the CTX+ (t(38)=1.082, p=0.286, R-Ext: 53.2±0.42, Ext: 52.5±0.44) but a difference in startle responses in the CTX- at trend (t(38)=-1.998, p=0.053, R-Ext: 48.2±0.51, Ext: 49.5±0.38). Across both groups we observed greater startle responses across both contexts in the early than late phase (t(39)=8.692, p<0.001, early: 54.1±0.41, late: 47.5±0.46), consistent with the normal habituation of startle responses over time. Importantly, across both groups we observed greater startle in the CTX+ than CTX- (t(39)=9.001, p<0.001, CTX+: 52.9±0.30, CTX-: 48.8 ± 0.34), indicating that both groups acquired differential contextual threat conditioned responses. Yet, as we observed an unexpected group x context effect we decided to explore this potential group difference further and tested FPS seperately for the early and late phase of acquisition, and compared startle responses in the CTX+ and CTX- with responses in the hallway as a control condition. A group (R-Ext, Ext) x context (CTX+, CTX-) rmANOVA for the early phase of acquisition revealed a trend-level interaction effect of group x context ($F_{1,38}$=53.311, p=0.077, η2=0.080) and a main effect of context ($F_{1,38}$=53.787, p<0.001, η2 = 0.586). A group (R-Ext, Ext) x context (CTX+, CTX-) rmANOVA for the late phase of acquisition only revealed a main effect of context ($F_{1,38}$=30.467, p<0.001, η2=0.445) and no trend for group x context interactions ($F_{1,38}$=1.643, p=0.208, η2=0.041). Thus, critically, in the late phase of acquisition, both groups showed comparible differences between startle responses in the CTX+ and CTX- indicating comparable acquisition of contextual conditioned threat responses.

On day two, both groups underwent succesful extinction of contextual threat conditioned FPS, which was preceded by an isolated reminder for the R-Ext group. During the reminder, we observed greater startle responses in the CTX+ than hallway (t(20)=3.114, p=0.005, startle responses of 61.8 ± 2.8 in the CTX+ and 49.26 ± 1.8 in the hallway, as participants did not traverse the CTX- during the reminder, a comparison between FSP in the CTX+ and CTX- was not possible). Thus, the reminder resulted in reactivation of the contextual threat conditioned memory in the R-Ext group. During the extinction task, both groups exhibited comparable extinction of FPS responses (Figure 2a). A group (R-Ext, Ext) x phase (early, late) x context (CTX+, CTX-) rmANOVA on FPS responses during the extinction task revealed an interaction of phase x context ($F_{1,37}$=8.552, p=0.006, η2=0.188) and a main effect of phase ($F_{1,37}$=217.726, p<0.001, η2=0.855), with no other main effects or interactions. Although there was a main effect of context at trend (p=0.058), there were no main or interaction effects of group (All p's>0.16). Follow-up paired t-tests revealed greater differential FPS responses in the early phase compared to the late phase (t(38)=2.787, p=0.008, early: 2.58±1.0, late: 0.18±0.46). Specifically, FPS responses in the CTX+ were greater than in the CTX- in the early phase (t(39)=2.185, p=0.035, CTX+: 55.5±0.80, CTX-: 53.3±0.54) but not the late phase (t(39)=-0.393, p=0.696, CTX+: 45.6±0.32, CTX-: 45.8±0.37), indicating that both groups intially exhibited retention of contextual threat conditioned FPS responses that fully extinguished over the course of the extinction task.

On day three, spontaneous recovery of FPS was tested under extinction conditions. To examine whether the lack of differential startle responses in the early phase could be due to a generalization of the startle potentiation to the CTX-, we further explored these effects by comparing responses in the CTX+ and CTX- with responses in the hallway for each phase. These unplanned comparisons revealed greater responses in the CTX+ and CTX- compared to the hallway in the early phase of spontaneous recovery (t(39)=2.481, p=0.018 for comparison between the CTX+ and hallway, CTX+: 54.1±1.3,

hallway: 50.1±0.80, and t(39)=2.765, p=0.009 for comparison of the CTX- and hallway, CTX+: 54.2±1.1), but no difference between the CTX+ and hallway or CTX- and hallway in the late phase of spontaneous recovery (all Ps > 0.1).

Scoring for the Spatial Memory Task

According to our pre-registration, we planned to calculate the percentage of correct answers for items in the CTX+ and CTX-. In order to differentiate between answers that were close but did not indicate the exact location of items, and answers that were wrong, we applied a graded scoring, where exactly correct answers were worth 1 point, close answers 0.5 points, remotely correct answers 0.2 points and completely incorrect answers were worth 0.1 points (see Figure 2.7).

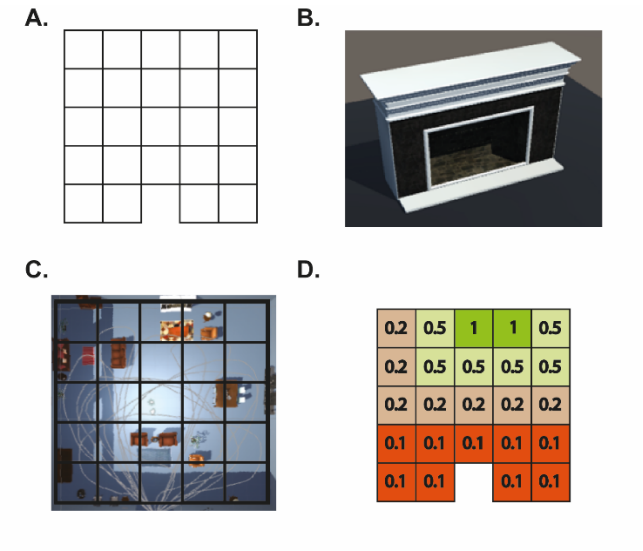


Figure 2.2. Impression of the spatial memory task. Participants were presented with an answer grid (A.) for every question and room separately. They were presented with pictures of objects that were located in the CTX+ and CTX- (B., for example), and asked to indicate the location of this object in each room on the answer grid. Correct locations were counted as any grid squares that contained a part of this object, as represented in C. A graded scoring system was applied (D.) where the exact correct location was worth 1 point and neighbouring grid locations were worth 0.5 points. Grid locations located two cells away from the correct location were worth 0.2 points and all other cells were worth 0.1 points.

As displayed in Figure 2.8, a Monte-Carlo simulation (1000 simulations) showed that the mean score based on chance level differs for the two different contexts for some of the questions. To correct for these differences, we performed a correction for chance level. Scores for each question and context for each participant were corrected by subtracting the chance-level mean and subsequently dividing by the maximum score minus the chance-level mean.

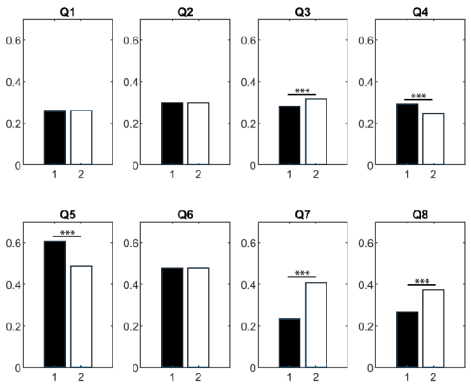


Figure 2.3. Mean scores for context one (black, 1) and context 2 (white, 2) for all 8 items of the spatial memory questionnaire resulting from a Monte-Carlo simulation. For questions 3, 4, 5, 7 and 8, the mean score at chance level is different for the two different contexts. Triple asterisk denote significance at the p<0.001 level.

Skin conductance responses

In our pre-registration, we specified that we would only include participants who showed successful conditioning during the acquisition phases, measured by a greater startle response in the threatening compared to the safe context. However, as non-responders on FPS are not necessarily the same as non-responders on skin conductance response (SCR) measures, we included for SCR analysis participants that showed a greater SCR in the threatening compared to the safe context. As we measured SCR both to the startle probe and to transitions into the contexts, we in- and excluded participants separately for the two SCR measures.

SCR in response to startle probes

For SCRs to startle probes, fourteen (out of twenty-seven) participants from the R-Ext and sixteen (out of thirty-three) participants from the Ext group showed differential SCRs during acquisition and were included in the analyses. To explore whether there were group-differences over the course of acquisition, we carried out a phase (early, late acquisition) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. This revealed a main effect of phase ($F_{1,29}$=5.997, p=0.021, η2=0.171) and of context ($F_{1,29}$=25.579, p<0.001, η2=0.469). Follow-up t-tests revealed that SCRs decreased from the early to the late phase (t(30)=2.472, p=0.019, early: 1.77±0.22, late: 1.36±0.18), and across the acquisition phase, SCRs in the CTX+ were larger than in the CTX- (t(30)=4.679, p<0.001, CTX+: 1.82±0.21, CTX-; 1.31±0.18). Both groups also showed comparable levels of extinction. A phase (early, late extinction) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA revealed an interaction effect of phase x context ($F_{1,29}$=3.962, p=0.024, η2=0.146) and a main effect of phase ($F_{1,29}$=14.278, p=0.001, η2=0.330). Follow-up t-tests revealed lower differential SCRs in the late phase of extinction as compared to the early phase (t(30)=2.336, p=0.026, early: 0.33±0.13, late: -0.07±0.10). During the early phase, SCRs to startle probes in the CTX+ were greater than in the CTX- (t(30)=2.588, p=0.015, CTX+: 2.38±0.27, CTX-: 2.04±0.24) but in the late phase there was no longer any difference (p=0.481, CTX+: 1.55±0.16, CTX-: 1.63±0.18), demonstrating that there was successful extinction of the contextual threat conditioned SCR.

A reminder did not prevent spontaneous recovery of the conditioned threat response. To test the effect of a reminder on spontaneous recovery of threat responses, SCRs were subjected to a phase (early, late recovery test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no interaction effects with or main effect of group (all Ps > 0.17), only a main effect of phase ($F_{1,29}$=16.790, p<0.001, η2=0.367) and a main effect of context at trend ($F_{1,29}$=3.979, p=0.56, η2=0.121). Follow-up t-tests showed differential responses did not change, while SCRs to in the CTX- dropped (t(30)=4.260, p<0.001, early: 2.22±0.21, late: 1.70±0.21) but not in the CTX+ (p=0.055, early: 2.26±0.26, late: 1.97±0.23. This demonstrates retention of the contextual threat conditioned SCR, as re-extinction to the CTX- occurred more rapidly than to the CTX+.

In the transition from late extinction to early spontaneous recovery, we see a generalized increase of SCRs, with no evidence for an effect of the reminder-extinction procedure. To test for the increase in SCRs, responses were subjected to phase (late extinction, early recovery test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no main or interaction effects of group (all Ps > 0.08), only a main effect of phase ($F_{1,29}$=7.748, p=0.009, η2=0.211). SCRs in both the CTX+ and CTX- were greater during early spontaneous recovery than during late extinction (t(30)=2.527, p=0.017, late extinction: 1.59±0.16, early spontaneous recovery: 2.24±0.25), which is consistent with a general increase in arousal at the start of a new experimental session.

The reinstatement test shows evidence for contextual threat conditioned SCRs in the CTX+ but does not reveal any effect of the reminder-extinction procedure. To test the effect of a reminder-extinction on reinstatement of threat responses, SCRs were subjected to phase (early, late reinstatement test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no main or interaction effects of group (all Ps > 0.08), only a main effect of phase ($F_{1,29}$=7.052, p=0.013, η2=0.196). Follow-up t-tests showed differential responses did not change, while SCRs to in the CTX- dropped (t(30)=2.430, p=0.021, early:

1.79±0.20, late: 1.34±0.13) but not in the CTX+ (p=0.083, early: 1.82±0.21, late: 1.50±0.15). Consistent with the pattern observed during spontaneous recovery, slower re-extinction to the CTX+ suggests that the contextual threat conditioned SCRs are retained.

To test for the increase in responses, SCRs were subjected to phase (late recovery test, early reinstatement test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no effects (all Ps > 0.64). As reinstated responses often extinguish rapidly, we also submitted reinstatement index scores (first trial of reinstatement test - last trial of recovery test) to a context (CTX+, CTX-) x group (reminder vs. no reminder) 2 x 2 repeated measures ANOVAs. There were no effects (all Ps > 0.76).

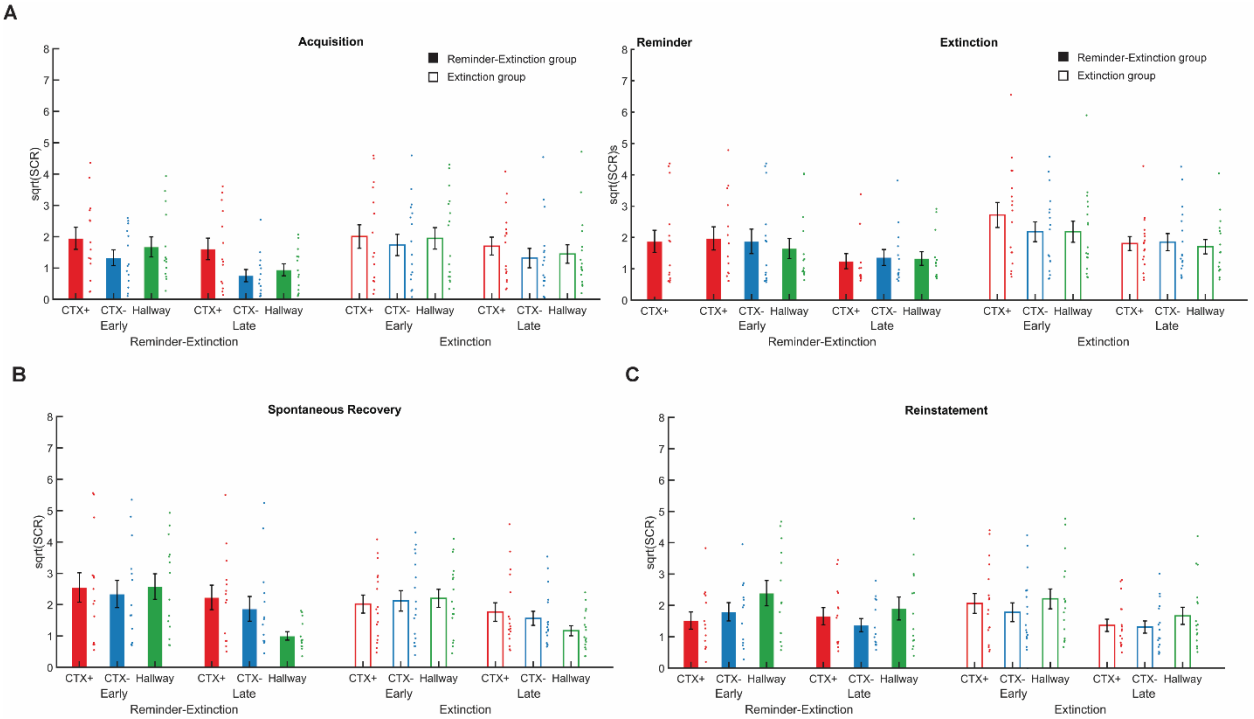


Figure 2.4. Skin conductance responses to startle probes

SCRs to transitions

For SCRs to startle probes, fifteen (out of twenty-seven) participants from the R-Ext and eighteen (out of thirty-three) participants from the Ext group showed differential SCRs during acquisition and were included in the analyses. Both groups show similar acquisition of contextual threat conditioned SCRs on day 1, as demonstrated by a phase (early, late acquisition) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA revealing only an interaction effect of phase x context ($F_{1,31}$=6.207, p=0.018, η2=0.167), a main effect of phase ($F_{1,31}$=20.043, p<0.001, η2=0.393) and a main effect of context ($F_{1,31}$=39.051, p<0.001, η2=0.557). Follow-up t-tests reveal a decrease in the differential SCR during acquisition (t(32)=2.525, p=0.017, early: 0.63±0.12, late: 0.21±0.09), while SCRs for transitions into the CTX+ remain greater than SCRs for transitions into the CTX- (t(32)=5.01, p<0.001 for the early phase, CTX+: 1.36±0.15, CTX-: 0.73±0.11, and t(32)=2.298, p=0.028 for the late phase, CTX+: 0.71±0.10, CTX-: 0.50±0.08). At the start of extinction, there was little evidence for retention for contextual conditioned SCRs to transitions into the CTX+. A phase (early, late extinction) x context (CTX+, CTX-) x group (R-Ext, Ext) 2x2x2 rmANOVA only revealed a main effect of phase at trend ($F_{1,31}$=3.963, p=0.055, η2=0.113). An exploratory t-test for differences in SCRs for transitions into the CTX+ and CTX- during the early phase of extinction revealed no difference. This suggests that as a measure of contextual threat conditioned memory, SCRs to transitions into the different contexts may be limited. This could be because participants never receive any shocks during the first five seconds in each context, as over the course of learning the latency of the SCRs shifts towards the moment at which shocks are anticipated (e.g. Prenoveau, Craske, Liao, & Ornitz, 2013). Although we carried out the pre-registered tests to

investigate spontaneous recovery and reinstatement of SCRs, these only revealed an effect of phase, and additional t-tests did not reveal any differential responses for CTX+ and CTX- and did not reveal any changes in differential responses.

To test the effect of a reminder on spontaneous recovery of threat responses, SCRs were subjected to phase (early, late recovery test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no main interaction effects of group (all Ps > 0.2), only a main effect of phase ($F_{1,31}$=6.766, p=0.014, η2=0.178). A follow-up t-test revealed a decrease in SCR responses during the spontaneous recovery test (t(32)=2.623, p=0.013, early: 1.54±0.23, late: 1.08±0.10).

To test for the increase in SCRs, responses were subjected to phase (late extinction, early recovery test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no main or interaction effects of group (all Ps > 0.17), only a main effect of phase ($F_{1,31}$=4.919, p=0.034, η2=0.137). A follow-up t-test revealed an increase in SCR responses during early spontaneous recovery as compared to late extinction (t(32)=2.262, p=0.031, late spontaneous recovery: 1.04±0.08, early reinstatement: 1.53±0.23).

To test the effect of a reminder on reinstatement of fear responses, SCRs were subjected to a phase (early, late reinstatement test) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. There were no main or interaction effects of group (all Ps > 0.19), only a main effect of phase ($F_{1,31}$=7.496, p=0.010, η2=0.195). A follow-up t-test revealed a decrease in SCR responses during the reinstatement test (t(32)=2.744, p=0.010, early reinstatement: 1.76±0.20, late spontaneous recovery: 1.08±0.10).

To test for the increase in responses, SCRs were subjected to phase (late recovery test, early reinstatement test) x context (CTX+, CTX x group (R-Ext, Ext) rmANOVA. There were no main or interaction effects of group (all Ps > 0.2), only a main effect of phase ($F_{1,31}$=12.263, p=0.001, η2=0.283). A follow-up t-test revealed an increase in SCR responses during the early reinstatement test as compared to the late phase of the spontaneous recovery test (t(32)=3.451, p=0.002, early: 1.04±0.08, late: 1.53±0.23).

As reinstated responses often extinguish rapidly, we also submitted reinstatement index scores (first trial of reinstatement test - last trial of recovery test) to a context (CTX+, CTX-) x group (R-Ext, Ext) 2x2 rmANOVA. There were no effects (all Ps > 0.16).

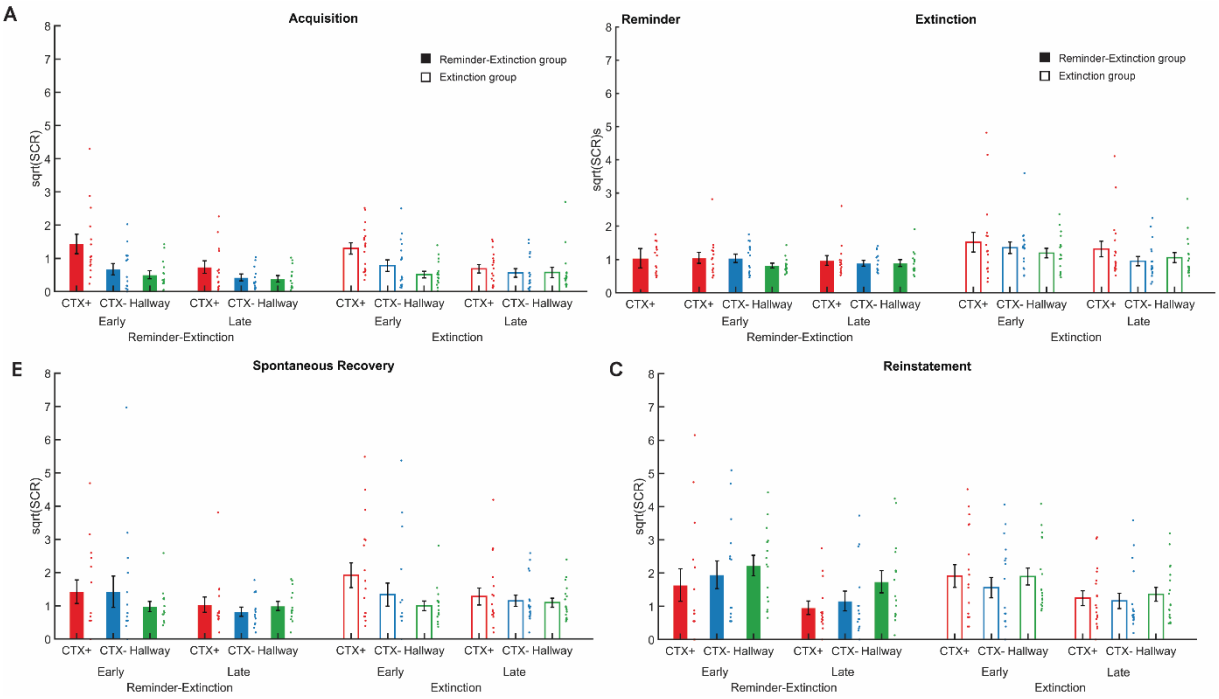


Figure 2.5. SCR in response to transitions

Heart rate changes during transitions

For both groups, heart rate shows a similar generalized conditioned threat response for both the CTX+ and CTX-. To explore whether HR responses showed evidence for conditioned threat responses to the CTX+, HR time courses were subjected to a timepoint x phase (early, late acquisition) x context (CTX+, CTX-) x group (R-Ext, Ext). This revealed a main effect of time point within the time course ($F_{2.642,124.184}$=21.325, p<0.001, η2=0.312). Follow-up t-test showed that relative to the baseline at 0s, HR at most time points (see table 2.2 below) showed a deceleration, but this deceleration is not specific to the CTX+.

Table 2.2. Significance levels for a comparison of each timepoint relative to baseline heart-rate during the early and late phase of the acquisition task. Results with significant comparisons (p<0.05) are shown in bold.

Timepoint	Early CTX+	Early CTX-	Late CTX+	Late CTX-
0.5s	t(48)=-1.911, p=0.062	t(48)=-0.114, p=0.910	t(48)=-1.523, p=0.134	t(48)=-1.333, p=0.189
1s	**t(48)=-2.145, p=0.037**	t(48)=-1.989, p=0.052	t(48)=-1.552, p=0.127	t(48)=-725, p=0.472
1.5s	**t(48)=-3.077, p=0.003**	t(48)=-1.947, p=0.057	t(48)=-1.777, p=0.082	**t(48)=-2.277, p=0.027**
2s	**t(48)=-4.074, p=0.000**	**t(48)=-2.652, p= 0.011**	**t(48)=-2.905, p=0.006**	**t(48)=-3.735, p=0.000**
2.5s	**t(48)=-3.791, p=0.000**	**t(48)=-3.806, p=0.000**	**t(48)=-3.078, p=0.003**	**t(48)=-5.131, p=0.000**
3s	**t(48)=-5.197, p=0.000**	**t(48)=-3.199, p=0.002**	**t(48)=-2.267, p=0.028**	**t(48)=-4.899, p=0.000**
3.5s	**t(48)=-5.182, p=0.000**	**t(48)=-3.150, p=0.003**	**t(48)=-3.028, p=0.004**	**t(48)=-3.362, p=0.002**
4s	**t(48)=-4.821, p=0.000**	**t(48)=-1.333, p=0.013**	**t(48)=-3.661, p=0.001**	**t(48)=-2.468, p=0.017**

At the beginning of extinction, HR in both groups shows a deceleration upon entry to the CTX+ and CTX-, but in the late phase, this is no longer the case (see table 3). To explore whether there was evidence for extinction of contextual threat conditioned HR responses to the CTX+, HR time courses were subjected to a phase (early, late extinction) x context (CTX+, CTX-) x group (R-Ext, Ext) rmANOVA. This revealed an interaction effect of phase x context x timepoint at trend ($F_{2.998,140.901}$=2.656, p=0.051, η2=0.053), an interaction effect of timepoint x group ($F_{2.693,126.593}$=4.101, p=0.010, η2=0.080), a main effect of phase ($F_{1,47}$=6.932, p=0.011, η2=0.129) and a main effect of timepoint ($F_{2.693,126.593}$=5.181,

p=0.003, η2=0.099). Follow-up t-test showed that relative to the baseline at 0s, HR at most time points (see table 2.3 below) showed a deceleration during early extinction for both CTX+ and CTX- entries, while this was no longer the case during late extinction.

Table 2.3. Significance levels for a comparison of each timepoint to baseline during the early and late phase of extinction of the contextual threat conditioned HR response. Results from significant comparisons (p<0.05) are shown in bold.

Timepoint	Early CTX+	Early CTX-	Late CTX+	Late CTX-
0.5s	**t(48)=-3.168 , p=0.003**	t(48)=-.943 , p=0.350	t(48)=1.333, p=0.189	t(48)=-0.227, p=0.821
1s	**t(48)=-3.042 , p=0.004**	**t(48)=-2.425, p=0.019**	t(48)=-0,903, p=0.371	t(48)=-0.836, p=0.407
1.5s	**t(48)=-3.244 , p=0.002**	**t(48)=- 3.999, p=0.000**	t(48)=-1.030 , p=0.308	t(48)=-1.281, p=0.206
2s	**t(48)=-2.494 , p=0.016**	**t(48)=-4.095, p=0.000**	t(48)=-1.563, p=0.125	t(48)=-0.982, p=0.331
2.5s	**t(48)=-3.378 , p=0.001**	**t(48)=-4.531, p=0.000**	t(48)=-1.705, p=0.095	t(48)=-1.500, p=0.140
3s	**t(48)=-3.680 , p=0.001**	**t(48)=-4.249, p=0.000**	t(48)=-0.917, p=0.364	t(48)=-0.801 , p=0.427
3.5s	**t(48)=-3.089 , p=0.003**	**t(48)=-4.585, p=0.000**	t(48)=-1.523, p=0.134	t(48)=-1.852 , p=0.070
4s	**t(48)=-3.844 , p=0.000**	**t(48)=-4.661, p=0.000**	t(48)=-1.419 , p=0.162	t(48)=-2.380 , p=0.021

During the late phase of spontaneous recovery, we observed a stronger deceleration upon entry of the CTX+ as compared to the CTX-, but this deceleration was not affected by a reminder. To test the effect of a reminder on spontaneous recovery of fear responses, HR time courses were subjected to phase (early, late recovery test) x context (CTX+, CTX-) x group (R-Ext, Ext) 2x2x2 rmANOVA. There was an interaction effect of phase, context and time-point ($F_{2.583,118.819}$=3.645, p=0.019, η2=0.073). We ran follow-up paired t-tests comparing HR in the CTX+ and CTX- for each time points separately for the early phase and the late phase, and found stronger deceleration for the CTX+ in the late phase for the 1.5s, 2.5s, 3s time points, (t(47)=-2.196, p=0.033, t(47)=-2.066, p=0.044 and t(47)=-2.159, p=0.036 respectively). This suggests that there is retention of a contextual threat conditioned HR response specific for the CTX+. There were no group interactions (all Ps > 0.2).

To test for spontaneous recovery of conditioned threat responses HR, responses were subjected to phase (late extinction, early recovery test) x context (CTX+, CTX-) x group (R-Ext, Ext) 2x2x2 rmANOVA. There was an interaction effect of phase x context x time-point ($F_{3.318,152.641}$=3.591, p=0.012, η2=0.072), and an interaction of context x time-point ($F_{2.701,124.259}$=3.189 =, p=0.031, η2=0.065). We ran follow-up paired t-tests comparing HR in the CTX+ and CTX- for each time points separately for the late phase of extinction and the early phase of spontaneous recovery, but we did not find any differences. There were no group interactions (all Ps > 0.5).

We also did not find any evidence for an effect of reinstatement on HR time courses. To test the effect of a reminder on reinstatement of contextual threat conditioned HR responses, HR time courses were subjected to phase (early, late reinstatement test) x context (CTX+, CTX-) x group (R-Ext, Ext) 2x2x2 rmANOVA. There were no effects (all Ps >0.06). To test for the increase in responses, HR time courses were subjected to phase (late recovery test, early reinstatement test) x context (CTX+, CTX-) x group (R-Ext, Ext) 2x2x2 rmANOVA. There were no effects (all Ps > 0.09).

Chapter 3. Reconsolidation-extinction in rodents: A reminder before extinction failed to prevent the return of conditioned threat responses irrespective of threat memory intensity in rats

Maxime C. Houtekamer, Marloes J.A.G. Henckens, Koen P. van den Berg, Judith Homberg, Marijn C.W. Kroes

Abstract

After retrieval, reactivated memories may destabilize and require restabilization processes to persist, referred to as reconsolidation. The reminder-extinction procedure has been proposed as a behavioral reconsolidation-based intervention to persistently attenuate threat conditioned memories. After presentation of a single reminder trial, the conditioned threat memory may enter a labile state, and extinction training during this window can prevent the return of conditioned threat responses. However, findings on this reminder-extinction procedure are mixed and its effectiveness may be subject to boundary conditions, including memory strength. Here, we systematically investigate whether more intense threat memories are less susceptible to disruption through a reminder-extinction procedure. Using a Pavlovian auditory threat conditioning procedure at three different shock intensities, rats acquired conditioned threat responses of variable 'strength'. Rats subsequently underwent either extinction preceded by a reminder or standard extinction. Although different shock intensities led to different strength threat memories, all groups showed reinstatement of conditioned threat responses irrespective of shock intensity or reminder condition. Hence, regardless of the intensity of the threat memory, the reminder-procedure was ineffective in preventing the return of threat responses in rats. We thus find no evidence that threat memory intensity is a potential modulator of the effectiveness of the reminder-extinction procedure.

Introduction

Upon reactivation, consolidated memories can re-enter a temporary period of lability requiring restabilization processes to persist, referred to as reconsolidation (Nader et al., 2000). Interventions targeting reconsolidation have the potential to persistently impair the reactivated memory (Nader & Hardt, 2009). As a result, reconsolidation-targeting interventions have been heralded as an opportunity to permanently change memories that contribute to stress- and anxiety-related disorders (Kroes et al., 2016). Yet, it remains an open question whether reconsolidation-targeting interventions can modify strong memories generated by highly aversive experiences, which characterize stress- and anxiety-related disorders.

Reconsolidation has been extensively studied using auditory threat conditioning (Beckers & Kindt, 2017; Nader et al., 2000; Nader & Hardt, 2009), and predominantly in rodents. In this paradigm, an auditory tone (conditioned stimulus, CS+) is coupled with an intrinsically aversive stimulus such as an electric shock (unconditioned stimulus, US), such that the CS+ by itself comes to evoke a conditioned threat response (e.g. freezing in rodents). After consolidation of the conditioned threat memory, its reactivation by presentation of a single unreinforced CS+ can once again render the memory sensitive to interventions that can persistently prevent the renewal, spontaneous recovery and/or reinstatement of conditioned threat responses (Debiec et al., 2002; Duvarci & Nader, 2004; Nader et al., 2000). Early experiments using protein-synthesis inhibitors suggested that after reactivation, memories enter a labile state and require de novo protein synthesis to persist. According to this reconsolidation account, administration of protein synthesis inhibitors after presentation of a reminder can persistently attenuate memories. Alternative accounts suggest that post-retrieval amnesia may be transient, and could reflect enhanced extinction or state-dependency (Alfei et al., 2020; Gisquet-Verrier et al., 2015; Lattal & Abel, 2004; Lewis, 1979, for reviews on alternative accounts, see e.g. Cahill & Milton, 2019; Gisquet-Verrier & Riccio, 2018). Most studies have used pharmacological interventions to target reconsolidation, yet more recently behavioral interventions, including the reminder-extinction paradigm (Monfils et al., 2009), have been proposed that may be preferable for clinical translation to patients as they are less invasive, safer, and more equitable (Kroes & Liivoja, 2018).

The behavioral reminder-extinction paradigm is a reconsolidation-targeting variant of standard extinction. In standard extinction procedures, after initial conditioning, the repeated presentation of the CS+ without aversive reinforcement results in a reduction of threat responses. Yet because standard extinction creates a novel safety memory that inhibits rather than erases the original threat memory, threat responses can recover over time or following stressful experiences (Bouton, 2002; Myers & Davis, 2002). This may explain the return of symptoms in patients with stress- and anxiety-

related disorders following exposure treatments that are based on the principles of Pavlovian extinction (Vervliet et al., 2013). In contrast, in the reminder-extinction procedure, the reminder is thought to reactivate the original threat memory rendering it labile, and as a consequence extinction training following the reminder may overwrite the original threat memory (Monfils et al., 2009). Several studies have found that the reminder-extinction procedure can prevent the recovery of threat responses, both in rodents (Auchter et al., 2017; Baker et al., 2013; Clem & Huganir, 2010; Flavell et al., 2011; Jones et al., 2013; Monti et al., 2017; Olshavsky et al., 2013; Pattwell et al., 2016; Piñeyro et al., 2014; Rao-Ruiz et al., 2011) and humans (Agren et al., 2012; Björkstrand et al., 2015; Chen et al., 2021; Feng et al., 2015; Hu et al., 2018; Johnson & Casey, 2015; Oyarzún et al., 2012; Schiller et al., 2013; Schiller et al., 2010; Steinfurth et al., 2014; Thompson & Lipp, 2017). However others have not replicated this result and observed a return of threat responses, both in rodents (Chalkia et al., 2020; Costanzi et al., 2011; Goode et al., 2017; Gräff et al., 2014; Ishii et al., 2012, 2015; Luyten & Beckers, 2017) and humans (Drexler et al., 2014; Fricchione et al., 2016; Golkar et al., 2012; Houtekamer et al., 2020; Kindt & Soeter, 2013; Klucken et al., 2016; Kredlow et al., 2018; Kroes et al., 2017; Ponnusamy et al., 2016; Soeter & Kindt, 2011; Zimmermann & Bach, 2020, for a review see Kredlow et al. 2016).

Several boundary conditions to reconsolidation have been proposed that may explain mixed results on the effectiveness of the reminder-extinction paradigm, including predictability of the reminder used to trigger reconsolidation, duration of the reminder-extinction training, housing conditions, and more fundamental: memory modality, memory age, and memory strength (Auber et al., 2013; Kroes et al., 2016; Zuccolo & Hunziker, 2019). In particular, an outstanding question in the field of memory reconsolidation is whether the 'strength' of memories determines their susceptibility to modifications via reconsolidation-targeting interventions (Alberini & Ledoux, 2013a; Kroes et al., 2016; Nader & Hardt, 2009; Robinson & Franklin, 2010). It has been suggested that 'stronger' threat memories as generated by highly intense aversive experiences might be less susceptible to modification by reconsolidation-targeting interventions (Auber et al., 2013; Duvarci & Nader, 2004; Haubrich et al., 2020b; Holehonnur et al., 2016; Kredlow et al., 2016; Wang et al., 2009), compromising their treatment potential for stress- and anxiety-related disorders, typically characterized by such strong memories. Accordingly, several studies have investigated the effectiveness of reconsolidation interventions for attenuation of strong vs. weak memories operationalized by varying the number of CS-US pairings, and found that stronger memories are resistant to disruption through reconsolidation (Haubrich et al., 2020b; Holehonnur et al., 2016; Wang et al., 2009). Shock intensities used in studies investigating the efficacy of the reminder-extinction procedure have varied, and contrary to experimental results from studies varying the number of CS-US pairings, a meta-analysis has suggested that the reminder-extinction procedure may have a larger efficacy for studies using a higher US intensity (Kredlow et al.,

2016). However, to date no systematic study of threat intensity on memory malleability during reconsolidation has been performed.

In this pre-registered study (Houtekamer et al., 2021), we investigated whether the reminder-extinction procedure can prevent the return of conditioned threat responses for threat memories of varying intensity. To operationalize the intensity of threat memories we used a Pavlovian threat conditioning procedure with three levels of shock intensity in which rats (n=78) were conditioned at either low, medium or high shock intensity (Phillips & LeDoux, 1992). To assess conditioned threat responses, freezing levels were measured during the presentation of the conditioned stimulus, an auditory tone. Next, in a between-subjects design, rats either underwent a reminder-extinction or standard extinction protocol. Critically, to test whether more intense threat memories would limit the efficacy of the reminder-extinction procedure to prevent the return of threat, we tested for the reinstatement of freezing responses at a subsequent long-term memory test. We found that the reminder-extinction procedure failed to prevent the recovery of threat responses regardless of threat memory intensity.

Results

Tone habituation

Following initial contextual habituation to the conditioning chambers, rats were habituated to the conditioned stimulus (CS+, tone) to ensure that rats showed comparable and minimal baseline freezing across conditions. Equivalent habituation to the CS+ was confirmed by the absence of any main or interaction effects of the assigned shock intensity (low, medium, high) or reactivation (reminder-extinction, extinction) conditions during two consecutive habituation sessions (all p's>0.1).

Acquisition of Pavlovian threat conditioned responses at different shock intensities

Rats in all groups acquired conditioned threat responses to the CS+ (auditory tone) where higher shock intensities resulted in more freezing (see Figure 3.1). Critically, a reactivation (Rem-Ext, Ext) x shock intensity (low, medium, high) x trial rmANOVA on freezing scores for the CS+ presentations during the acquisition task revealed a main effect of shock intensity ($F_{(2,72)}$=7.434, p=0.001, η_p^2 = 0.171), a main effect of trial ($F_{(2,144)}$=189.303, p<0.001, η_p^2 = 0.724), and an interaction effect of trial and shock intensity ($F_{(4,114)}$=4.073, p=0.004, η_p^2 = 0.102), with no other main effects or interactions (all p's > 0.4). Across all groups, freezing levels increased with each trial (t(77)=-9.740, p<0.001 for trial 1 to 2 and t(77)=-8.993, p<0.001 for trial 2 to 3). A follow-up one-way ANOVA for freezing levels at each trial revealed effects of shock intensity at the first trial ($F_{(2,75)}$=4.071, p=0.021) and the third trial ($F_{(2,75)}$=11.349, p<0.001). Unexpectedly, freezing levels were higher at medium shock intensity than at low and high shock intensity during the first trial (t(50)=2.32, p=0.024 for medium compared to low and t(50)=2.25, p=0.029 for medium compared to high shock intensity, low: 5.6%±1.5%, medium:

12.9%±2.8, high:5.2%±2.0%). As animals had shown equivalent habituation and had not yet experienced a shock at the time of the first CS+ presentation this unexpected group difference likely reflects a chance effect. Critically, during the final trial of conditioning, freezing scores were lower in the low shock intensity group ($t(50)=-3.51$, $p=0.001$ and $t(50)=-4.0$, $p<0.001$ compared to the medium and high shock respectively, low: 53.3%±7.0%, medium: 82.1%±4.4%, high: 83.7%±3.2%), but not different between the medium and high shock intensity groups ($t(50)=-0.286$, $p=0.886$). Comparable freezing between the medium and high group might be the result of freezing scores during acquisition having reached ceiling levels. Regardless, all groups thus showed successful acquisition of the conditioned threat response and low-intensity resulted in lower levels of freezing.

Reactivation of the conditioned threat response

The next day, rats in the Rem-Ext group were reminded by a single unreinforced presentation of the CS+, while rats in the Ext group were not (see Figure 3.1). After a 10-minute break (Rem-Ext), during which the rats were returned to their home cage, animals (all groups) were returned to the conditioning chamber for extinction training during which the CS+ was presented either 17 (Rem-Ext) or 18 (Ext) times without US reinforcement, to equalize the total number of CS+ re-exposures.

To explore whether the reminder trial evoked a conditioned threat response of comparable strength as the response at the end of the acquisition task, we carried out a repeated measures ANOVA for the last trial of acquisition and the reactivation trial with shock intensity as between-subjects factor. This

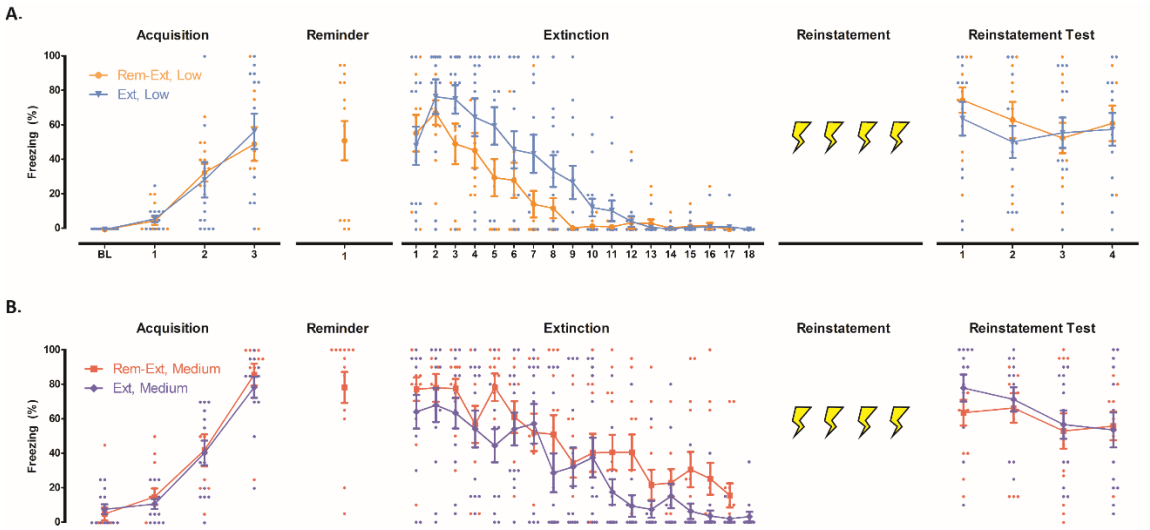


Figure 3.1. Freezing levels during the acquisition, reminder extinction and reinstatement of cue-conditioned Pavlovian threat responses. During the three CS+ presentations co-terminating with the US, rats in all experimental groups (6 groups, n = 13 per group) acquired a cue-conditioned threat response reflected in increased levels of freezing (**A, B, C, acquisition**, see SI for a more detailed description of the acquisition of conditioned threat responses in individual animals). Freezing responses were stronger in groups exposed to medium (**B**) and high (**C**) shock intensities as compared to low (**A**) shock intensity. The following day, rats assigned to the reminder-extinction (Rem-Ext) groups were placed back in the conditioning chamber and presented with one CS+, without US (**A, B, C, reminder)**. 10 minutes after the retrieval trial, animals underwent extinction training through 17 (Rem-Ext) or 18 (extinction (Ext)) uncoupled presentations of the CS+, resulting in decreased freezing levels (**A, B, C, extinction)**. On day three, animals received four unsignalled shocks in a reinstatement procedure. All groups showed reinstated freezing during the reinstatement test on day five **(A,B,C, reinstatement test).** All groups are shown as separate lines. Data presented as mean ± S.E.M and dots represent individual datapoints

revealed a main effect of shock intensity ($F_{(2,36)}$=6.374, p=0.004, η^2 = 0.262), without an effect of session (p=0.430) or session x shock intensity interaction (p=0.716). Post-hoc tests revealed lower levels of freezing in the low-intensity group as compared to the medium (p<0.001, low: 53.8%±6.8%, medium: 80.3%±4.3%) and high intensity group (81.8%±2.7%), while freezing levels in the medium and high did not differ (p=0.825). Hence, we observed reactivation of threat memory.

Specificity of the conditioned threat response

To investigate whether rats expressed threat-responses to the CS over and beyond threat responses to the context itself, we scored freezing levels during a 20-second window prior to the onset of the first CS. We found that during the reminder phase, pre-CS freezing levels (i.e., contextual threat responses) across groups were rather substantial (pre-CS: 53.6%±5.8%), but significantly lower than CS-evoked freezing levels (t(38)=-4.674, p<0.001, CS-evoked:68.7%±5.5%). Pre-CS freezing during the extinction phase was similarly lower than CS-evoked freezing levels (t(75)=-5.636, p<0.001, pre-CS: 50.1%±4.5%, CS-evoked: 67.3%±3.8%). Pre-CS levels of freezing during the reinstatement test, however, did not significantly differ from freezing levels during the first CS-presentation (*Z*=-1.172, *p*=0.241, pre-CS: 69.5%±4.0%, CS-evoked: 72.7%±3.1%). We explored freezing during the reinstatement test in more detail by also measuring freezing levels in the 20s window prior to each CS onset. This analysis indicated that across the full reinstatement test average pre-CS freezing levels were significantly lower than CS-evoked freezing (t(77)=-4.232, p<0.001, pre-CS: 58.4%±2.5%, CS-evoked: 66.05%±2.4%). A stimulus (pre-CS, CS-evoked) x trial (1-4) x shock intensity (low, med, high) x group (Rem-Ext, Ext) revealed a main effect of stimulus ($F_{(1,68)}$=16.744, p<0.001, η2=0.198) and decrease in freezing over trials ($F_{(3,204)}$=6.041, p<0.001, η2 =0.082) with no other effects or interactions (all p's>0.1). Hence, during the long-term memory test for reinstatement the rats displays threat responses to the conditioned cue above and beyond threat responses to the context. Thus, rats shows reinstatement of the acquired cue-conditioned, not context-related, threat responses irrespective of reminder cue before extinction.

Extinction of the conditioned threat responses

In line with previous studies (Luyten & Beckers, 2017; Monfils et al., 2009), we explored whether all groups show equivalent extinction. To check for equivalent threat recall and extinction, freezing scores of all 18 retrieval-extinction trials (R-EXT group 1 reminder + 17 extinction trials, EXT group 18 extinction trials) were submitted to a repeated measures ANOVA with trial as within subject factor and shock intensity (low, medium, high) and reactivation (Rem-Ext, Ext) as between-subjects factors (Luyten & Beckers, 2017; Monfils et al., 2009). Freezing levels during extinction changed over trials ($F_{(8.408,571.714)}$=53.647, p<0.001, η2 = 0.441) and were affected by shock intensity ($F_{(2,68)}$=9.342, p<0.001, η2 = 0.216) and an interaction between reminder and shock intensity ($F_{(2,68)}$=5.236, p=0.008, η2 =

0.133). A follow-up t-test across groups showed the expected decrease in freezing responses from the first to the last trial of extinction ($t(73)=12.814$, $p<0.001$). The low intensity group showed lower levels of freezing compared to the medium intensity group ($t(50)=-3.785$, $p=0.001$, low: 24.5%±2.3%, medium: 40.5%±3.6%) and high intensity group ($t(48)=-4.226$, $p<0.001$, high: 46.6%±4.9%), but freezing levels were not different between the medium and high intensity groups ($t(48)=-1.033$, $p=0.307$). More importantly, we sought to explain the interaction effect between reminder and US intensity. Freezing rates in the Rem-Ext and Ext groups were comparable at low and high shock intensities ($p=0.088$ and $p=0.796$ respectively) but differed for medium intensity, at which the Rem-Ext groups displayed higher freezing ($t(24)=2.752$, $p=0.011$, Rem-Ext: 49.2%±4.8%, Ext: 31.8%±4.2%).

As pre-registered and in line with previous studies (Luyten & Beckers, 2017; Monfils et al., 2009), freezing scores for the last four extinction trials were submitted to a repeated measures ANOVA with trial as within-subject factor and shock intensity (low, medium, high) and group (Rem-Ext, Ext) as between-subject factors. This revealed a main effect of trial ($F_{(3,204)}=2.809$, $p=0.041$, $\eta^2 = 0.040$) and shock intensity ($F_{(2,68)}=6.325$, $p=0.003$, $\eta^2 = 0.157$) and an interaction between reminder and shock intensity ($F_{(2,68)}=6.391$, $p=0.003$, $\eta^2 = 0.158$). Thus, contrary to our expectations, freezing scores during the last four trials of extinction differed for rats conditioned at different shock intensities depending on the presentation of a reminder. At low and high shock intensity, the Rem-Ext and Ext groups showed comparable levels of freezing during the last four trials of extinction (low shock intensity: $t(24)=0.125$, $p=0.901$, Rem-Ext: 1.2%±0.6%, Ext: 1.1%±0.5%, high shock intensity: $t(20)=-1.653$, $p=0.114$, Rem-Ext: 10.3%±5.9%, Ext: 30.9%±10.9%), while at medium intensity, the Rem-Ext group showed higher levels of freezing than the Ext group ($t(24)=2.651$, $p=0.014$, Rem-Ext: 23.8%±7.2%, Ext: 4.0%±1.9%). Across the last four extinction trials, freezing was significantly lower in the low intensity group as compared to the medium ($t(50)=3.075$, $p=0.003$, low: 1.1±0.4%, medium: 13.9±4.1%) and high intensity group ($t(50)=3.277$, $p=0.002$, high: 20.6±6.5%), but comparable between the medium and high intensity groups. Despite a main effect of trial, a paired t-test across all groups provided no evidence that freezing levels changed from the fifteenth until the last extinction trial (trial 15: 12.8%±3.0%, trial 18: 8.1%±2.5%).

Thus, our exploratory and preregistered analyses indicate that all groups showed extinction of conditioned threat responses. The low intensity group displayed less freezing during and at the end of extinction compared to the medium and high intensity groups, but we observed no difference between the medium and high intensity groups. Surprisingly we found less freezing during extinction and at the end of extinction for animals in the no-reminder compared to the reminder group within the medium intensity group, but no effect of reminder on extinction in the low or high intensity group.

Reinstatement and recovery of the conditioned threat response

All groups received reinstatement shocks on day 5 and were tested for reinstatement of conditioned threat responses on day 6 to test for long-term effects of the reminder-extinction procedure on threat memory. In line with previous studies (Luyten & Beckers, 2017; Monfils et al., 2009) we explored the reinstatement of threat responses by comparing the average of the last four trials of extinction and the average of the four trials of the reinstatement test in a phase (extinction, reinstatement test) x group (Rem-Ext, Ext) x shock intensity (low, medium, high) rmANOVA (Figure 3.1). This revealed a main effect of phase ($F_{(1,68)}$=241.116, p<0.001, η^2 = 0.780) and shock intensity ($F_{(2,68)}$=8.639, p<0.001, η^2 = 0.203) and a reminder x shock intensity interaction ($F_{(2,68)}$=85.903, p=0.005, η^2 = 0.143). As expected, freezing levels were higher during the reinstatement test as compared to the last four trials of extinction (t(73)=15.416, p<0.001, extinction: 11.4%±2.6%, reinstatement test: 65.1%±2.5%) indicating that conditioned threat responses were successfully reinstated across all groups. Following up on the main effect of shock intensity, we found freezing was greater in the high shock intensity group compared to the low shock intensity group (t(46)=3.821, p<0.001, low: 30.7%±2.3%, high: 47.3%±3.8%), greater for the medium intensity compared to the low intensity group at trend (t(49)=1.989, p=0.052, medium: 37.8±2.7) and lower for the medium compared to the high intensity group at trend (t(46)=-1.956, p<0.056). At low and medium shock intensity, the Rem-Ext and Ext group showed no difference in average freezing during late extinction and the reinstatement test (low intensity: t(24)=0.657, p=0.517, medium intensity: t(24)=1.374, p=0.182), yet at high intensity, the Rem-Ext group showed lower levels of freezing (t(20)=-2.640, p=0.016, Rem-Ext: 38.2%±3.5%, Ext: 56.3%±5.9%). However, across all shock intensities, the absence of an interaction effect between phase and shock intensity with reminder indicates that, although overall freezing levels are modulated by shock intensity in interaction with the reminder, the reinstatement of threat responses was comparable across shock intensities and reactivation groups.

In addition, we explored whether freezing differed between groups over the course of the reinstatement test by submitting all trials (1-4) of the reinstatement test to a group (Rem-Ext, Ext) x shock intensity (low, medium, high) rmANOVA. Freezing levels decreased over trials ($F_{(1,71)}$=9.253, p=0.003, η2 =0.115) and were different between shock intensities ($F_{(2,71)}$=3.848, p=0.026, η2 =0.098), but were not affected by previous presentation of a reminder (all p's>0.2). Rats previously conditioned at high shock intensity revealed higher levels of freezing than rats previously conditioned at low intensity ((t(50)=2.601, p=0.012) and medium intensity (t(50)=2.385, p=0.021), while freezing levels for rats conditioned at low and medium intensity did not differ (p=0.712). In conclusion, results from the reinstatement test suggest that threat responses recovered regardless of whether rats had received a

reminder before extinction or not. Furthermore, rats appeared to have retained a memory representation of the intensity of the original aversive conditioning experience.

As per our preregistration, we also investigated whether the reinstatement of threat responses might differ between groups by calculating a reinstatement index (Figure 3.2) where we subtracted the average freezing level during the last 4 trials of extinction from the average level of freezing during the reinstatement test. Reinstatement index scores were submitted to a two-way ANOVA with shock intensity and reminder as a factor, but this did not reveal any main effects or interactions ($p>0.1$), showing that the magnitude of reinstatement was not affected by shock intensity, or the presentation of an isolated reminder before extinction, irrespective of shock intensity. Thus, although we found that the reminder high intensity group displayed less freezing than the no-reminder high intensity group overall during the last trials of extinction and trials of the reinstatement test, the reminder had no effect on the reinstatement of threat responses per se (Figure 3.2).

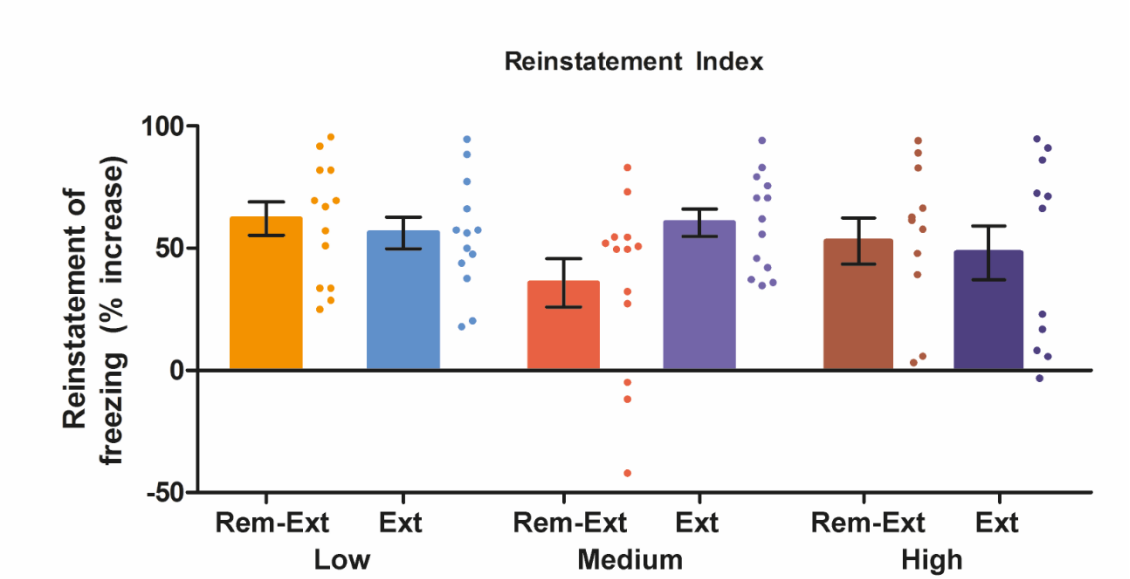


Figure 3.2. Presentation of a single isolated reminder trial before extinction did not prevent the reinstatement of conditioned threat responses, irrespective of shock intensity. A reinstatement index was calculated by subtracting freezing during the final 4 extinction trials from the average level of freezing during the reinstatement test. Reinstatement indices are displayed separately for rats conditioned at low, medium and high shock intensity for the extinction (Ext) and reminder-extinction (Rem-Ext) groups. Rats in all groups showed similar reinstatement scores, indicating that the presentation of a single CS+ as reminder did not prevent the reinstatement of a conditioned threat response, irrespective of shock intensity. Data presented as mean ± S.E.M. Dots represent individual data points.

Discussion

The present pre-registered study systematically investigated the effect of the reminder-extinction procedure on the return of threat responses for Pavlovian conditioned threat memories of different intensities in rats. Threat responses during the acquisition, reminder and extinction phases and the reinstatement test were affected by shock intensity used during the acquisition phase. Yet, we found

that the reminder-extinction procedure did not prevent the reinstatement of conditioned threat responses, irrespective of threat memory intensity. Specifically, rats assigned to both the reminder and the reminder-extinction group showed successful and comparable acquisition and extinction of Pavlovian threat conditioned memories, where observed that rats reinforced with low intensity shocks showed lower levels of freezing compared to rats reinforced with medium or high intensity shocks. Rats in the reminder-extinction group showed retention and 'reactivation' of the conditioned threat response during the reminder, again with more freezing in the medium and high intensity groups than low intensity group. At the long-term memory test following reinstatement, our reinstatement index revealed comparable return of threat responses in all groups, with the high intensity group displaying more absolute freezing than the medium and low intensity groups. Collectively, our results imply that the reminder-extinction procedure failed to prevent the return of conditioned threat responses, irrespective of threat memory intensity. We thus failed to replicate previous reports that the reminder-extinction procedure can prevent the return of threat responses in rodents (Auchter et al., 2017; Baker et al., 2013; Cahill et al., 2018; Clem & Huganir, 2010; Flavell et al., 2011; Jones et al., 2013; Jones & Monfils, 2016; Liu et al., 2014; Monfils et al., 2009; Monti et al., 2017; Olshavsky et al., 2013; Pattwell et al., 2016; Piñeyro et al., 2014; Rao-Ruiz et al., 2011). Nevertheless, our results are in line with previous non-replications, including a pre-registered direct replication attempt (Luyten & Beckers, 2017), as well as conceptual replication attempts (Chan et al., 2010; Costanzi et al., 2011; Flavell et al., 2011; Goode et al., 2017; Gräff et al., 2014; Ishii et al., 2012, 2015; MacPherson et al., 2013; Ponnusamy et al., 2016; Stafford et al., 2013; Xu et al., 2013). Our objective was to investigate whether threat memory intensity influences the efficacy of the reminder-extinction procedure in preventing the return of threat responses, in line with suggestions that memory strength may be a boundary condition to reconsolidation (Alberini & Ledoux, 2013; Auber et al., 2013; Chen et al., 2021; Kroes et al., 2016; Robinson & Franklin, 2010; Suzuki et al., 2004; Wang et al., 2009; Zuccolo & Hunziker, 2019). As we found no evidence that the reminder-extinction procedure prevented the return of threat responses, irrespective of threat memory intensity, our results provide no evidence that memory intensity is a boundary condition to the efficacy of the reminder-extinction procedure.

One explanation for our lack of an observed effect of the reminder-extinction procedure on the return of threat responses may be that our reminder failed to reactivate the threat memories and render the memories labile during a reconsolidation process. However, we found significant freezing in response to the reminder tone that reflected the intensity of the originally acquired threat memory. As previous studies that have reported positive effects of reconsolidation targeting interventions, including the reminder-extinction procedure, have used comparable reminder procedures with similar freezing responses to the reminder as evidence of memory reactivation (Baker et al., 2013; Cahill et al., 2019;

Monfils et al., 2009), we doubt whether failed reactivation can explain our results. Alternatively, even if the reminder successfully reactivated the threat memory, a prediction error may also be required for successful destabilization (Alfei et al., 2015; Chen et al., 2021; Junjiao et al., 2019; Monti et al., 2017; Pedreira, 2004; Sevenster et al., 2014, but see Cahill et al., 2018 for an opposite account, for reviews see Exton-McGuinness et al., 2015; Krawczyk et al., 2017; Sinclair & Barense, 2019). In addition, a recent study suggests that for stronger memories to destabilize a reminder should generate greater prediction errors (Chen et al., 2021). Indeed, it is a possibility that in the current experiment, the reminder did not generate sufficient prediction error to destabilize any of the threat memories. Although it could be that the memories of differential strength created in the current experiment, required different degrees of prediction error to destabilize, we may not be able to observe this effect if the generated prediction error is already too weak to destabilize the low-intensity threat memory. Nonetheless, it should be noted that the reactivation procedure used here (a single unreinforced CS presentation) would be expected to evoke a prediction error, entropy, and/or novelty signal as the US is omitted. In addition, given that we followed the reactivation procedure outlined by Monfils et al. (2009), a 'failure to evoke prediction error' explanation fails to reconcile our findings with those of Monfils et al. (2009).

Another possibility is that full extinction is required for the reminder-extinction procedure to be effective in overwriting threat memories and preventing the return of threat responses. Although we aimed to instate threat memories of comparable strength to those reported by Monfils and colleagues (Monfils et al., 2009) in the group of rats conditioned with medium-intensity shocks, as well as weaker and stronger memories in the other groups, that would fully extinguish (see SI), not all our experimental groups showed full extinction. From a theoretical stance, if the reminder-extinction procedure leads to a persistent attenuation of threat through a persistent update of the threat memory to a safety memory, it is tempting to think that full extinction may be required for its efficacy. However, residual cued freezing responses at the end of extinction were also reported by Monfils and colleagues (2009). Indeed, investigated experimentally through a comparison of selectively bred rats with different extinction profiles, the effect of the reminder-extinction procedure has been shown not to depend on successful extinction learning (Auchter et al., 2017). Thus, it seems unlikely that incomplete extinction can explain the current results.

A third explanation is that our Pavlovian threat conditioning procedure created qualitatively different threat memories that are more resistant to the reminder-extinction procedure. It has previously been shown that stronger memories, created through an increased number of tone-shock pairings, are more resistant to disruption through reconsolidation-based interventions (Haubrich et al., 2020; Holehonnur et al., 2016; Wang et al., 2009). Specifically, activation of noradrenergic projections from the locus

coeruleus to the amygdala during encoding of such strong threat memories seems to limit memory lability (Haubrich et al., 2020). Although we had aimed to create threat memories in our medium intensity group of comparable strength (i.e., freezing levels) to those reported by Monfils et al. (2009), this required us to use higher shock amplitudes (see SI). The use of higher shock intensities in the current study could have engaged noradrenergic projections to the amygdala already at low shock intensity, thereby limiting the malleability of threat memories created at all three different shock intensities. Yet in our low intensity group we see relatively rapid extinction of threat responses which makes us doubt that threat memory in these animals is particularly resistant to modification due to qualitative differences.

Alternatively, the use of higher shock intensities during Pavlovian conditioning may increasingly involve not just the amygdala but also the hippocampus in the formation of the threat memory (Phillips & LeDoux, 1992). It has been suggested that hippocampus-dependent memories are less susceptible to disruption through reconsolidation-based interventions than amygdala-dependent memories (Alberini, 2005; Kroes et al., 2016; Kroes et al., 2017). Yet we note that our low intensity group received a lower shock amplitude (0.5 mA) compared to Monfils et al. (2009, 0.7 mA), and showed full extinction (unlike rats in Monfils et al., 2009), suggesting the creation of a weaker memory, yet the reminder-extinction procedure failed to prevent the return of threat responses in this group. An explanation of our result in terms of our procedure producing qualitatively different memories that were more hippocampus-dependent and therefore more resistant to modification by the reminder-extinction procedure thus seems to be limited.

A final interpretation of our results that may be most in line with the existing literature is that, contrasting previous reminder-extinction studies reporting positive findings in rodents (Auchter et al., 2017; Baker et al., 2013; Cahill et al., 2018; Clem & Huganir, 2010; Flavell et al., 2011; Jones et al., 2013; Jones & Monfils, 2016; Liu et al., 2014; Monti et al., 2017; Olshavsky et al., 2013; Pattwell et al., 2016; Piñeyro et al., 2014; Rao-Ruiz et al., 2011) and humans (Johnson & Casey, 2015; Oyarzún et al., 2012; Schiller et al., 2010; Steinfurth et al., 2014; Thompson & Lipp, 2017), the reminder-extinction procedure does not persistently disrupt threat responses (Costanzi et al., 2011; Fricchione et al., 2016; Golkar et al., 2012; Goode et al., 2017; Gräff et al., 2014; Ishii et al., 2012, 2015; Kindt & Soeter, 2013; Klucken et al., 2016; Kredlow et al., 2018; Kroes et al., 2017; Luyten & Beckers, 2017; Meir Drexler et al., 2014; Ponnusamy et al., 2016; Soeter & Kindt, 2011; Zimmermann & Bach, 2020). There is a general trend for small and non-significant effects of the extinction-reconsolidation paradigm across animal studies (Kredlow et al., 2016), and a recent direct replication attempt of the reminder-extinction effect in cued threat conditioning in rats as reported by Monfils et al. (2009), failed to reveal any differences between standard extinction and reminder-extinction procedures (Luyten & Beckers, 2017).

Collectively these non-replication studies indicate that the efficacy of the reminder-extinction procedure is hard to reproduce and may depend on subtle experimental parameters that are yet to be understood.

One limitation to our present study is that we carried out the same reinstatement procedure for all groups using quite high intensity foot-shocks. In choosing the reinstatement procedure, we reasoned that it would be best to use the same shock intensity for all groups, to avoid the possibility that differences in the return of threat could be attributed to differences in shock intensity during reinstatement. In addition, we chose a shock intensity for reinstatement that was novel to all animals and higher than any of the shock intensities used during acquisition to assure that a potential absence of the return of threat responses could not be explained by a relatively weak reinstatement procedure. A potential risk of the strong shock used during the reinstatement procedure, is that a non-associative sensitization might have occurred. To exclude the possibility of sensitization, additional control groups receiving un- or backward-paired presentations of the CS and US could have been included (see e.g. Brooks et al., 1995). However, we observed a return of threat responses following reinstatement in all groups and a main effect of shock intensity on overall freezing responses, as well as re-extinction, reflecting the intensity of the shock during acquisition on the reinstatement test. This suggests that the original threat memory was retained. It thus appeared that our reinstatement procedure preserved the differences in the original threat memories without imposing additional variation. Furthermore, we found greater freezing to the CSs than during the pre-CS periods during the reinstatement test indicating that freezing response to the CS occurred over and beyond responses to the context alone. These findings indicate that rats recovered threat memory for the intensity of the acquisition experience and expressed cue-dependent responses over and beyond context-related freezing. Hence, the recovery of threat responses that we observed are unlikely to be explained by sensitization alone and therefore we did not perform any further control experiments for sensitization effects.

Another limitation is that because we did not use no-reinstatement control groups, we cannot exclude that the return of threat that we observe in all groups is at least in part also driven by spontaneous recovery. While the reminder-extinction procedure may be have been able to prevent the return of threat responses when driven exclusively by either spontaneous recovery or reinstatement, its effectiveness may be masked when both processes are at play. Furthermore, this also means that we cannot exclude that the relative contribution of reinstatement and spontaneous recovery to the return of threat may have differed between our different shock intensity groups. Yet, whether reinstatement alone or a combination of reinstatement and spontaneous recovery drove the return of threat responses we observed, the current study does not provide evidence that the reminder before extinction has attenuated or prevented the return of threat.

Finally, we note the unexpected difference in freezing levels at the end of extinction within the medium intensity groups, and the non-significant yet numerical difference in the high intensity groups. As we had no a priori hypotheses for these differences, we hold these to be unfortunate chance findings. Yet these differences do complicate the interpretation of our findings as it makes it difficult to differentiate whether threat responses simply returned or returned to the same degree between groups. To aid this interpretation we have provided both test results of group comparisons within the extinction task and the reinstatement test as well as results for the reinstatement index. Yet the possibility remains that the reminder-extinction procedure may be effective in attenuating the return of threat responses as opposed to preventing their return altogether. We welcome future studies to further investigate this matter.

To conclude, our results indicate that a reminder before extinction failed to prevent the return of threat responses irrespective of threat memory intensity in rats. These findings question whether threat memory intensity forms a boundary condition for the reminder-extinction procedure and add to the collection of studies that cast doubt on the replicability of the reminder-extinction effect. The reminder-extinction procedure has been proposed as a behavioral alternative to pharmacological reconsolidation-targeting interventions to treat stress- and anxiety-disorders. Considering the small effect sizes of previous studies reporting that the reminder-extinction procedure can prevent the return of threat and the growing number of non-replication studies, we suggest that it would be wise to first further test pre-clinically whether the reminder-extinction procedure really works, and under which specific conditions, before embarking on translations to clinical populations. It nevertheless remains interesting to further investigate how reconsolidation and other processes could interact with extinction learning as it may reveal new treatment strategies for patients suffering from stress- and anxiety related disorders.

Materials and Methods

Animals

Adult male Sprague-Dawley rats (250-275g, Charles River) were housed individually throughout the experiment on a 12:12h light/dark cycle (lights on at 7 a.m.). Food and water were available *ad libitum* in the home cage. All experimental procedures were approved by the Central Committee on Animal Experiments (Centrale Commissie Dierproeven, CCD, The Hague, The Netherlands).

Apparatus

All experiments were carried out in two 30.5 x 24.1 x 21 cm conditioning chambers (Med Associates, Vermont) housed individually within a sound-attenuating cubicle, equipped with a white and near infrared house light. The white house light (10 lux) was illuminated during all behavioral sessions. Conditioning chambers contained a metal grid floor connected to a scrambled shock generator (Model

EW-414S, Med Associates, Vermont) to deliver foot shocks. An infra-red light was illuminated during the presentations of the tone to facilitate scoring of freezing behavior. A video camera mounted in front of the conditioning chamber, on the inside of the doors closing the sound-attenuating cubicle, was used to record behavior.

Tone habituation

On day 1, all rats were habituated for 30 minutes to the conditioning chamber once in the morning and once in the afternoon. On day 2, all rats were habituated to the conditioned stimulus (CS+, tone) in the conditioning chamber, once in the morning and once in the afternoon. Each habitation session consisted of three 20-second presentations of the CS+ with a variable inter-trial interval (60-300s, average of 180s). The first tone was preceded by a 10-minute acclimatization period and the task ended 3 minutes after the last tone presentation. The total duration of the tone habituation phase was twenty-three minutes.

Acquisition of Pavlovian threat conditioning at distinct shock intensities

Our study design followed that of Monfils et al. (2009) as close as possible. In line with this study, all tasks were carried out in the same operant conditioning chamber. Auditory conditioning cues were played through a speaker within the operant conditioning chamber, at a frequency of 3000 Hz, and a volume of 85 dB. Conditioning trials consisted of 3 presentations of a 20 second tone, co-terminating with a 1 second foot shock. Rats were conditioned at low (0.5 mA), medium (1.0 mA) or high (2.0 mA) intensity (these intensities were chosen based on pilot data to create three behaviorally distinct extinction profiles where that of the medium group would reflect behavioral responses of Monfils et al., 2009, see supplementary information). Note that although the high intensity shock is higher than generally used in previous studies, the low and medium shock intensities fall well within the range of shock intensities of previous reminder-extinction publications (0.3-0.5 mAs as reviewed by Kredlow et al, 2016 and 1.0mA used in Haubrich et al. 2020). The inter-trial intervals were variable (range 60-300s) with an average of 180 seconds. The first tone was preceded by a 10-minute acclimatization period and the task ended 3 minutes after the last tone presentation. The total duration of the acquisition phase was twenty-three minutes.

Reminder-extinction and standard extinction

Rats in the reminder-extinction condition (Rem-Ext) were subjected to a reminder before extinction. The reminder consisted of a single unreinforced CS+ presentation of 20 seconds, preceded by a 2-minute interval and followed by a 1-minute interval. The total duration of the reminder task was three minutes and twenty seconds. Following a 10 min break, the Rem-Ext group received extinction training consisting of 17 (Rem-Ext condition) unreinforced CS presentations, whereas the standard extinction (Ext) group received 18 (Ext condition) presentations to equate the number of unreinforced CS

presentations animals were exposed to, in line with Monfils et al., 2009. In line with the previous tasks, a 60-300s variable inter-trial interval was used with an average of 180 seconds. The extinction session was preceded by a 3-minute acclimatization period and ended immediately after offset of the last trial. The total duration of the extinction session one hour and 3 minutes for the Ext condition and one hour for the Rem-Ext condition.

Reinstatement procedure

To reinstate the conditioned threat response, rats were placed in the conditioning chamber and received 4 unsignalled USs at a novel shock intensity that was kept constant for all experimental groups (2.2 mA). To avoid that the potential differential return of threat responses could be attributable to differences in intensity of the shock during reinstatement, we opted to keep the intensity of the reinstatement shocks constant for all experimental groups. To avoid that the intensity of the shock was familiar to some animals but not others and we opted for a novel shock intensity for all animals. Here, we chose for a higher, as compared to a lower, shock intensity than any of the animals had experienced to prevent that potential absence of the return of threat responses could be explained by a weak reinstatement procedure. Reinstatement trials consisted of four 1-second, unsignalled foot shocks. The inter-trial intervals were variable (range 60-300s) with an average of 180 seconds. The reinstatement procedure started with a 10-minute acclimatization period, and continued for three minutes after offset of the last trial. The total duration of the reinstatement procedure was 23 minutes.

Reinstatement test

To test the extent of reinstatement of conditioned threat responses after the reinstatement procedure, rats were placed in the conditioning chamber for a reinstatement test consisting of 3 unreinforced CS+ presentations. We chose for reinstatement as a test for threat recovery as Monfils et al., 2009 reported that the reminder-extinction procedure prevented reinstatement on a test after 24h (compared to spontaneous recovery after one month), allowing us to limit the number of days required for our study. The inter-trial intervals were variable (range 60-300s) with an average of 180 seconds. In line with the reinstatement procedure, a 10-minute acclimatization period preceded the first trial, and the task continued for three minutes after offset of the last trial. The total duration of the reinstatement test was 23 minutes.

Assessment of freezing

The time spent immobile, with the exception of breathing and ‘scanning’ behavior, during the CS+ presentation was scored as “freezing”. The duration of freezing was measured with a digital stopwatch by an observer blind to the experimental conditions (M.C.H). To assess the reliability of scoring, all behavior was scored by two additional blinded observers (K.P.B., M.C.W.K.), each scoring a different subset of animals. Freezing scores were averaged per session separately for each observer and showed

a strong average correlation (*r*= 0.77 for observer 1&2 and *r*=0.91 for observer 1&3*)* across observers. Freezing scores by observer 1 thereby deemed reliable and were used for further analysis.

Procedure

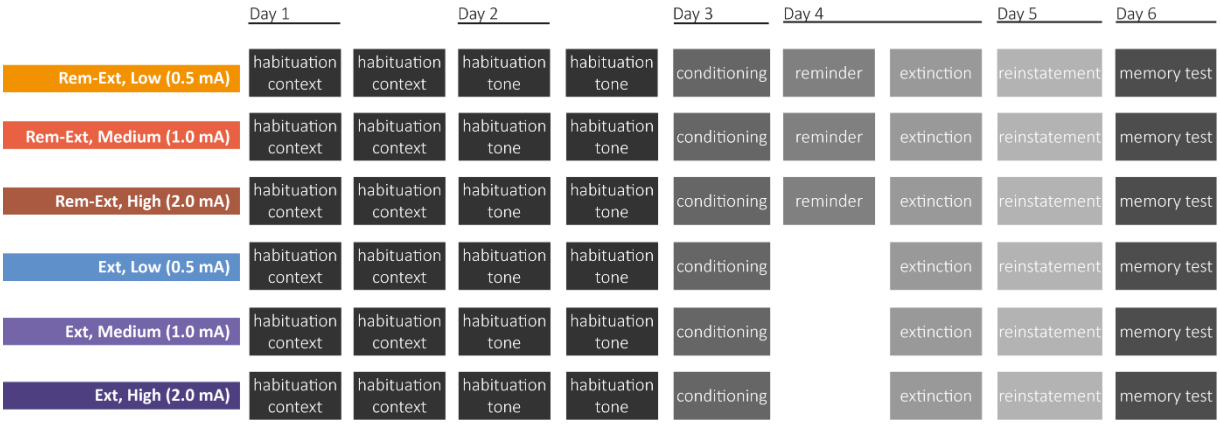


Figure 3.3. Overview of the experimental design. A between-subjects design with extinction either preceded by a reminder (Rem-Ext) or without a reminder (Ext) and three different levels of shock intensity during acquisition (low, medium and high). All groups underwent habituation to the context (day 1) and tone (CS+) (day 2) followed by auditory cued threat conditioning. On day 4, the Rem-Ext group received a reminder, while the Ext group did not. After a 10-minute break, all groups underwent extinction training. On day 5, four unannounced presentations of the foot shock were administered to reinstate the threat memory. On day 6, long-term memory for the CS+ induced threat response was tested.

In total, 78 rats were exposed to behavioral testing in a six-group between-subjects design taking place across 6 days (see Figure 3.3). Rats were assigned to either the reminder-extinction (Rem-Ext) or extinction (Ext) condition with a low, medium or high shock intensity (n=13 per group). The study was counterbalanced so that an equal number of rats was assigned to each group in each testing batch. Upon arrival, rats were first habituated for a week to individual housing conditions, followed by a week of habituation to human handling (once a day, 5-10 minutes). Rats were then pseudo-randomly assigned to experimental groups according to a predetermined allocation sequence.

On day 1, all rats were habituated for 30 minutes to the conditioning chamber once in the morning and once in the afternoon. On day 2, all rats were habituated to the conditioned stimulus (CS+, tone) in the conditioning chamber, once in the morning and once in the afternoon. Each habitation session consisted of three presentations of the CS+. On day 3, all rats were conditioned by 3 presentations of the CS+ that co-terminated with the unconditioned stimulus (US, electrical foot-shock stimulation). Rats were pseudo-randomly assigned to different groups and received low, medium or high electrical stimulation to acquire threat memories of different intensity. On day 4, rats assigned to the Rem-Ext condition were reminded of the CS+ by a single unreinforced presentation, while rats assigned to the Ext condition were not. After 10-minute interval, for which rats in the Rem-Ext group were returned to the home-cage, all rats underwent extinction training. All rats received reinstatement on day 5 and a reinstatement test on day 6 to test for the possible return of threat responses.

Statistics

All statistical analyses were performed using SPSS Statistics version 23.0 (SPSS Inc., IBM, Armonk, NY, USA). Data are presented as mean ± standard error of the mean (SEM). Scores for critical tests (mean level during the acquisition and extinction phases, and the reinstatement index) were checked for outliers to ensure no individual values were more than 3 standard deviations away from the mean. This criterion did not lead to the exclusion of any subjects. Behavioral data were analyzed using a repeated measures analysis of variance (ANOVA) with trial as repeated measure and reminder (Rem-Ext, Ext) and shock intensity (low, medium, high) as between-subject factors. Significant ANOVA contrasts and interaction effects were followed up by one-way ANOVAs and independent sample t-tests. For all tests a p-value of 0.05 as cut-off value for significance was considered. Statistics were Greenhouse-Geisser or Huyn-Feldt corrected for non-sphericity when appropriate. The number of animals included in our study had been chosen based on power-analyses as reported in our pre-registration. The proper comprehension of our results required us to run statistical tests additional to those described in our preregistration. We will refer to these additional statistical tests as 'exploratory' and for the sake of readability occasionally present these before the results of our preregistered analyses.

Supplementary Information – Chapter 3

Pilot Experiment

We aimed to create a conditioned threat memory at comparable strength as the memories associated in the original paper by Monfils and colleagues[7], as well as a weaker and a stronger memory. To achieve this, we performed a pilot experiment comparing four different shock intensities (0.3, 0.5, 1.2 and 2.0 mA, see Figure 3.4). Rats conditioned at the lowest threat intensity (0.3 mA) showed negligible levels of freezing during the first trial of the reinstatement test (9.2%, n=6), while rats conditioned at medium intensity (0.5 mA) showed modest levels of freezing (27.0%, n=6). Rats conditioned at 1.2 mA showed 50.8% freezing on the first trial of extinction (n=6), and rats conditioned at 2.0 mA showed 66.3% of freezing (n=4). As freezing levels at the 1.2 mA and 2.0 mA seemed more similar than freezing levels at 1.2 mA and 0.5 mA, we decided to use 0.5 mA, 1.0 mA and 2.0 mA to create maximally distinguishable levels of conditioned threat.

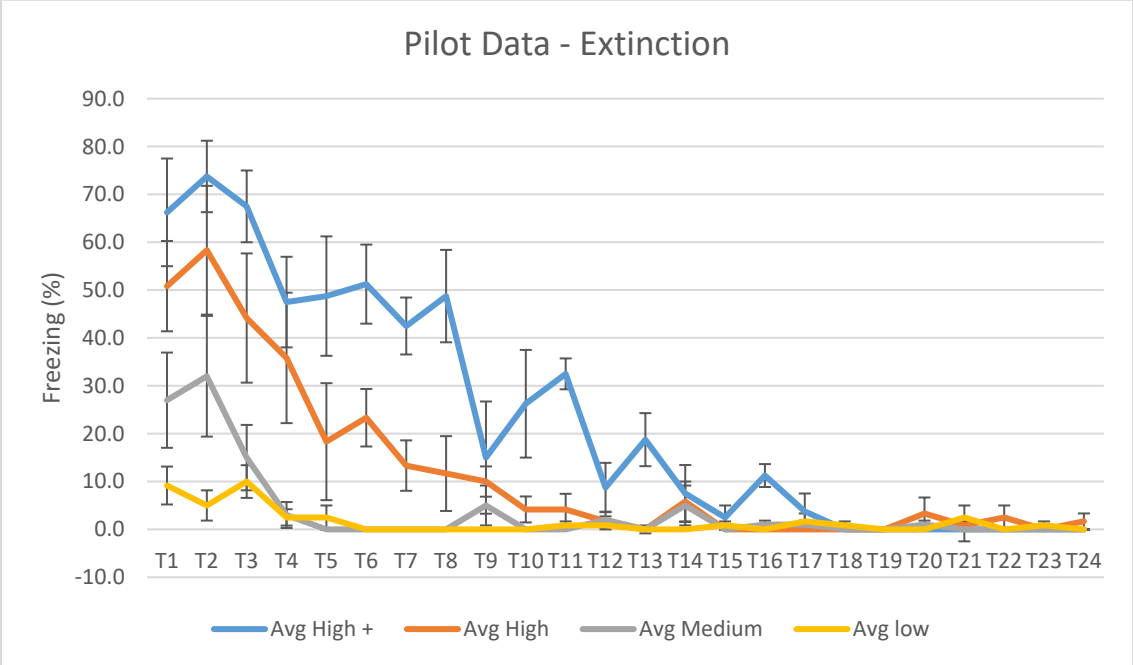


Figure 3.4 Freezing levels evoked by four different stimulus intensities in a pilot experiment. Rats in all groups (4, groups, n=4-6 per group) were exposed to three tone-shock pairings. On the subsequent day, all animals underwent 24 extinction trials. Groups conditioned at different shock intensities are displayed as separate lines. Data presented as mean.

Tone Habituation

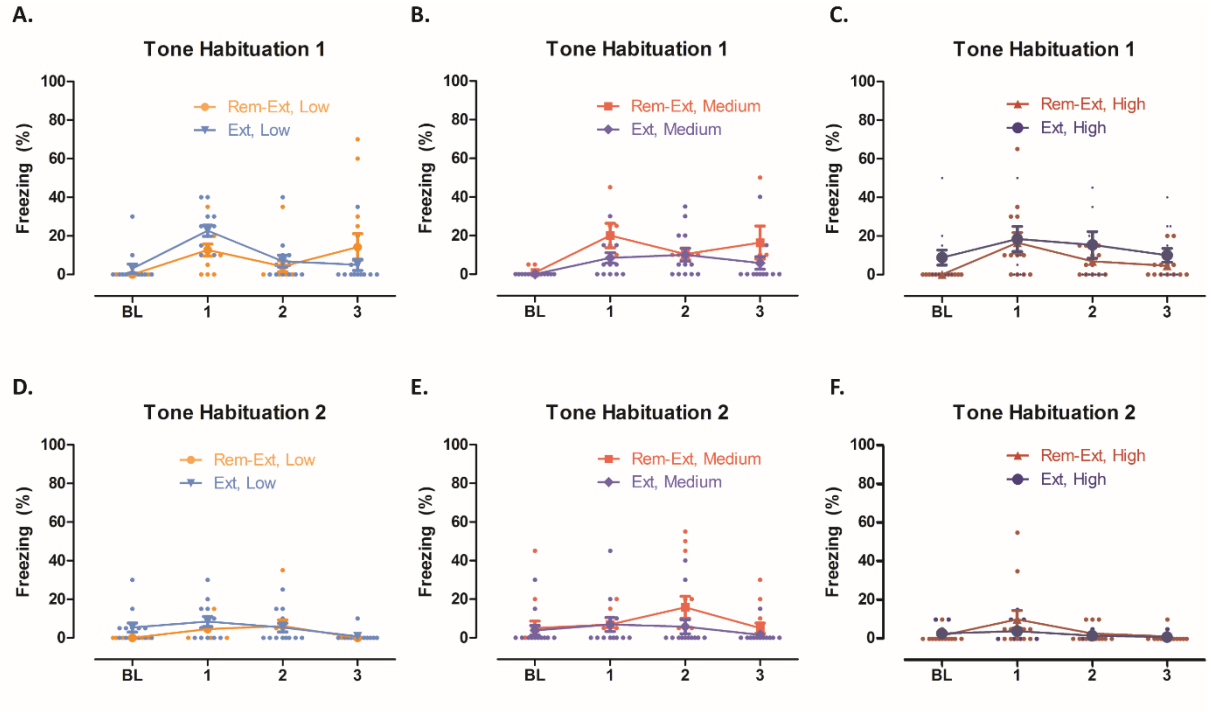


Figure 3.5. All groups show comparable habituation to the auditory tone later used as conditioned stimulus. Rats in all experimental groups (6 groups, n = 13 per group) were exposed to 3 presentations of the auditory cue (CS+) in a separate first (**A-C**) and second (**D-F**) habituation session. Freezing scores to the tone decreased during habituation sessions. All groups are shown as separate lines. Data presented as mean ± S.E.M. Dots represent individual data points.

During the first habituation session, rats show marginally increased freezing during auditory cue (CS+) presentation as compared to freezing at baseline level, measured during the 20 seconds preceding the onset of the first tone (See Figure 3.5). Freezing to the auditory cue (CS+) decreased over trials ($F_{(2,140)}$=6.939, p=0.001, η2 = 0.090). Specifically, freezing levels decreased from the first to the second trial (t(75)=3.711, p<0.001, from 16.4% ± 1.9% freezing to 9.0% ± 1.6%), but did not change significantly from the second to the third trial (t(75)=-0.88, p=0.930, 9.0% ± 1.6% freezing on the second trial and 9.15% ± 2.0% on the third trial). Freezing levels during the second habituation session also decreased over trials ($F_{(1.728,124.449)}$=8.424, p=0.001, η2 = 0.105). During the second habituation session freezing levels did not change from the first to the second trial (t(77)=0.327, p=0.744, from 6.65% ± 1.2% freezing to 6.10% ± 1.4%) but showed a significant drop from the second to the third trial (t(77)=3.884, p<0.000, 6.10% ± 1.4% freezing on the second trial and 1.4% ± 0.5% freezing on the third trial). To explore whether tone-induced freezing levels were successfully reduced to baseline levels, a paired t-test was used to compare baseline freezing levels before presentation of the first tone in the first habituation session to freezing levels during the last tone presentation in the second session. Freezing was similar during the last tone presentation as compared to baseline freezing (p=0.541). Thus, tone habituation was successful in reducing tone-induced freezing to low levels, and this was not different between groups.

Acquired threat responses in individual animals

We explored whether all animals display acquisition of threat responses on an individual level by ensuring that animals displayed freezing behaviour either during the acquisition phase, during presentation of the reminder, or during the first half of the extinction phase. All animals displayed freezing behaviour during the conditioning phase. For the sub-set of ten animals that displayed an average freezing level below 20% during the three acquisition trials, we explored freezing behaviour during the first three reminder-extinction trials. On average, these animals showed 56.3% freezing during the first three reminder-extinction trials. Only one of the animals showing freezing levels below

20% during the acquisition phase also showed freezing levels below 20% during the first three reminder-extinction trials, showing 8.3% freezing during the acquisition phase and 7.5% freezing during the early reminder-extinction trials. Hence, we conclude that all individual animals show acceptable acquisition of conditioned threat responses.

To investigate context-elicited freezing as opposed to CS-elicited freezing, freezing levels were scored during a 20-second window prior to the onset of the first CS. We found that during the reminder phase, pre-CS freezing levels were rather substantial (pre-CS: 53.6%±5.8%), but significantly lower than CS-evoked freezing levels ($t(38)=-4.674$, $p<0.001$, CS-evoked:68.7%±5.5%). Pre-CS (i.e., contextual) freezing during the extinction phase is similarly lower than CS-evoked freezing levels ($t(75)=-5.636$, $p<0.001$, pre-CS: 50.1%±4.5%, CS-evoked: 67.3%±3.8%). Pre-CS levels of freezing during the reinstatement test however did not significantly differ from freezing levels during the first CS-presentation ($Z=-1.172$, $p=0.241$, pre-CS: 69.5%±4.0%, CS-evoked: 72.7%±3.1%). We explored freezing during the reinstatement test in more detail by also measuring freezing levels in the 20s window prior to each stimulus onset. This analysis indicated that average pre-CS freezing levels were significantly lower than CS-evoked freezing ($t(77)=-4.232$, $p<0.001$, pre-CS: 58.4%±2.5%, CS-evoked: 66.05%±2.4%). A trial (1-4) x shock intensity (low, med, high) x group (Rem-Ext, Ext) revealed a decrease in pre-CS freezing over trials ($F_{(3,204)}=6.970$, $p<0.001$, $\eta 2 =0.093$) and no other effects or interactions (all p's>0.06).

Chapter 4. Counterconditioning in humans: Unravelling the neurocognitive mechanisms underlying counterconditioning in humans

Maxime C. Houtekamer, Lisa Wirz, Jette de Vos, Joseph E. Dunsmoor, Judith Homberg, Marloes J.A.G. Henckens, Erno J. Hermans

Abstract

Counterconditioning aims to attenuate emotional memories by establishing a new association of opposite valence. Aversive-to-appetitive counterconditioning (CC) holds promise for improved treatment of stress-related disorders. While the neurocognitive mechanisms underlying CC are largely unexplored, previous studies suggest qualitatively different mechanisms from extinction.

In this functional magnetic resonance imaging (fMRI) study, we compared the neural mechanisms underlying CC and extinction between subjects. To test whether CC and extinction result have different efficacies, we also measured physiological threat responses, valence and arousal ratings and recognition memory.

Participants underwent differential categorical threat conditioning. In a hybrid version of the monetary incentive delay task, participants responded to targets superimposed on category exemplars. During the CC task, conditioned exemplars were reinforced with monetary rewards, while reinforcement was omitted in the extinction task.

The next day, recovery of differential conditioned threat responses was assessed. As expected, we observed spontaneous recovery after extinction but not CC, suggesting enhanced efficacy of CC. Interestingly, CC not only strengthened recognition memory for conditioned exemplars presented during CC but also retroactively strengthened recognition memory for the prior fear conditioning task. While the ventromedial prefrontal cortex (vmPFC) was activated during regular extinction, participants undergoing CC showed persistent CS+-specific deactivations in the vmPFC and hippocampus and CS+-specific activation of the nucleus accumbens (NAcc). These data suggest that CC leads to more efficient safety learning with a distinct neural substrate: increased reward-processing in the NAcc along with a deactivation of the vmPFC that is generally involved in extinction, resulting in enhanced retention.

Introduction

Trauma-related disorders such as posttraumatic stress disorder are prevalent and highly detrimental to the individual's quality of life (Kessler et al., 2005). To treat these disorders, patients are undergo exposure therapy in a safe therapeutic environment to allow threat responses to fade away (Scheveneels et al., 2016). Although exposure therapy may be successful initially, relapse often occurs and is the major remaining challenge in optimizing treatment efficacy. Research suggests that exposure therapy creates a safety memory that competes for expression with the original threat memory (Bouton, 2004; Myers & Davis, 2002), suggesting that relapse may occur because of relatively weak learning and retention of the safety memory. Therefore, identifying mechanisms that can be used to strengthen safety learning is a key step in advancing treatment for trauma-related disorders. A promising approach to strengthen safety learning is to create new, positive associations with experiences that were previously linked to aversive experiences. However, while there are indications that establishing positive associations can prevent relapse, the underlying mechanisms are poorly understood (for a review, see Keller et al., 2020).

To study threat responses in a controlled setting, experiments typically use aversive Pavlovian conditioning, in which a neutral stimulus (conditioned stimulus, CS; e.g., a picture) is coupled with a biologically aversive unconditioned stimulus (US; e.g., an electrical shock), after which the CS alone also elicits a conditioned threat response. Conditioned threat responses to the CS can be attenuated using extinction, during which the CS is repeatedly presented in absence of the US. However, early theories have suggested that the aversive responses may more easily be inhibited by engaging appetitive systems (Dickinson & Pearce, 1977; Rescorla & Solomon, 1967). Indeed, experiments have shown that coupling a CS to a positive US after threat conditioning, a process known as aversive-to-appetitive counterconditioning (CC), speeds up the attenuation of the conditioned response (Dickinson & Pearce, 1977; Pearce & Dickinson, 1975), reduces threat expectancy (Kang et al., 2018; Newall et al., 2017) and leads to more positive valence ratings (Jozefowiez et al., 2020; Luck & Lipp, 2018; van Dis et al., 2019) immediately post-CC. Tests for spontaneous recovery, reinstatement, and renewal can subsequently be used to evaluate the return of threat over time, after unsignalled presentation of the US, or in a novel context, respectively (Bouton, 2002, 2004), to investigate whether CC persistently attenuates threat responses. While some studies found comparable spontaneous recovery (van Dis et al., 2019) and reinstatement (Luck & Lipp, 2018; van Dis et al., 2019) of threat responses after CC compared to standard extinction, another study indicated that CC can attenuate spontaneous recovery of threat responses (Keller & Dunsmoor, 2020). The difference in observed efficacy of CC may depend on the positive valence evoked during the CC phase, and could be lower in the former studies due to the use of auditory startle probes during the CC (de Haan et al., 2018; van Dis et al., 2019) and the use

of positive verbal instructions (Luck & Lipp, 2018) as opposed to a conditioning procedure with positive stimuli (Keller & Dunsmoor, 2020). Whether extinction and CC engage distinct neural mechanisms is largely unexplored.

Extinction learning appears to be mediated by activation of the vmPFC, which in turn inhibits the expression of threat responses by suppressing amygdala activity (Morgan et al., 1993; Phelps et al., 2004; Quirk et al., 2000; Quirk et al., 2003). When the process of extinction is enhanced by replacing aversive with novel, neutral outcomes, the ventromedial prefrontal cortex (vmPFC) was found to be engaged more effectively than during standard extinction (Dunsmoor, Kroes, Li, et al., 2019). Interestingly, both standard and enhanced extinction have similar effects on episodic memory. Both result in a drop in recognition of CS+ items presented during extinction compared to CS+ items presented during the acquisition phase (Dunsmoor et al., 2018). If CC is another form of enhanced extinction, it may likewise be mediated by a stronger engagement of extinction networks compared to regular extinction, while resulting in comparable drops in episodic memory.

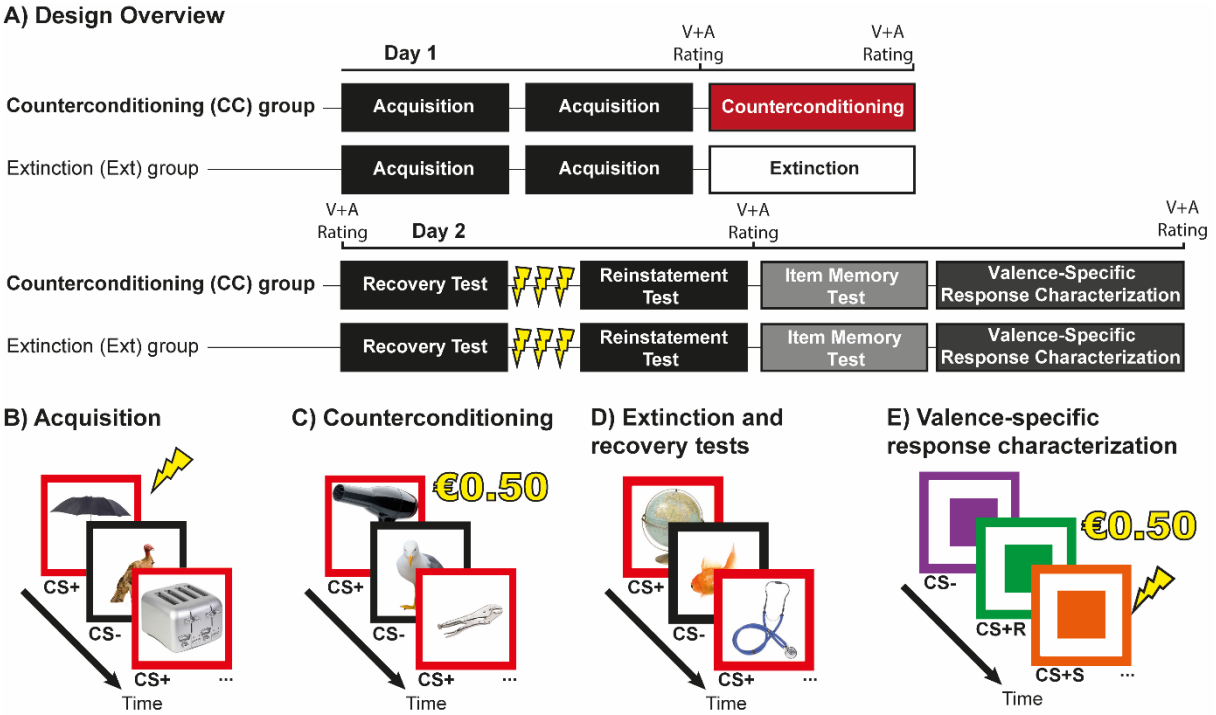
Recent work by Keller and Dunsmoor (2020), however indicates that CC and extinction have opposite effects on episodic memory. Item recognition was strengthened for CS+ items from the counter-conditioned category compared to the extinguished category, suggesting that compared to extinction CC can enhance episodic memory representations and potentially provide stronger retrieval competition against a threat memory. At a neural level, counterconditioning has been associated with activation of the ventral striatum (Bulganin et al., 2014; Correia et al., 2016), a region known to be involved in the anticipation and receival of reward (Diekhof et al., 2012). Taken together, CC may lead to the formation and consolidation of a positive memory that provides stronger competition against retrieval of the threat memory compared to regular extinction, which could be mediated by activation of reward-related neural circuits.

We expected that CC leads to a more persistent attenuation of threat responses compared to extinction. This could be mediated by two possible neural mechanisms; through enhanced engagement of extinction circuitry, reflected by increased engagement of the vmPFC, or through a shift towards reward networks, reflected by activation of the ventral striatal activation. Based on earlier studies, we expect engagement of reward networks to strengthen item recognition, while enhanced engagement of extinction does not. To investigate this, we performed a two-day fMRI study comparing CC versus regular extinction in a between-subjects design. Participants underwent differential threat conditioning to a semantic category (animals or objects). Subsequently, participants underwent aversive-to-appetitive CC with monetary reinforcement for the CS+ (CC group) or regular extinction (Ext group). We tested threat memory retrieval the next day. We measured skin

conductance responses (SCRs) and pupil dilation responses (PDRs) as indicators of physiological arousal evoked by threat- and reward-anticipation. To distinguish between arousal responses evoked by threat and reward, we collected explicit valence ratings. Item recognition memory for pictures presented during the acquisition and CC/extinction tasks was assessed in a surprise memory test 1 day later.

Results

To investigate whether CC prevents the return of threat responses compared to regular extinction, we used a 2-day between-subjects design (**Figure 4.1A-F**). During the CC task, participants in the CC group were able to obtain monetary rewards dependent on how quickly they responded to a cue superimposed on category exemplars from the CS+ category, similar to the monetary incentive delay (MID) task (Knutson et al., 2000). To maximize task similarity between tasks and groups, the cued-response element was kept consistent between tasks from acquisition to reinstatement, although response-time contingent monetary rewards were only presented to the CC group during the CC task (**Figure 4.1F**). The presentation of shocks during the acquisition task was not contingent on response times. On day 2, the return of threat responses was assessed using a test of spontaneous recovery followed by a reinstatement procedure and test. To characterize PDRs and SCRs during the anticipation of shock- and reward-reinforcement independently from prior conditioning, a separate valence-specific response characterization task was included at the end of the experiment (**Figure 4.1E**). In the valence-specific response characterization task, we observed that both threat and reward-anticipation induce strong arousal-related PDRs and SCRs (see Supplementary Information). However, PDRs allowed for a better differentiation between the two. Therefore, we decided to focus on PDRs and will refer to the supplementary information for details on the analysis of SCRs. During the acquisition task, both groups showed comparable and successful acquisition of differential conditioned threat responses (see Supplementary Information).



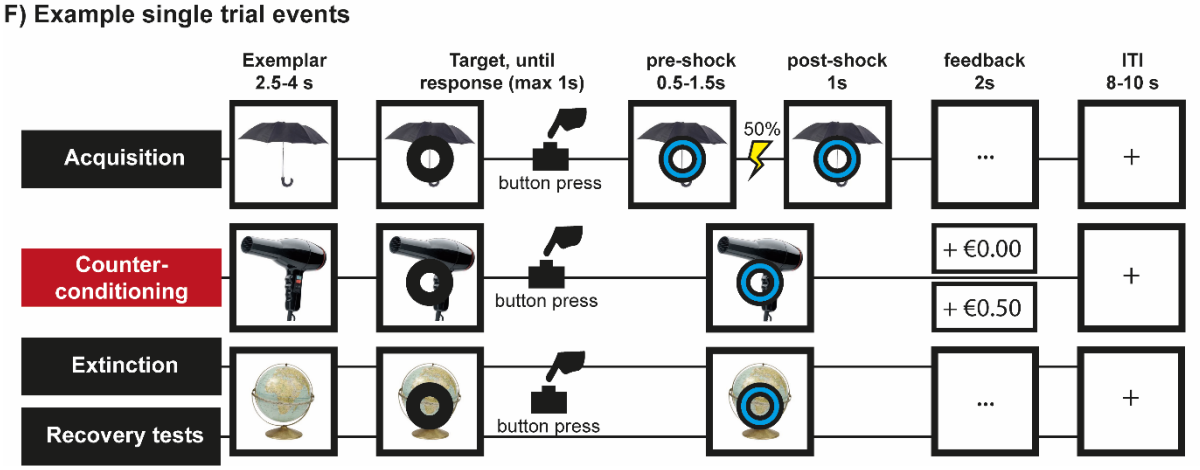


Figure 4.1. Overview of the experimental design. (A). Participants were assigned to the counterconditioning (CC) or extinction (Ext) group. On day 1, participants performed a category localizer paradigm (not described in this manuscript), two blocks of acquisition of category-conditioned threat responses separated by a 30 second break and either a CC or extinction task, depending on the assigned group. On day 2, participants performed a spontaneous recovery test, followed by a reinstatement procedure and test, an item memory test and a valence-specific response characterization task. Valence and arousal ratings for the different categories (stimuli in case of the valence-specific response characterization task) were taken before or after the tasks as indicated by 'V+A Rating'. All tasks were performed in an MRI scanner. **(B)** During the acquisition task, participants viewed trial-unique exemplars of objects and animals. Exemplars of one category (CS+ animals or objects counterbalanced) were paired with a shock in 50% of trials, while CS- trials were not reinforced. **(C)** Participants in the CC group were able to earn a monetary reward if they responded quickly enough to exemplars in the CS+ category. **(D)** Participants in the Ext group underwent extinction. During the extinction task, recovery test and reinstatement test, neither CS+ nor CS- exemplars were paired with a shock. **(E)** In the valence-specific response characterization task, participants viewed three different coloured squares. One colour was associated with shock (CS+S), one colour with reward (CS+R) and one colour served as CS-. The trial structure was otherwise identical to the acquisition and CC tasks. **(F)** In all Pavlovian tasks, trial onset was marked by presentation of a unique category exemplar. After a variable interval, a target appeared, to which participants were instructed to respond as quickly as possible. Upon response, the cue shifted in colour to confirm that the response had been registered. In the acquisition task, participants could receive a shock after a variable interval of 0.5-1.5-seconds after the response window had elapsed (indicated as 'pre-shock). The category exemplar and cue remained on screen 1 second after potential shock administration (indicated as 'post-shock'). During the CC tasks, participants receive visual feedback for 2 seconds on the monetary rewards (+€0.50 approximately the fastest 70% of trials, +0.00 on other trials),

whereas during the other tasks, participants view neutral feedback consisting of three dots. Trials are separated by an 8-10 s intertrial interval, during which a fixation cross is displayed in the centre of the screen.

Behavioural and physiological findings

Extinction and appetitive counterconditioning

After fear acquisition, participants in the CC group underwent appetitive CC, while participants in the Ext group underwent extinction. Across both groups and phases (early vs. late), we observed retention of conditioned differential PDRs (rmANOVA, CS-type (CS+, CS-) x Phase (Early, Late) x Group (CC, Ext), main effect CS-type: $F_{(1,34)}=15.393$, $p<0.001$, $\eta^2=0.312$, **Figure 4.2**), as well as a decrease in PDRs over the course of the task (main effect phase: $F_{(1,34)}=10.121$, $p=0.003$, $\eta^2=0.229$). These findings were in contrast to our expectation of a CS-type x Phase x Group interaction, with differential PDRs extinguishing during extinction in the Ext group, while being sustained in the CC group due to increased reward anticipation. Extinction in the Ext group however already occurred during the early phase (paired t-test, early CS+ vs. CS-, $p=0.233$), and differential responses did not change towards the late phase ($p=0.979$). As a result, we found distinct differential conditioned PDRs throughout the CC/extinction task between groups (CS-type x Group interaction: $F_{(1,34)}=6.053$, $p=0.019$, $\eta^2=0.151$), with participants undergoing CC showing retention of differential conditioned responses (paired t-test average CS+ vs. CS-, $t(20)=3.602$, $p=0.002$, CS+: 1.07±0.04, CS-: 1.04±0.04), whereas differential PDRs were extinguished in participants undergoing extinction (paired t-test average CS+ vs. CS-, $p=0.246$, CS+: 1.05±0.04, CS-: 1.04±0.04). The valence-specific response characterization showed that

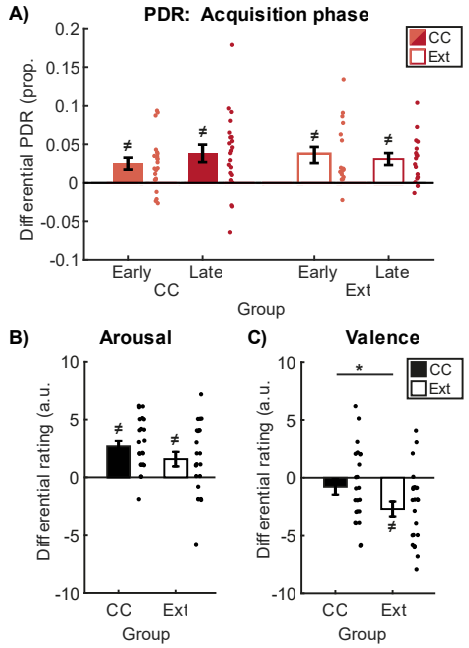


Figure 4.2. Differential PDRs during CC/extinction and explicit ratings of arousal and valence provided after the counterconditioning or extinction phase. (A) Differential PDRs for the early (light red) and late (dark red) phase of counterconditioning (CC, solid bars) or extinction (EXT, open bars). Participants undergoing CC showed increased differential PDRs as compared to participants undergoing extinction. (B) Arousal and (C) valence ratings displayed separately for participants assigned to the counterconditioning (CC, solid bars) and extinction (EXT, open bars) groups. Participants that had undergone CC gave stronger differential arousal scorings than participants that had undergone extinction. In addition, participants that underwent CC showed flipped differential valence ratings: while valence differential valence ratings were negative after extinction, the direction reversed to positive differential ratings after CC. Error bars represent ± standard error of the mean. *=p<0.05, **=p<0.01, ***=p<0.001, ≠ indicates that the bar is significantly different from 0.

differential PDRs can also be indicative of anticipation of reward (**Supplementary Figure 4.10A**). Thus, while PDRs in the Ext group indicate that differential conditioned threat responses were successfully extinguished, differential PDRs persist in the CC group, likely reflecting reward anticipation. Differential

SCRs persisted during the late phase of both CC and extinction but were no longer detectable in the last two trials and were comparable between groups (see supplementary information).

Valence and arousal ratings provide further support for extinction of differential responses in the Ext group and positive, reward-induced arousal for CS+ items in the CC group (**Figure 4.2B-C**). Differential valence ratings for the CS+ and CS- differed between groups after the CC/extinction task (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), CS-type x Group interaction: $F_{(1,44)}=12.054$, p=0.001, $\eta^2=0.215$). Participants in the CC group rated CS+ stimuli more positive than CS- stimuli (t(21)=3.469, p=0.002, CS+: 7.5±0.30, CS-: 5.41±0.38), while participants in the Ext group gave both categories similar valence ratings (p=0.245, CS+: 5.63±0.32, CS-: 6.21±0.28). Differential arousal ratings for the CS+ and CS- also differed between groups (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), CS-type x group interaction: ($F_{(1,44)}=20.862$, p<0.001, $\eta^2=0.322$). Participants in the CC group reported higher arousal levels for the CS+ category than for the CS- category (t(21)=6.370, p<0.001, CS+: 6.64±0.20, CS-: 3.45±0.38) while participants in the Ext group gave similar arousal ratings for the CS+ and CS- categories (p=0.290, CS+: 4.21±0.43, CS-: 3.80±0.40). Taken together, more positive valence and higher arousal ratings for the CS+ in the CC group as compared to the Ext group further support the interpretation of increased differential PDRs reflecting arousal induced by reward anticipation.

CC prevents differential spontaneous recovery

To investigate whether CC prevented the spontaneous recovery of differential conditioned threat responses, we compared PDRs in the last two trials of the CC/extinction phase and the first two trials of the spontaneous recovery test in a CS-type (CS+, CS-) x Group (CC, Ext) x Phase (last two trials of CC/extinction, first two trials of the spontaneous recovery test) rmANOVA. We expected the Ext group to show an increase in PDRs from the extinction task to the spontaneous recovery task, while we expected PDRs for the CC group to remain stable or decrease. Critically, differential spontaneous recovery of PDRs differed between groups (Group x CS-type x Phase interaction: $F_{(1,28)}=6.329$, p<0.018, $\eta^2=0.184$, **Figure 4.3**). While the CC group showed a decrease in differential PDRs from CC to spontaneous recovery (t(14)=-1.807, p=0.046, one-tailed, CC: 0.34±0.2, spon: -0.01±0.18), the Ext group showed an increase in differential PDRs (t(14)=1.850, p=0.043, one-tailed significance, extinction: 0.11±0.01, spon: 0.04±0.02). To conclude, while we observed differential spontaneous recovery in the Ext group, we did not find evidence for differential spontaneous recovery in the CC group, suggesting that CC attenuated the recovery of threat-responses compared to regular extinction.

However, since participants undergoing CC showed persistent differential PDRs during the last two trials of the CC phase, while participants undergoing extinction did not, we additionally explored whether there was differential responding during the first two trials of the spontaneous recovery test. During the first two trials of the spontaneous recovery test, participants in the CC group showed

decreased differential PDRs as compared to the Ext group (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), CS-type x Group interaction: $F_{(1,29)}=3.901$, p=0.029, one-tailed, $\eta^2=0.119$). Further exploration within the groups confirmed that participants in the CC group did not show retention of differential responses (paired t-test, CS+ and CS- responses during the first two trials of the spontaneous recovery test, p=0.219, one-tailed), while the Ext group did show increased responses to the CS+ as compared to the CS- (t(14)=1.958, p=0.35, one-tailed). Thus, both the differential spontaneous recovery of PDRs between sessions, and differential responding within the first two trials of the spontaneous recovery test suggested that CC prevented spontaneous recovery of differential responses compared to extinction. SCRs did not show differential recovery and were comparable between groups (see supplementary information).

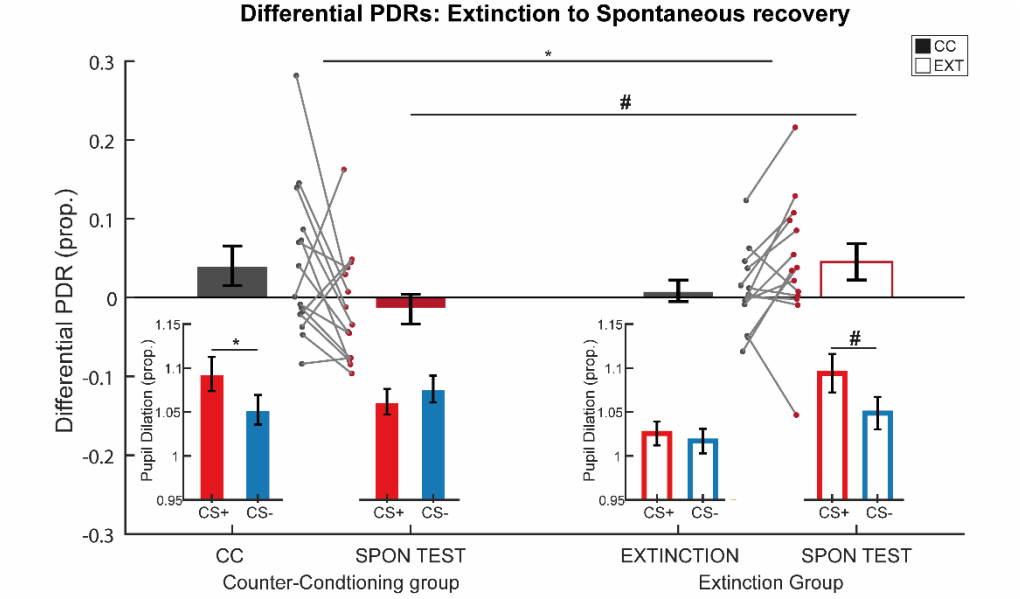


Figure 4.3. Differential PDRs during last two trials of extinction (grey) and the first two trials of the spontaneous recovery test (dark red). Differential PDRs show selective spontaneous recovery after extinction (Ext group, open bars) but not after CC (CC group, solid bars). During the first two trials of the spontaneous recovery test, differential PDRs are increased in the Ext group as compared to the CC group. Insets show PDRs to the CS+ (red) and CS- (blue) during the last two trials of extinction and the first two trials of the spontaneous recovery test. While the Ext group shows differential responding during the spontaneous recovery test, the CC group does not. Error bars represent ± standard error of the mean. *=p<0.05, #=p<0.05 one-tailed significance.

CC also appeared to have lasting beneficial effects on valence ratings compared to extinction. At the start of the second testing day, differential valence ratings continued to differ between groups (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), CS-type x Group interaction: $F_{(1,44)}=5.160$, p=0.028, $\eta^2=0.105$). While participants in the CC group gave similar valence ratings to both categories (p=0.179, CS+: 6.3±0.34, CS-: 5.4±0.35), participants in the Ext group gave more negative valence ratings to the CS+ category than to the CS- category (t(23)=-1.964, p=0.031 one-tailed test, CS+: 5.5±0.30, CS-: 6.3±0.24), also illustrative of relapse of threat associations.

Surprisingly, while participants in the CC group showed heightened differential arousal ratings immediately after CC as compared to ratings from participants who had undergone extinction (**Figure 4.2B**), participants in both groups gave comparable differential arousal ratings at the start of the second day immediately before the spontaneous recovery test (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), main effect of CS-type: $F_{(1,44)}$=10.932, p=0.002, η^2=0.022, CS+: 4.8±0.28, CS-: 3.9±0.24). Likewise, response times to the CS+ and CS- during the first two trials of the spontaneous recovery task were similar across both groups (all p's>0.2). These findings may suggest that differential arousal evoked by the categories was similar in both groups immediately before and during the spontaneous recovery test.

The spontaneous recovery test was followed by a reinstatement procedure, consisting of three unsignalled shocks, and a reinstatement test. However, mean PDRs decreased from spontaneous recovery to reinstatement (t(30)=3.063, p=0.005, last two trials of spontaneous recovery: 1.04±0.01, first two trials of reinstatement: 1.01±0.01). Given that we did not observe successful reinstatement in either group, our reinstatement test was not informative on whether CC can lead to a more persistent attenuation of fear as compared to regular extinction. A full description of PDR and SCR results of the reinstatement test can be found in the supplementary information.

Neuroimaging findings

Distinct stimulus type-specific activation for extinction and appetitive counterconditioning

After acquisition, the CC group underwent appetitive CC, while the Ext group underwent regular extinction. Whole brain analysis revealed that over the course of this task, stimulus-specific activation changed differentially between the two groups in a large cluster encompassing multiple regions in the medial temporal lobe (Group x CS-type x Phase interaction, cluster size = 1760 mm^3, p=0.034, whole-brain FWE-corrected, **Figure 4.5B** and **Table 4.1**). We further investigated the anatomical location of the cluster using our ROIs to probe for activity and found that the effect encompassed the amgydala. To further investigate the interaction effect in the amygdala, we extracted parameter estimates from the complete bilateral amygdalae (Automated Anatomic Labelling, AAL, atlas in the WFU PickAtlas toolbox in MN152 space) and performed post-hoc comparisons. In the early phase, stimulus type-specific responses differed between the groups (t(1,44)=2.173, p=0.035, CC: 0.18±0.08, Ext: -0.073±0.08). Specifically, the CC group showed increased amygdala activation to the CS+ as compared to the CS- (t(23)=2.210, p=0.037) while that was not the case in the Ext group (p=0.390). In the late phase, differential responses were comparable between groups (p=0.503).

Whole-brain analysis further revealed a number of clusters showing distinct CS-specific activations between groups throughout the task, including the anterior cingulate, cuneus, nucleus accumbens,

caudate, thalamus and inferior frontal gyrus (**Figure 4.5A**, **Table 4.1**). The group and stimulus-specific activation of the NAcc was in line with a priori expectations for the CC phase (**Figure 4.5C**). To further explore this effect, averaged parameter estimates from the bilateral NAcc ROI (mask acquired from the IBASPM 71 atlas in the WFU PickAtlas toolbox in MNI152 space) were extracted. Across the bilateral NAcc, differential activation was increased in the CC as compared to the Ext group ($t(44)=2.731$, $p=0.009$, CC: 0.37±0.10, Ext: 0.04±0.06), with the CC showing increased NAcc activation to the CS+ compared to the CS- ($t(23)=6.194$, $p<0.001$, CS+: 0.59±1.12, CS-: 0.16±0.09) whereas the Ext group did not ($p=0.574$).

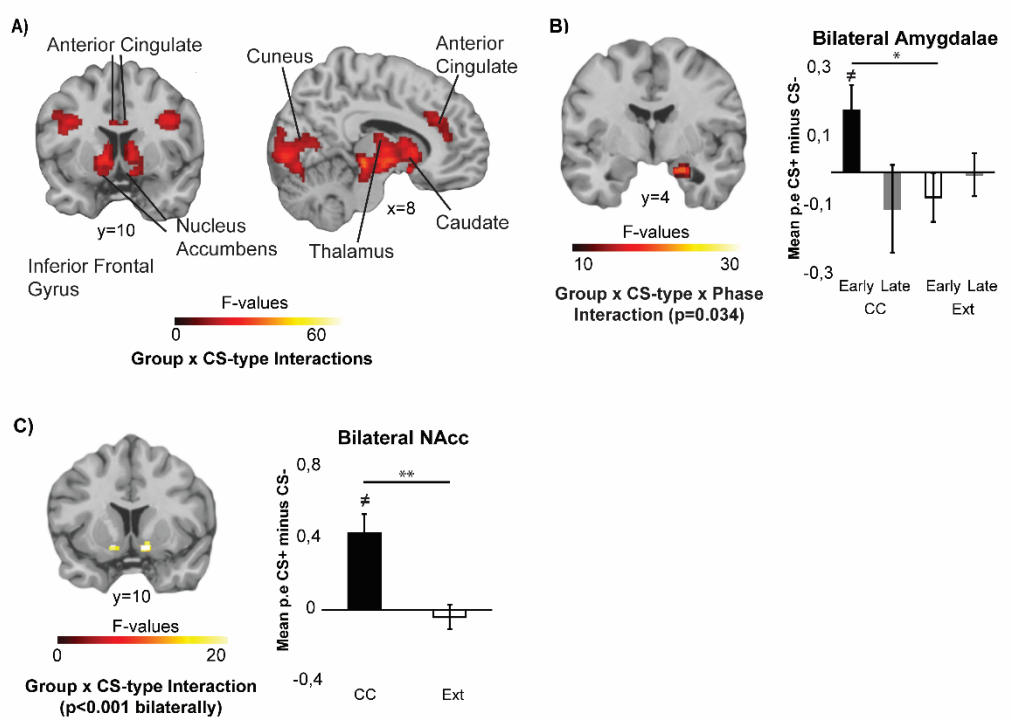


Figure 4.5. Stimulus-type specific activation differs between participants undergoing CC and extinction. A. Whole-brain Group x CS-type interaction effects revealed distinct stimulus-specific activation of regions including the anterior cingulate, cuneus, nucleus accumbens, caudate, thalamus and inferior frontal gyrus during the counterconditioning vs. extinction phase. Panel A displays group F-images (see table 1 for directions) thresholded at FWE-corrected $p<0.05$, cluster-forming threshold $p=0.001$. **B**. The right amygdala showed a Group x CS-type x Phase interaction during the CC/extinction task, where CC appears to be accompanied by a decreasing activation of the amygdala while extinction does not**. C.** The bilateral NAcc showed a Group x CS-type interaction during the CC/extinction task, where the CC group showed increased NAcc activation in response to the CS+ compared to the CS- while the Ext group did not. Panel B and C display group F-images thresholded at FWE-SVC $p<0.05$, cluster-forming threshold $p=0.001$, along with post-hoc tests on parameter estimates from the full ROI included in analysis. ** $p<0.01$, * $p<0.05$, ≠ indicates that the bar is significantly different from 0.

Table 4.3. Whole-brain main effects of group (CC, Ext), CS type (CS+, CS-) and phase (early, late) and interactions, during the counterconditioning/extinction task. Cluster-forming threshold p=0.001, FWE-corrected at p<0.05, clusters were labelled using the taliarach deamon atlas and the AAL atlas for ROIs. For each cluster, the peak voxel coordinates (MNI space) and regions are reported, and additional regions contained within the cluster are added in italics. See supplementary table 4.4 for main effects of CS-type.

		Peak MNI coordinate						
Region	*Cluster*	*x*	*y*	*z*	*Size (mm3)*	*pFWE (cluster)*	*Peak F-value*	*Direction*
Group x CS-type x phase								
Parahippocampal Gyrus BA34R ***Parahippocampal Gyrus Amygdala, Uncus BA34R***	1	18	-8	-20	1760	0.034	23.40	CS+<CS- difference increases from early to late phase for CC, not for Ext
Group x CS-type								
Lateral Geniculum Body LR, Caudate Head LR, Thalamus LR, Lentiform Nucleus LR	1	2	-26	-18	29920	<0.001	73.15	
Cuneus L ***Lingual Gyrus BA17/BA18 LR, Posterior Cingulate LR, Cuneus BA18R, Cuneus BA30L Declive R***	2	-6	-96	2	23272	<0.001	43.50	
Inferior Frontal Gyrus BA47L ***Insula BA13 L***	3	-36	18	-6	4504	0.009	30.62	
Extra-Nucleus R	4	30	26	2	3136	0.016	37.67	
Superior Temporal Gyrus L ***Superior Temporal Gyrus BA41 L, Transverse Temporal Gyrus L***	5	-60	-44	14	9088	0.002	43.56	CS+>CS- for CC, less for Ext
Transverse Temporal Gyrus BA41 R ***Superior Temporal Gyrus R, Superior Temporal Gyrus BA42/BA22R***	6	44	-22	12	7784	0.003	42.17	
Anterior Cingulate BA32R ***Anterior Cingulate BA32L, Cingulate Gyrus R***	7	6	30	26	8880	0.002	27.90	
Precentral Gyrus L ***Inferior Frontal Gyrus L***	8	-36	0	30	3624	0.014	30.10	
Precentral Gyrus R ***Sub-Gyral R***	9	40	2	32	4056	0.011	40.64	
Precentral Gyrus BA6L ***Middle Frontal Gyrus BA6L***	10	-44	-6	52	2184	0.028	24.34	CS+>CS- for CC, less for Ext
Angular Gyrus R ***Supramarginal Gyrus R***	11	54	-60	36	1944	0.032	24.18	CS+<CS- for CC, less for Ext
Group x Phase								
No significant clusters								
CS-type x Phase								
No significant clusters								
Group								
No significant clusters								
Phase								
Inferior Frontal Gyrus R ***Inferior Frontal Gyrus BA45 R***	1	30	26	8	4848	0.006	40.27	
Insula L ***Superior Temporal Gyrus BA22, Precentral Gyrus L***	2	-28	26	0	4368	0.007	38.41	Early Phase > Late Phase
Postcentral Gyrus L	3	-54	-24	22	1768	0.031	23.75	

Contrast estimates in further a priori defined regions of interest (ROIs) during the CC/Ext task were submitted to a Group (CC, Ext) x CS-type (CS+, CS-) x Phase (early, late) rmANOVA (**Figure 4.6**). The bilateral hippocampi (right hippocampus cluster size: 664 mm^3, p=0.001, FWE-SVC, left hippocampus

cluster size: 112 mm^3, p=0.024, FWE-SVC) and the left vmPFC (mask defined as bilateral gyrus rectus and medial orbital gyri, cluster size = 160 mm^3, p=0.013, FWE-SVC) showed differentially changing CS-type-specific activations between groups (Group x CS-type x Phase interaction). While CS+-specific suppression of these regions appeared to increase during the CC task, this was not the case during the extinction task. Post-hoc comparisons on averaged parameter estimates in the bilateral hippocampi confirmed that stimulus-specific suppression increased during the course of the task in the CC group (t(23)=3.280, p=0.003, early CS+-CS-: 0.054±0.07, late: -0.150±0.07), but not in the Ext group (p=0.266). Post-hoc comparisons across the vmPFC ROI also revealed increased CS+-specific suppression in the CC group compared to the Ext group (t(44)=2.221, p=0.032, CC: -0.189±0.06, Ext: -0.070±0.10). While the extinction group showed increased CS+-specific activation from the early to the late phase of the extinction task (t(21)=2.235, p=0.036, early CS+: -0.149±0.08, late CS+: 0.040±0.09), the CC group did not (p=0.120). During the late phase the CC group showed increased vmPFC deactivation to the CS+ compared to the CS- (t(23)=3.174, p=0.004, late CS+: -0.284±0.06, late CS-: -0.095±0.05), while the Ext group did not (p=0.503). Thus, across both the hippocampus and the vmPFC, counterconditioning induced increased stimulus-specific suppression, while extinction did not.

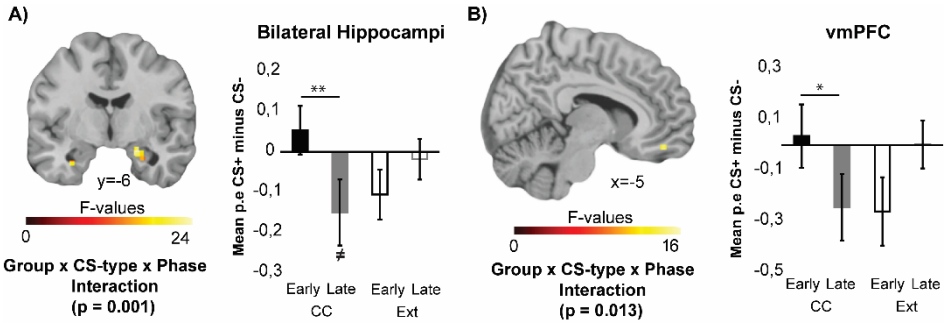


Figure 4.6. ROI analyses during the CC/extinction task reveal distinct activity in the hippocampus and left vmPFC. During the CC/extinction task, stimulus-specific activation of the hippocampus (C) and left vmPFC (**D**) changes differently between groups. ** p<0.01, * p<0.05, ≠ indicates that the bar is significantly different from 0. Group F-images thresholded at FWE-SVC p<0.05, cluster-forming threshold p=0.001, along with post-hoc tests on parameter estimates from the full ROI included in analysis.

During the spontaneous recovery task, a priori defined regions of interest did not reveal any effects (see Supplementary Information).

Counterconditioning retrospectively enhances item recognition for conditioned exemplars

Following the reinstatement test and re-extinction, participants completed a surprise item recognition test approximately 24 hours after acquisition and the CC/extinction task. Threat conditioning has previously been shown to enhance 24-hour item recognition for category exemplars presented during the acquisition phase (Dunsmoor et al., 2012). However, this enhancement for CS+ items does not extend to items presented during an extinction session separated from the acquisition phase by a short break (Dunsmoor et al., 2018). We therefore analysed item recognition for the CS+ and CS- during

acquisition and the CC/extinction phase separately to examine whether the groups differed in recognition memory performance (**Figure**).

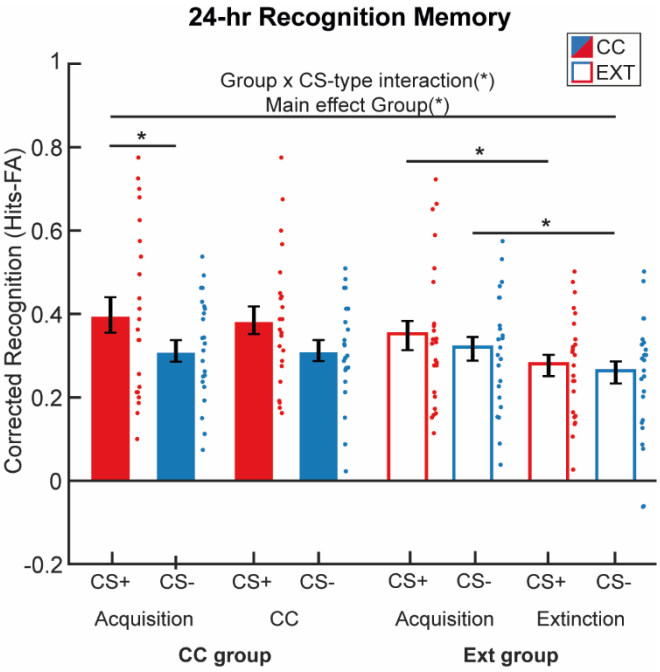


Figure 4.7. Twenty-four hours recognition memory results. During acquisition and extinction on the first day of the experiment, participants viewed trial-unique exemplars from two semantic categories (objects, animals) that served as CS+ and CS-. The next day, participants completed a surprise memory test for these items, mixed with an equal number of novel exemplars. Participants recognized relatively more items from the CS+ category, and participants that underwent CC showed improved item recognition compared to participants in the Ext group. Error bars represent ± standard error of the mean. *=p<0.05.

Corrected recognition scores (pHits – pFA) were subjected to a task (acquisition, CC/extinction task) x CS-type (CS+, CS-) x Group (CC, Ext) rmANOVA including CS+-category (animals, tools) as covariate. Overall, participants showed better memory for items from the CS+ category (main effect of CS-type: ($F_{(1,43)}$=11.550, p=0.001, η^2=0.212) and participants who underwent CC showed better memory as compared to participants who underwent extinction (main effect of Group: ($F_{(1,43)}$=5.829, p=0.020, η^2=0.119). Stimulus-type specific item recognition differed between the CC and Ext groups (CS-type x Group interaction: $F_{(1,43)}$=4.482, p=0.040, η^2=0.094). While participants in the CC group showed better recognition memory for the CS+ category compared to the CS- category (t(22)=2.447, p=0.023, CS+: 0.40±0.04, CS-: 0.32±0.02), participants in the Ext group did not (p=0.384, CS+: 0.30±0.03, CS-: 0.28±0.02). Although the effect of stimulus-type was stronger for tools as CS+, this was not different between groups (see Supplementary Information). Thus, across the acquisition and CC/extinction phase, participants who underwent CC showed a stronger enhancement of CS+ memory compared to the participants that underwent extinction.

To further investigate to what extent CC retroactively affected memory for items presented during the acquisition task, we examined item recognition during acquisition and the CC/extinction tasks

separately. While fear conditioning increased memory for CS+ items presented during the acquisition task across both groups (main effect CS-type: ($F(_{1,42})$=18.147, p=<0.001, η^2=0.301), subsequent CC enhanced this effect (Group x CS-type interaction: ($F(_{1,42})$=5.112, p=0.029, η^2=0.109). Post-hoc tests revealed increased item memory for the CS+ category compared to the CS- category presented during acquisition in the CC group (t(21)=2.341, p=0.029) but not in the Ext group (p=0.122). As the acquisition task was identical between groups, it appears that CC in comparison to extinction retroactively enhanced memory for CS+ items. For items presented during the CC/extinction task, overall item recognition was better in the CC group compared to the Ext group (main effect group: $F(_{1,42})$=5.112, p=0.029, η^2=0.109, post-hoc: CC>Ext (t(43)=2.765, p=0.008, CC: 0.35±0.03, Ext: 0.26±0.02). Thus, compared to regular extinction, CC enhanced recognition of items presented during CC, but interestingly also strengthened the emotional memory enhancement of CS+ exemplars presented during acquisition, suggesting that immediate CC may alter consolidation of a prior fear conditioning episode.

Following previous work (Dunsmoor et al., 2018; Dunsmoor, Murty, et al., 2015; Keller & Dunsmoor, 2020), we explored stimulus type-specific decreases in item recognition between tasks, as well as within-phase differences between item recognition for the CS+ and CS-, in each group. As expected, a post-hoc paired samples t-test showed that participants in the Ext group remembered significantly more CS+ items from the acquisition phase as compared to the extinction phase (t(22)=2.238, p=0.036, acquisition: 0.33±0.03, extinction: 0.27±0.03). In contrast, participants who had undergone CC, remembered CS+ items presented during acquisition and CC equally well (p=0.437, acquisition: 0.41±0.04, CC: 0.38±0.03). Thus, while recognition memory for items encoded during the extinction task was substantially weaker than memory for items from the acquisition task, this was not the case for items presented during CC.

Discussion

This study aimed to test whether CC, as compared to regular extinction, can lead to a more persistent attenuation of threat responses, and to investigate whether this is mediated by neural mechanisms reflecting extinction-related enhanced engagement of the vmPFC or engagement of reward-focused networks. We found that CC prevented differential spontaneous recovery of PDRs compared to regular extinction, supporting that CC reduces recovery of threat responses. fMRI results suggest that CC engages different neural mechanisms compared to extinction. Most notably, while the extinction group showed an increase in CS+-specific vmPFC activation during extinction, the CC group showed CS+-specific deactivation of the vmPFC that persisted during the late phase of CC. Furthermore, CC led to increased NAcc activation for the CS+ as compared to the CS-, while extinction did not. Lastly, phase- and stimulus-specific activation of the hippocampus and the amygdala differed between extinction

and CC. Compared to extinction, CC led to increased activation of the amygdala in the early phase, and increasing stimulus-specific deactivation of the hippocampus over the course of the early and late phases. In addition, CC retrospectively enhanced item recognition for conditioned exemplars presented during acquisition and strengthened memory for conditioned exemplars presented during CC compared to extinction.

The mechanism underlying CC appears to be qualitatively different from the mechanism underlying regular extinction. Regular extinction is associated with activation of the vmPFC (Milad et al., 2007; Phelps et al., 2004) that may inhibit the expression of threat responses by suppressing amygdala activity (Morgan et al., 1993; Phelps et al., 2004; Quirk et al., 2000; Quirk et al., 2003). In comparison to regular extinction, novelty facilitated extinction, a form of enhanced extinction where aversive events are replaced with novel, neutral outcomes, shows stronger CS+-specific vmPFC activation (Dunsmoor, Kroes, Li, et al., 2019). If CC was similarly mediated by enhanced recruitment of extinction networks, we would have expected increased activation of the vmPFC, yet we observed a deactivation of the vmPFC in response to CS+-presentation in the CC group, opposing this hypothesis. Interestingly, deactivation of the vmPFC during CC was also found for a form of counterconditioning induced via real-time fMRI decoded neurofeedback (Koizumi et al., 2017; Taschereau-Dumouchel et al., 2018). During neurofeedback CC sessions, participants implicitly learned to obtain monetary rewards by generating a representation of the target CS+ in the visual cortex (Koizumi et al., 2017). After neurofeedback CC, reductions in threat responses were stronger in participants showing stronger vmPFC deactivation, suggesting that vmPFC disengagement can be involved in the reduction of fear (Koizumi et al., 2017). Taken together, both our findings and previous neurofeedback studies of CC suggest that in contrast to enhanced extinction, CC disengages the vmPFC. Given that we replicate this finding using a different approach, that includes direct exposure to the CS+, vmPFC disengagement may be a central mechanism of CC. The observed pattern of activity, including vmPFC deactivation during CC further bears resemblance to activity patterns during goal-directed eye movements used in EMDR, a technique that has also been shown to improve extinction learning (de Voogd et al., 2018). A similar activity pattern and effect has also been found for working memory-like tasks, such as a game of Tetris (Holmes et al., 2009; James et al., 2015; Price et al., 2013). Similar to working memory-like tasks, goal-directed attention to the MID task during CC may engage executive-control networks while deactivating the salience network (Liang et al., 2016; Qin et al., 2009; Seeley et al., 2007).

The mechanisms underlying CC could be similar to avoidance learning. In avoidance learning, aversive outcomes signalled by a conditioning stimulus can be prevented by performing an instrumental avoidance action. Like CC, active avoidance is associated with stimulus-specific activation of the ventral striatum and is more effective than extinction in persistently diminishing threat responses (Boeke et

al., 2017; Delgado et al., 2009). For CC, a behavioural study in rats has previously shown that CC is more resistant to renewal when the delivery of the appetitive US is contingent on an action (Thomas et al., 2012). However, while active avoidance is associated with increased vmPFC activation in response to the avoided CS (Boeke et al., 2017; Delgado et al., 2009), we observed a deactivation of the vmPFC during CC. Thus, while both CC and active avoidance may enhance recruitment of the ventral striatum through goal-driven avoidance or reward-directed actions, distinct vmPFC activation suggest that they are distinct mechanisms.

The CC procedure led to clear CS+-specific activation of the NAcc, which is in line with expectations for reward anticipation in tasks with a monetary incentive delay aspect (Knutson & Cooper, 2005). CS+-specific activation of the NAcc was not seen in participants undergoing extinction, which may indicate that this is reward-related activation that is specific to CC. However, previous work in rodents revealed an amygdala-ventral striatum (NAcc) pathway that is activated during extinction training (Correia et al., 2016). The recruitment of this pathway was shown to be enhanced during CC, and reduced the return of fear(Correia et al., 2016), suggesting that CC may in fact enhance activation of reward-related networks that are weakly activated by extinction. Indeed, fMRI studies in humans that modelled prediction error for omitted aversive outcome during extinction training (i.e. outcomes "better-than-expected") showed involvement of the NAcc (Esser et al., 2021; Raczka et al., 2011; Thiele et al., 2021). Possibly, activation of the NAcc during extinction is limited to early extinction trials generating prediction errors. Nevertheless, based on our findings, it appears that sustained CS+-specific activation of the NAcc is a distinct mechanism underlying CC but not extinction.

CC strengthened memory for the conditioned category compared to regular extinction. Both reward and threat conditioning can enhance item recognition for CS+ category (Dunsmoor et al., 2012; Patil et al., 2017). On the contrary, after within-session extinction, item recognition of CS+ exemplars presented during extinction drops compared to acquisition, even when extinction was enhanced through novelty-facilitated extinction (Dunsmoor et al., 2018). In contrast to extinction, within-session CC was previously shown to enhance memory, suggesting that CC has a unique, strengthening effect on memory (Keller & Dunsmoor, 2020). In the current study, we replicate this finding, showing strengthened memory after CC compared to extinction. While enhanced recognition of items presented during CC could be mediated by attentional prioritization (Talmi et al., 2008), CC also retrospectively strengthened memory for items presented during acquisition, suggesting that CC may also alter the consolidation of a prior fear conditioning episode. Retroactive enhancement of memory consolidation for related items has previously been shown for conceptually-related neutral items presented prior to threat conditioning (Dunsmoor, Murty, et al., 2015) and reward conditioning (Patil et al., 2017). At a neurobiological level, these findings have been related to the tag-and-capture

hypothesis suggesting that memories for neutral events can be strengthened if they are followed by salient events, thanks to an initially short-lived synaptic "tag" that allows later events to stabilize the memory (Ballarini et al., 2009; Dunsmoor, Murty, et al., 2015; Frey & Morris, 1997). At a system's level, retroactive memory strengthening has been linked to reverse replay (Braun et al., 2018). Specifically, animal research indicates that reward increases reverse replay (Ambrose et al., 2018; Diba & Buzsáki, 2007; Foster & Wilson, 2006), and that reward-induced reverse replay occurs concurrently with firing of midbrain dopamine neurons (Gomperts et al., 2015). Interestingly, spontaneous replay is also involved in regular extinction, where unexpected omission of the feared outcome drives spontaneous reactivations of neural activity patterns evoked in the vmPFC. These spontaneous reactivations are predictive of extinction recall and can be amplified through pharmacological enhancement of dopaminergic activity (Gerlicher et al., 2018). Yet while dopaminergic modulation during extinction may be limited to extinction prediction error signals during the early phase (Esser et al., 2021; Raczka et al., 2011; Thiele et al., 2021), dopaminergic modulation may be sustained throughout the MID-based CC task applied in this study. While we did not measure dopaminergic activity directly, activation of the NAcc during reward anticipation is predictive of dopamine release within the NAcc (Buckholtz et al., 2010; Schott et al., 2008; Weiland et al., 2014, 2017). Given the increased stimulus-specific activation of the NAcc in the CC group, it is likely that dopaminergic activity in the current study was enhanced during CC compared to regular extinction. The enhanced dopaminergic modulation could strengthen memories through replay (Ambrose et al., 2018; Singer & Frank, 2009), or may increase synaptic plasticity directly, potentially explaining enhanced item recognition after CC compared to regular extinction (Atherton et al., 2015; Braun et al., 2018; Brzosko et al., 2015). In line with these findings, research in humans shows that reward systematically modulates memory for neutral objects in a retroactive manner, with objects closest to the reward being prioritized (Braun et al., 2018). It could be that reward-conditioning during CC similarly drives reward-driven reverse replay, which enhances episodic memory for conceptually related items presented during the preceding acquisition task.

It was previously suggested that the effectiveness of CC may be limited because prior threat conditioning interferes with reward learning, leading to lower activation of the reward network compared to reward-conditioned items that were not previously associated with threat (Bulganin et al., 2014). Here we show that CC successfully induced reward anticipation, as evidenced by decreased response latencies, stimulus-specific activation of the NAcc and the enhancement of item recognition. While these effects may be stronger for items that were not previously fear conditioned, prior fear conditioning does not seem to block the efficacy of CC.

Several limitations of the current study are worth considering. First, while the monetary incentive aspect during CC clearly induced positive valence, it also increased physiological arousal, making it difficult to isolate the individual effects of positive valence and reward-induced arousal. While the current results are in line with previous work in CC using low-arousal, positive-valence pictures (Keller & Dunsmoor, 2020), we cannot exclude the possibility that the current findings (in part) reflect differences in task engagement between participants. However, we may ask whether it is meaningful to tease out individual effects of valence and arousal since arousal may facilitate reward processing. Indeed, striatal responses in response to obtained monetary rewards are dependent on salience and are increased when rewards are dependent on active responses compared to passive delivery (Zink et al., 2004). Second, although we included a reinstatement procedure in the experiment, neither the Ext nor the CC group showed differential reinstatement. It is worth noting however, that reinstatement paradigms in humans may not reliably produce differential reinstatement after extinction (Haaker et al., 2014). Third, it is important to note that CC/extinction was carried out within minutes after the acquisition phase, and the effects of CC and extinction may differ when carried out after the acquisition memory has been consolidated (Chang & Maren, 2009; Devenport, 1998; Maren, 2014; Myers et al., 2006). Fourth, whole-brain analysis of the CS-specific activation during the spontaneous recovery test in the Ext group did not yield any clusters above threshold, while physiological results indicated spontaneous recovery of differential threat responses. Given that recovered threat responses are often quick to extinguish, it may be that threat-evoked neural activity was too brief to be detected. Activation of stimulus-specific clusters in the fusiform gyrus and the caudate during the spontaneous recovery phase in the CC group, could support attentional prioritization of the CS+ items because of CC the previous day.

In conclusion, our findings show that appetitive CC improves the retention of safety memory over standard extinction. This effect is associated with stimulus-specific deactivation of the vmPFC and hippocampus, along with increased activation of the NAcc and creates a stronger safety memory compared to extinction. These findings may inform development of future treatments for fear- and anxiety disorders. While a large body of research focuses on enhancing regular extinction, this study indicates that another promising and potentially longer-lasting approach may be to engage reward-circuits. Although further work is needed, a major advantage of CC-based interventions over extinction-based interventions may be that CC could be more tolerable as it may shift attention away from the experience of fear.

Materials and Methods

Participants

Forty-eight healthy right-handed volunteers (15 males, 33 females; age [22.71±0.44]) with no neurological or psychiatric history, uncorrected hearing and normal or corrected-to-normal vision completed the study. Exclusion criteria were pregnancy, disorders of the autonomic system, heart conditions, recreational drug use and any contraindications for MRI. Participants provided written informed consent and were paid 55 euros for their participation. Participants in the CC group were able to earn an additional 14 euros. This study was approved by the local ethical review board (CMO region Arnhem-Nijmegen).

Design and procedure

This study was a two-day between-subjects experiment carried out in the fMRI scanner (see **Figure 4.1** for an overview of the design). Participants were assigned to either the CC or extinction (Ext) group according to a predetermined allocation sequence. At the start of each session, two Ag/AgCl electrodes attached to the medial phalanges of the second and third digit of the left hand, a pulse oximeter was attached to the first digit of the left hand to measure finger pulse and respiration a respiration belt was placed around the abdomen to measure respiration. All measures were taken using a BrainAmp MR system and recorded using the BrainVision Recorder software (Brain Products GmbH, Munich, Germany). The first day consisted of individual adjustment of the electrical shock followed by a single fMRI session that included the following tasks: an object localizer task (17 min, see supplementary information), a category threat conditioning task (23 min) and the CC or extinction task (23 min). The second session took place the following day and consisted of three runs: the spontaneous recovery and reinstatement test (12 min), item recognition test (29 min), valence localizer (17 min).

Pavlovian conditioning paradigm

Note, that CC included an instrumental and not Pavlovian conditioning procedure. This was done because of pragmatic constraints in studies with humans. For example, we could not food deprive humans to make an appetitive reward truly reinforcing and make the participants anticipate the reward. Previous work (Patil et al., 2017; Zink et al., 2004) and our pilot studies indicated that to maximize reward anticipation and evoke conditioned responses, the reward conditioning needed to be instrumental.

The acquisition, counterconditioning, extinction, spontaneous recovery and reinstatement tasks consisted of a categorical differential delay fear conditioning paradigm (Dunsmoor et al., 2012) with elements of the monetary incentive delay task (Knutson et al., 2000). Participants viewed trial-unique exemplars of pictures from two categories (animals or objects, see **Figure 4.1**). In a counter-balanced manner, exemplars from one category served as CS+ (reinforced) stimuli, while exemplars form the

other category served as CS- (unreinforced stimuli). Each trial started by presentation of the stimulus. After a variable delay of 2.5-4s, a cue appeared, to which participants were instructed to respond as quickly as possible with a button press. After a button was pressed, or when the 1s response window had elapsed, the colour of the cue shifted from black to blue. 0.5-1.5s after the response window elapsed, CS+ items presented during the acquisition phase could be reinforced with a shock. During the acquisition phase, 50% of the CS+ pictures was followed by a shock. After 1s, the stimulus was replaced by neutral feedback during the acquisition, extinction, and recovery tasks. During the CC phase, neutral feedback was replaced by monetary feedback. During the CC phase, participants could obtain a €0.50 reward for their quickest responses to the cues presented on top of CS+ stimuli. The response time target was dynamically adjusted to achieve a reinforcement rate of approximately 70%. Reward was withheld during the first three CS+ trials during the CC phase to make the transition from the acquisition to the extinction phase more gradual. The inter-trial interval (ITI) varied randomly between 8 and 10s. Pictures were presented in a pseudorandom order with no more than three repetitions of the same category. The acquisition, extinction and CC blocks consisted of 40 CS+ and 40 CS- presentations each. The spontaneous recovery block consisted of 15 CS+ and 15 CS+ presentations, and the reinstatement test consisted of 5 CS+ and 5 CS- presentations.

Item recognition memory test

Participants carried out a surprise recognition memory test compromised of 160 pictures (80 CS+, 80 CS-) shown during the acquisition and CC/extinction phases, as well as 160 category-matched new items (80 CS+, 80 CS-). Participants rated on a 6-point scale whether the picture was 'definitely old', 'probably old', 'maybe old', 'maybe new', 'probably new', 'definitely new'.

Valence-specific response characterization

The valence-specific response characterization task consisted of an adapted version of the conditioning paradigm used during the acquisition phase. Instead of category items, participants were presented with squares in three different colours. One of the stimuli was reinforced with shocks (CS+-shock, 50% reinforcement rate), one stimulus was reinforced with monetary rewards (CS+-reward, approximately 70% reinforcement rate, response time target adjusted dynamically) and the last stimulus was not reinforced (CS-). Each stimulus was presented 40 times in a pseudorandom order, with no more than three repetitions of each stimulus. Colours and reinforcement (shocks vs. rewards) were counterbalanced across participants.

Peripheral stimulation

Electrical shocks were delivered using two Ag/AgCl electrodes attached to the medial phalanges of the second and third digit of the right hand using a MAXTENS 2000 (Bio-Protech) device. Shock intensity varied in 10 intensity steps between 0 to 40 V and 0 to 80 mA. Shock duration was 200 ms. Shock

intensity was calibrated using an ascending staircase procedure starting with a low voltage setting near a perceptible threshold and increasing to a level deemed "maximally uncomfortable but not painful" by the participant, in keeping with prior threat conditioning protocols (Dunsmoor, Murty, et al., 2015; Kroes, Dunsmoor, Mackey, et al., 2017; LaBar et al., 1998).

Arousal and valence ratings

Arousal and valence ratings were acquired using self-assessment manikin scales. The arousal scale ranged from 1 (=extremely calm) to 10 (=extremely excited). The valence scale ranged from 1 (=extremely negative) to 10 (=extremely positive). The valence and arousal ratings were collected for the two categories (animals and tools) after the acquisition phase, after the CC/extinction phase, at the start of day 2 immediately before the spontaneous recovery test and after the reinstatement test. For the stimuli used in the valence localizer, valence and arousal ratings were collected immediately after the valence-specific response characterization.

Retrospective shock and reward estimation

Participants were asked to estimate the number of shocks, the number of rewards and the reinforcement rate.

SCR pre-processing and analysis

EDA data was pre-processed using in-house software; radio frequency (RF) artefacts were removed and a low-pass filter was applied (de Voogd et al., 2016b, 2016a). Skin conductance responses (SCR) were automatically scored with additional, blinded, manual supervision using Autonomate(Green et al., 2014). SCR amplitudes (measured in µSiem) were determined for each trial as the maximum response with an onset between 0.5 and 7.5s after stimulus onset and maximum rise time of 14.5s. Shock- and reward- reinforced trials were excluded from analysis. All response amplitudes were square-root transformed and normalized according to each participant's mean UCS response prior to statistical analysis. The average SCRs were computed per stimulus type, task, phase (early, late), and participant.

PDR pre-processing and analysis

Pupil dilation was measured with a MR-compatible eye-tracker from SensoMotoric Instrument (MEye Track-LR camera unit, SMI, SensoMotoric Instruments) and sampled at a rate of 50 Hz. Data were analysed using in-house software (Hermans et al., 2013) implemented in Matlab R2018b (MathWorks), based on previously described methods (Siegle et al., 2003). Eyeblink artifacts were identified and linearly interpolated 100 ms before and 100 ms after each identified blink. Data from scan runs missing 50% time points or more were excluded. After interpolating missing values, time series were band-pass filtered at 0.05 to 5 Hz (by subtracting the mean and dividing by the standard deviation) within

each participant and run to account for between-subjects variance in overall pupil size. Event-related pupil diameter responses were calculated by averaging pupil diameter during 3.5 to 7 sec period after stimulus onset, divided by the 1 sec pre-stimulus pupil diameter (-1 to 0 sec). The average PDRs were computed per stimulus type, task, phase (early, late), and participant.

MRI data acquisition

MRI scans were acquired using a Siemens (Erlangen, Germany) 3T MAGNETOM PrismaFit MR scanner equipped with 32-channel transmit-receiver head coil. The manufacturer's automatic 3D-shimming procedure was performed at the beginning of each experiment. Participants were placed in a light head restraint within the scanner to limit head movements during acquisition. Functional images were acquired with multi-band multi-echo gradient echo-planar (EPI) sequence [51 oblique transverse slices; slice thickness, 2.5mm; TR, 1.5s; flip angle, 75°; echo times, 13.4, 34.8, and 56.2 ms; FOV, 210 x 210 mm2; matrix size 84x84x64, fat suppression]. To account for regional variation in susceptibility-induced signal drop out, voxel-wise weighted sums of all echoes were calculated based on local contrast-to-noise ratio after which echo-series are integrated using PAID weighting(Poser et al., 2006). Field maps were acquired (51 oblique transverse slices; slice thickness, 2.5mm; TR, 0.49 s; TE, 4.92 ms and 7.48 ms; flip angle, 60°; FOV, 210 x 210 mm2; matrix size 84x84x64) at the start of each session to allow for correction of distortions due to magnetic field inhomogeneity. A high resolution structural image (1mm isotropic) was acquired using a T1-weighted 3D magnetization-prepared rapid gradient echo sequence [MP-RAGE; TR, 2300 ms; TE, 3.03 ms; flip angle, 8°; 192 contiguous 1 mm slices; FOV = 256 x 256 mm2].

fMRI analysis

Anatomical and functional data were pre-processed using fMRIPrep 20.0.6 (Esteban et al., 2019). The complete boilerplate can be found in the supplementary methods. In brief, MRI data were pre-processed in standard stereotactic (MNI152) space. Pulse and respiration data were processed offline using in-house software and visually inspected to remove artefacts and correct peak detection, and corrected pulse and respiration data were used for retrospective image-based correction (RETROICORplus) of physiological noise artefacts in BOLD-fMRI data (Glover et al., 2000). Identical transformations were applied to all functional images, which were resliced into 2 mm isotropic voxels. After pre-processing in fMRIPrep, functional images were smoothed with a 6 mm FWHM Gaussian kernel (using SPM12; http://www.fil.ion.ucl.ac.uk/spm; Wellcome Department of Imaging Neuroscience, London, UK).

For the acquisition, extinction/cc and spontaneous recovery phases, BOLD responses to CS+, and CS- during the early phase (first half of the trials) and late phase (second half of the trials) were modelled in 4 separate regressors using box-car functions. Additionally, during all these phases, target

presentation, button press and shocks were modelled using stick functions, and feedback presentation and breaks were modelled using box-car functions and included as nuisance regressors. For the category localizer, BOLD responses to animals, objects, and phase-scrambled blocks were modelled in 3 separate regressors using box functions. All first-level models also included six movement parameter regressors (3 translations, 3 rotations) derived from rigid-body motion correction, 25 RETROICOR physiological noise regressors, high-pass filtering (1/128 Hz cut-off), and AR(1) serial correlations correction. First-level contrasts were calculated for early and late CS+ and CS- separately for the acquisition, CC/extinction, and spontaneous recovery phases.

For the acquisition and CC/extinction, first-level contrast were entered into a second-level Group (extinction, cc) x CS-type (CS+, CS-) x Phase (early, late) mixed factorial model using the Multilevel and Repeated Measures (MRM) toolbox (McFarquhar et al., 2016). For the spontaneous recovery test, BOLD-responses from the early phase were entered into a second-level Group (extinction, cc) x CS-type (CS+, CS-) mixed factorial model. Thresholding was achieved using nonparametric permutation testing (5,000 iterations), with a cluster-setting threshold of $p<.001$ for whole-brain analysis and familywise error (FWE) correction at $p<0.05$ at cluster-level for whole-brain analysis and voxel-level for ROI-analysis (Amygdala, Hippocampus, vmPFC, NAcc). Activations are displayed on the single-subject high-resolution T1 volume provided by the Montreal Neurological Institute (MNI).

Region of interest definition

Based on a priori hypotheses, results for the amygdala, NAcc, hippocampus and the ventromedial prefrontal cortex are corrected for reduced search volumes using small volume. Masks were created using the WFU PickAtlas toolbox (Maldjian et al., 2003) in combination with the Automated Anatomical Labeling atlas (Tzourio-Mazoyer et al., 2002) for the bilateral amygdala, bilateral hippocampus and vmPFC (Frontal_Med_orb_L&R and Rectus L&R). The *IBASPM 71* anatomical atlas toolbox (Alemán-Gómez et al. 2006) was used to create a mask for the bilateral NAcc.

Statistical testing

Statistical analyses of behavioural and physiological variables were performed in SPSS (IBM SPSS Statistics Inc.). Dependent measures were submitted to repeated measure ANOVAs and statistics were Greenhouse-Geisser or Huyn-Feldt corrected for non-sphericity when appropriate. Significant findings from ANOVAs were followed-up by paired- and independent samples t-tests. We report partial eta-square as measure of effect size. Means ± s.e.m are provided where relevant unless otherwise indicated.

Deviations from the pre-registration

We pre-registered to sample SCRS in a 0.75 and 3.15 s window after stimulus onset. However, visual inspection of SCR responses during the acquisition phase indicated that response latencies shifted towards the late phase of the trial. We therefore opted to use a longer window (0.5s to 7.5s for stimulus onset) and exclude reinforced trials. The pre-registration erroneously stated that pupil-dilation data would be z-scored and later divided by the pre-stimulus average. PDR data were not z-scored but were only normalized to a 1-sec pre-stimulus baseline. In line with the SCR data, response onset latencies were later than expected. Based on visual inspection of the data from the acquisition phase, we decided to use a window around the expected shock onset: 3.5-7s after stimulus onset. Reinforced trials were excluded. Results for SCR, retrospective reinforcement estimations and the reinstatement test can be found in Supplementary Information. Due to an error in the scripts for the item recognition test, trial-by-trial data were not recorded for the first 12 participants. Therefore, analysis of the memory data focused on averaged data for the early and late phase of acquisition and CC/extinction, leaving out planned change point analyses on bins of 4 trials.

While we planned to extract a vmPFC mask for ROI analysis based on a [CS- > CS+ shock] contrast of BOLD responses during the valence-specific response characterization task to identify “extinction regions”, this did not yield ventromedial prefrontal clusters that survived correction. Instead, in line with our other ROIs, we opted to create a mask based on the AAL atlas. Due to time constraints, native-space and functional connectivity analyses were not carried out for this manuscript.

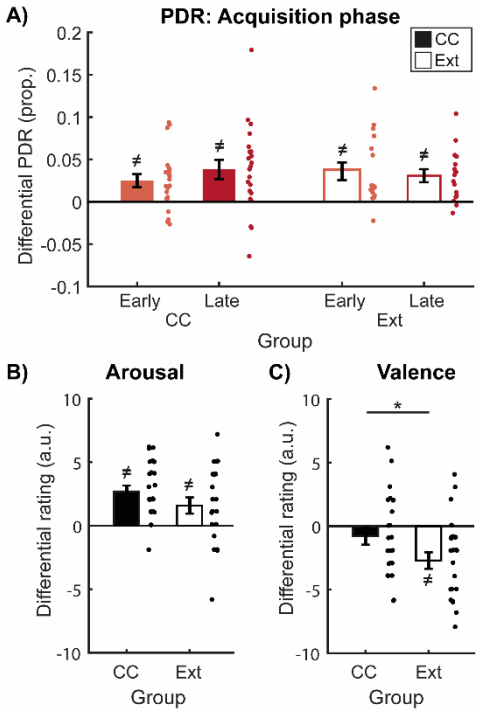
Supplementary Information – Chapter 4

Acquisition phase

Physiological and behavioural evidence for acquisition of conditioned threat responses

Participants pre-assigned to the CC and Ext groups underwent an identical threat acquisition procedure. To verify that participants pre-assigned to both groups acquired conditioned threat memories of comparable strength, we compared PDRs, explicit valence and arousal ratings, and response times between groups. During the acquisition task, participants pre-assigned to both groups showed stable and comparable differential conditioned threat responses as measured by PDRs (**Supplementary Figure 4.8A,** rmANOVA, CS-type (CS+, CS-) x Phase (Early, Late) x Group (CC, Ext), main effect CS-type: $F_{(1,37)}$=41.172, p<0.001, η^2=0.533, other main effects and interactions: all p's>0.2). Both groups also acquired comparable differential skin conductance responses (main effect CS-type: $F_{(1,42)}$=58.633, p<0.001, η^2=0.583), although SCRs showed habituation over the course of the task (main effect phase: $F_{(1,42)}$=66.907, p<0.001, η^2=0.614, all other p's>0.3). SCRs during the acquisition phase demonstrate succesful and comparable acquisition of conditioned threat responses between groups. Thus, both SCRs and PDR demonstrated comparable acquisition of conditioned threat responses between groups.

Successful threat acquisition was further confirmed by valence and arousal ratings for the CS+ and CS- categories at the end of the acquisition task. Arousal ratings for the CS+ category exceeded arousal ratings for the CS- category (**Supplementary Figure 4.8B**, rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), main effect CS-type: $F_{(1,44)}$=27.573, p<0.001, η^2=0.385), and did not differ between groups (all p's>0.2). Similarly, the CS+ category was given lower valence (less positive) ratings than the CS- category (**Supplementary Figure 4.8C**, rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), main effect CS-type: $F_{(1,44)}$=12.626, p<0.001, η^2=0.223). Although there was no main effect of group on valence ratings (p>0.7), the effect of CS-category unexpectedly differed between the CC and Ext group (CS-type x Group interaction: $F_{(1,44)}$=4.512, p=0.039, η^2=0.093), due to more positive ratings to the CS- category in the Ext group (CC: 5.8±0.4, Ext: 6.9±0.3, t(44)=2.156, p=0.037). Nevertheless, valence ratings for the CS+ category were comparable between groups (CC: 5.1±0.5, Ext: 4.2±0.4, p>0.1), suggesting that the strength of the acquired threat memory is likely similar between groups.



Supplementary Figure 4.8. Differential PDRs during acquisition and explicit ratings of arousal and valence provided after acquisition. (A) Differential PDRs for the early (light red) and late (dark red) phase of the acquisition task, (B) arousal and (C) valence ratings, displayed separately for participants assigned to the counterconditioning (CC, solid bars) and extinction (EXT, open bars) groups. Both groups showed comparable differential PDRs and arousal ratings during the acquisition task. For arousal ratings, increased numerical ratings indicate higher levels of arousal. For valence ratings, increased numerical ratings indicate more positive valence. Participants in both groups showed negative differential valence ratings, although the Ext group showed stronger differential valence ratings. Error bars represent ± standard error of the mean. *, p<0.05. ≠. Significantly different from 0

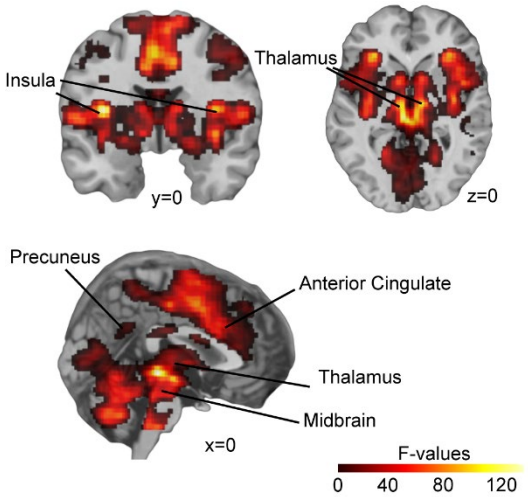
Participants in both groups reported higher estimated reinforcement rates for the CS+ category as compared to the CS- category (rmANOVA, CS-type (CS+, CS-) x Group, $F(_{1,45})=82.176$, $p<0.001$, $\eta^2=0.646$). The reported reinforcement rates did not differ between groups (all p's>0.3).

To keep all experimental tasks similar between groups, participants in both groups were asked to respond to targets that were superimposed on the stimuli as quickly as possible. To verify that both groups performed similarly on this task, we compared response times for the different stimuli between the groups. During the acquisition task, participants responded faster to targets in CS+ trials compared to CS- trials (rmANOVA CS-type (CS+, CS-) x Group (CC, Ext), main effect of stimulus-type: $F(_{1,45})=10.839$, $p=0.002$, $\eta^2=0.194$), with no differences between groups (all p's>0.058).

fMRI results acquisition phase

Successful acquisition of conditioned threat responses

The acquisition of conditioned fear on the first day reliably activated networks associated with fear conditioning. Whole-brain analysis identified regions that were more responsive to the CS+ versus the CS- category (**Figure 4.8** and Supplementary Table 4.3 for a complete overview of findings). We observed differential BOLD responses in a large number of brain areas, including the bilateral insula, posterior and anterior cingulate, thalamus, precuneus (undirected test, cluster size = 425400 mm^3, $p<0.001$, whole-brain FWE-corrected) and the bilateral amygdala (right cluster size = 1088 mm^3, $p<0.001$, FWE-SVC, left cluster size = 736 mm^3, $p<0.001$, FWE-SVC).



Supplementary Figure 4.8. Differential threat responses during acquisition revealed CS-specific activation of clusters encompassing a range of regions including the bilateral insula, thalamus, precuneus, anterior cingulate and midbrain. Group F-image of the effect of CS type, thresholded at cluster-level FWE-corrected p<0.05, cluster-forming threshold p=0.001, displayed on the single-subject high-resolution T1 volume provided by the Montreal Neurological Institute (MNI).

Supplementary Table 4.3. Whole-brain main effects of group (CC, Ext), CS type (CS+, CS-) and phase (early, late) and interactions, during the acquisition task. Cluster-forming threshold p=0.001, FWE-corrected at p<0.05, clusters were labelled using the taliarach deamon atlas and the AAL atlas for ROIs. For each cluster, the peak voxel coordinates (MNI space) and regions are reported, and additional regions contained within the cluster are added in italics.

Region		Peak MNI coordinate						Direction
	luster				*ize (mm3)*	*FWE (cluster)*	*eak F-value*	
CS-type x phase								
Parahippocampa Gyrus L *Insula L, Parahippocamal Gyrus Hippocampus L, Claustrum L,*		18	10	16	656	.005	5.86	

Lentiform Nucleus Putamin L, Uncus L, Postcentral Gyrus BA43 L								
Parahippocampa Gyrus Amygdala R *Inferior Frontal Gyrus R, Subcallosal Gyrus BA34R*	0	4	22	116	.033	5.34		
Culmen L *Declive L, Lingual Gyrus L*	8	54	16	720	.027	0.57		
Parahippocampal Gyrus L *Parahippocampal gyrus BA36L/BA30L, Culmen L*	20	42	2	236	.006	1.65		Early CS+ > Late CS+
Medial Frontal Gyrus BA11 L *Anterior Cingulate BA32L, Medial Frontal Gyrus BA10 R, BA11 R*	4	8	14	104	.046	7.23		
Superior Temporal Gyrus L *Middle Temporal Gyrus BA21/BA22 L*	52	0	14	056	.015	8.06		
Lingual Gyrus BA18/BA19 R	6	68	2	584	.018	5.21		
Insula R *Inferior Parietal Lobule R, Superior Temporal Gyrus BA22 R, Postcentral gyrus BA3 R, Superior Temporal Gyrus BA22 R, Precentral Gyrus BA4/BA6, Inferior Parietal Lobule BA40, Middle temporal gyrus, Superior temporal gyrus BA42*	8	6	8	4488	.001	7.62		
Parahippocampal Gyrus R	4	36	4	432	.035	2.26		
Inferior Frontal Gyrus BA45 R	0	2	4	4	392	.036	1.37	
Precentral Gyrus L	1	60	8	2	640	.006	2.09	
Inferior Parietal Lobule BA40L *Postcentral gyrus BA2L*	2	56	36	2	104	.046	0.42	
Precuneus L *Postcentral gyrus L, cingulate gyrus L*	3	14	42	4	496	.034	9.85	
Precuneus R *Paracentral Lobule Ba7 R, Precuneus R, Cingulate gyrus R, Superior Parietal Lobule BA7 R*	4	0	52	4	528	.006	4.86	
Medial Frontal gyrus L (23) *Medial frontal gyrus BA6LR, Paracentral Lobule L*	5	6	20	4	840	.025	1.47	
CS-type								
Postcentral Gyrus L *Inferior Parietal Lobule LR, Insula LR, Postcentral gyrus R, Cingulate Gyrus LR, Thalamus LR, Caudate LR, Inferior- Middle- and Superior Frontal Gyrus LR, Posterior Cingulate R, Precentral Gyrus LR, Precuneus L, Delice R, Culmen R, Cuneus L, Superior Temporal Gyrus LR, Anterior Cingulate LR, Parahippocampal Gyrus BA27 R, Lentiform nucleus LR*	50	20	6	25400	0.001	95.37		CS+>CS-
Posterior Cingulate BA31 L *Precuneus M*	4	56	4	816	.021	6.17		CS+<CS-
Corpus Callosum M *Corpus Callosum R*			2	296	.049	5.45		

Angular Gyrus R *Angular Gyrus BA39 R, Precuneus R*	6	66	0	432	.024	1.42		
Angular Gyrus BA39L	54	68	0	584	.010	6.02		
Superior Frontal Gyrus BA9L *Superior Frontal Gyrus BA8L, Middle frontal gyrus BA6L*	18	0	2	200	.007	3.18		
Phase								
Superior Temporal Gyrus LR, *Inferior Parietal Lobule R, Middle Temporal Gyrus LR, Inferior- Middle- and Superior Frontal Gyrus LR, Caudate LR, Middle Occipital Gyrus LR, Cingulate Gyrus LR, Anterior Cingulate LR, Declive LR, Precuneus LR, Insula LR, Culmen LR, Superior Temporal Gyrus LR, Lingual Gyrus LR, Fusiform Gyrus LR, Angular Gyrus R, Claustrum LR, Thalamus LR, Parahippocampal Gyrus LR, Cuneus LR*	64	38	2	84632	0.001	7.44		Early>Late

Decreased response times in CS+ trials during CC indicate motivation to obtain rewards

As a measure of motivation to obtain rewards during the CC task, we compared response times to trials of the different stimulus types between participants undergoing CC and extinction. Similar to the acquisition task, both groups were quicker to respond to CS+ trials as compared to CS- trials during the CC/extinction task (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), main effect of CS-type: $F_{(1,44)}=42.736$, $p<0.001$, $\eta^2=0.493$), yet the difference was larger for participants undergoing CC (Group x CS-type interaction: $F_{(1,44)}=8.733$, $p=0.005$, $\eta^2=0.166$). While response times in CS- trials were comparable between groups ($p=0.958$, CC: 0.40±0.02, Ext: 0.40±0.01), participants undergoing CC were quicker to respond during CS+ trials as compared to participants undergoing extinction ($t(39.536)=2.314$, $p=0.026$, CC: 0.35±0.01, Ext: 0.39±0.01). Decreased response times to CS+ trials in the CC group as compared to the Ext group suggest that the obtained monetary reward was motivating participants in the CC group to respond as quickly as possible.

Counterconditioning and extinction are reflected in SCRs

Differential SCRs were still apparent during the CC/extinction phase (rmANOVA, CS-type (CS+, CS-) x Phase (Early, Late) x Group (CC, Ext), main effect CS-type: $F_{(1,40)}=17.609$, $p<0.001$, $\eta^2=0.306$). To verify that successful extinction was reached by the end of the phase, we explored SCRs in the late phase separately, but found that differential SCRs persisted during the second half of the CC/extinction phase ($F_{(1,41)}=12.166$, $p=0.001$, $\eta^2=0.229$). Finally, we explored whether the last two trials of the extinction phase showed evidence of residual differential SCRs. In the last two trials of the extinction, across both groups, there is no evidence for differential SCRs (all p's>0.2). Thus, while differential SCRs persist during the late phase of the extinction task, differential responses are no longer apparent in the last two trials. Throughout the CC/extinction, there is no evidence for different SCRs between groups, suggesting that participants in both groups undergo a comparable but slow extinction of differential SCRs.

Overlapping stimulus-specific activation during CC and extinction

A number of clusters showed comparable stimulus-specific activations during CC and extinction (Supplementary Table 4.4).

Supplementary Table 4.4. Whole-brain main effect of CS-type during the counterconditioning/extinction task. Cluster-forming threshold p=0.001, FWE-corrected at p<0.05, clusters were labelled using the taliarach deamon atlas and the AAL

atlas for ROIs. For each cluster, the peak voxel coordinates (MNI space) and regions are reported, and additional regions contained within the cluster are added in italics.

Region	luster	Peak MNI coordinate			ize (mm3)	FWE (cluster)	eak F-value	Direction
CS-type								
Caudate Head L *Thalamus LR, Caudate Head R, Substantia Nigra LR*		10	0	2	5136	.001	0.98	
Insula R *Inferior Frontal Gyrus R, Precentral Gyrus BA44 R, Inferior Frontal Gyrus BA45 R*		8	6		2800	.001	9.75	
Inferior Frontal Gyrus L *Insula BA13 L*		32	8		808	.004	2.01	
Lingual Gyrus L *Inferior Occipital Gyrus L, Cuneus L, Middle Occipital Gyrus L*		24	80	12	696	.006	2.81	
Superior Temporal Gyrus R *Transverse Temporal Gyrus R*		0	18		744	.012	0.65	
Lingual Gyrus R *Cuneus R*			94		688	.012	7.81	CS+>CS-
Superior Temporal Gyrus L *Transverse temporal Gyrus L*		44	24		864	.011	4.50	
Anterior Cingulate BA32 R *Medial Frontal Gyrus BA8 R, Anterior Cingular LR, Cingulate Gyrus BA32 R*			8	0	064	.004	6.61	
Superior Temporal Gyrus R *Supramarginal Gyrus R, Inferior Parietal Lobule BA40R*	0	4	34	4	720	.012	5.88	
Superior Temporal Gyrus L	1	60	46	6	504	.023	6.62	
Cingulate Gyrus L *Posterior Cingulate BA23R, Posterior Cingulate L*	3	6	20	0	928	.011	8.33	
Angular Gyrus L *Middle Temporal Gyrus L, Angular Gyrus BA39 L*	2	44	64	2	104	.010	3.58	
Inferior Temporal Gyrus BA21 L, *Middle Temporal Gyrus BA21 L*		64	10	22	696	.039	7.05	
Angular Gyrus R *Supramarginal Gyrus R*	4	4	66	4	392	.050	8.35	CS+<CS-
Postcentral Gyrus BA40R *Precentral Gyrus Ba4/BA3 R*	5	4	40	8	704	.038	9.84	
Middle Frontal Gyrus BA8/BA6 L	6	24	6	8	576	.008	4.03	

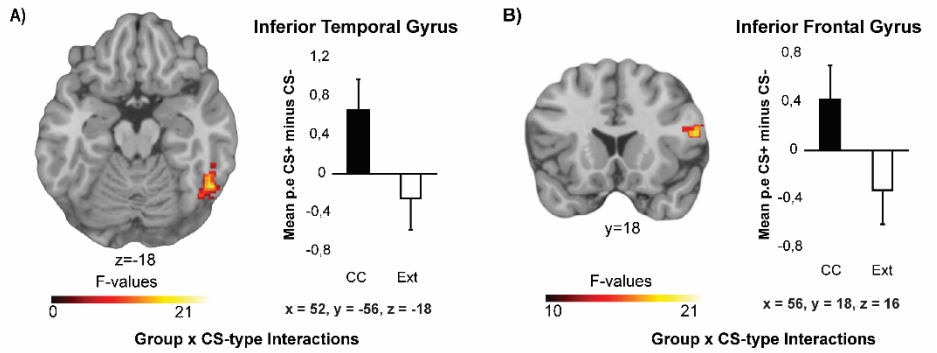
SCRs do not show evidence for differential spontaneous recovery

To investigate whether CC can prevent spontaneous recovery of differential SCRs, SCRs during the last two trials of extinction and the first two trials of the spontaneous recovery test are submitted to a CS-type (CS+, CS-) x Phase (last 2 trials of the CC/extinction phase, first two trials of the spontaneous

recovery test) x Group (CC, Ext) rmANOVA. SCRs showed a generalized increase from the last two trials of extinction to the first two trials of the spontaneous recovery test (main effect phase: F(1,38)=32.392, p<0.001, η^2=0.460)). There is evidence for differential SCRs across both phases (main effect CS-type: (F(1,38)=9.560, p=0.004, η^2=0.201), as CS+ stimuli evoked higher SCRs than CS- stimuli (t(43)=2.518,p=0.016, CS+:0.41±0.03, CS-:0.35±0.03), yet we do not find evidence for CS+-specific spontaneous recovery or effect of group (all p's>0.4). Thus, although there is a generalized increase in responding from the end of extinction to the start of the spontaneous recovery test, SCRs do not show differential recovery and are comparable between groups.

Spontaneous recovery test

During the spontaneous recovery test, CS-specific activation differed between groups in the inferior temporal gyrus (cluster size = 2008 mm^3, p=0.020, FWE-corrected, **Figure 4.9A**) and the inferior frontal gyrus (cluster size = 1920 mm^3, p=0.022, FWE-corrected, **Figure 4.9B**). Separate analysis of the spontaneous recovery phase within each group did not reveal any suprathreshold clusters in the Ext group, while a number of clusters showed stimulus-specific activation in the CC group. Specifically, the CC group showed stimulus-specific activation in the bilateral fusiform gyri, superior parietal lobes and inferior frontal gyri, and in the right thalamus, caudate, middle frontal gyrus, and angular gyrus (see **Table 4.5)**. A priori defined regions of interest (ROIs) during the spontaneous recovery task were submitted to a Group (CC, Ext) x CS-type (CS+, CS-) x Phase (early, late) ANOVA but did not reveal any effects.



Supplementary Figure 4.9. During the spontaneous recovery test, stimulus type-specific activation of the inferior temporal and frontal gyri differed between groups. The inferior temporal Gyrus (**A**) and Inferior frontal gyrus (**B**) show increased CS+-specific activation in the CC group as compared to the Ext group. Group F-images thresholded at FWE-corrected p<0.05, cluster-forming threshold p=0.001, displayed on the single-subject high-resolution T1 volume provided by the Montreal Neurological Institute (MNI) and parameters estimates from peak voxels.

Supplementary Table 4.5. Peak voxel coordinates and statistics of activations during the spontaneous recovery phase in the CC group. Clusters were labelled using the AAL atlas. For each cluster, the peak voxel coordinates and regions are reported, and additional regions contained within the cluster are added in italics. Clusters are whole-brain FWE-corrected at p<0.05.

Region	*luster*	Peak MNI coordinate			*ize (mm3)*	*FWE (cluster)*	*eak T-value*	*Direction*
CS-type								
Thalamus R								
Parahippocampal Gyrus R		0	22	4	160	0.001	.70	CS+>CS-
Inferior temporal Gyrus R		6	52	2	856	0.001	.56	
Fusiform gyrus R								
Inferior frontal gyrus, triangular R		8	6	6	992	0.001	.93	

	Superior parietal lobe R							
	Angular Gyrus R	6	60	0	048	0.001	.36	
	Inferior frontal gyrus,							
orbital part L		0	6	6	120	.019	.31	
	Fusiform gyrus L							
	Lingual gyrus	38	80	18	176	.015	.18	
	Caudate R							
				0	952	0.001	.87	
	Middle Frontal gyrus R							
		7		4	92	.03	.86	
	Superior parietal lobule							
L		32	58	8	480	.004	.53	
	Angular gyrus L							

A reinstatement procedure did not trigger reinstatement of differential conditioned threat responses

To test whether CC additionally reduced reinstatement of conditioned threat, participants received three unsignalled shocks to trigger reinstatement of differential conditioned threat responses. However, across both the CC and Ext group, we did not observe reinstatement of differential conditioned PDRs. On the contrary, PDRs showed a generalized decrease from the last two trials of the spontaneous recovery test to the first two trials of the reinstatement test (main effect of phase ($F_{(1,29)}$=9.104, p=0.005, η^2=239)). Mean PDRs decreased from spontaneous recovery to reinstatement (t(30)=3.063, p=0.005, last two trials of spontaneous recovery: 1.04±0.01, first two trials of reinstatement: 1.01±0.01). Given that we did not observe successful reinstatement in either group, our reinstatement test does not inform us about whether CC can lead to a more persistent attenuation of fear as compared to classic extinction.

SCRs showed a generalized increase from the spontaneous recovery phase to the reinstatement test (rmANOVA, CS-type (CS+, CS-) x Group (CC, Ext), x phase (spontaneous recovery test, reinstatement test), main effect phase: F(1,39)=25.758, p<0.001, η^2=0.398, last two trials of spontaneous recovery: 0.22±0.04, first two trials of reinstatement: 0.38±0.03). Across the last two trials of the spontaneous recovery test and the first two trials of the reinstatement test, differential SCRs differ between the counterconditioning and extinction group (interaction effect of stimulus type and group: F(1,39)=4.967, p=0.032, η^2=0.113). Yet, there is no evidence for differential reinstatement between groups (no CS-type x Phase x Group interaction, p=0.218). Moreover, mean SCRs to CS+ and CS- stimuli do not differ within either group (all p's>0.12).

After the reinstatement test and subsequent re-extinction, valence ratings continued to differ between groups (main effect group: $F_{(1,44)}$=8.602, p=0.005, η^2=0.164). Participants in the CC group gave overall lower mean ratings than participants in the Ext group (CC: 5.5±0.18, Ext: 6.2±0.16), but there was no main effect or interaction of CS-type (all p's>0.3). Differential arousal ratings differed between groups after the reinstatement test and subsequent re-extinction (CS-type x group interaction: $F_{(1,44)}$=8.977, p=0.004, η^2=0.169). Although participants in both groups gave higher arousal ratings to the CS+ category as compared to the CS- category, the difference was larger for participants that underwent CC (t(44)=2.996, p=0.004, CC: 2.0±0.44, ext:0.25±0.36).

CS+-specific enhancement of recognition memory depends on CS+ category

Corrected recognition scores (pHits – pFA) were subjected to a task (acquisition, CC/extinction task) x CS-type (CS+, CS-) x Group (CC, Ext) rmANOVA including CS+-category (animals, tools) as covariate. Although the effect of CS-type differed depending on the category used as CS+ (CS-type x CS+-category interaction: $F_{(1,42)}$=19.400, p<0.001, η^2=0.316) and task (CS-type x CS+-category x task interaction: $F_{(1,43)}$=5.375, p=0.042, η^2=0.095) where the effect of stimulus-type was stronger for tools as CS+, this was not different between groups.

To further investigate to what extent CC retroactively affected memory for items presented during the acquisition task, we examined item recognition during acquisition and the CC/extinction tasks

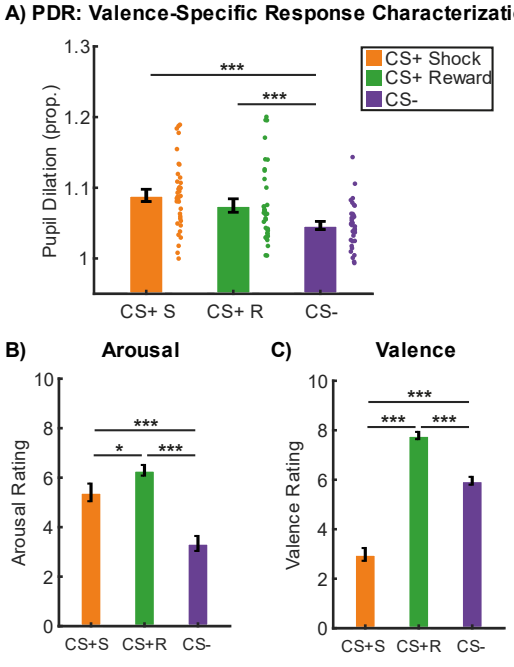
separately. Retrospective memory enhancement for the CS+ items compared to CS- items differed depending on the CS+ category during both the acquisition task (CS-type x CS+ category interaction: $F_{(1,42)}=29.730$, $p<0.001$, $\eta^2=0.414$, CS+ category main effect: $F_{(1,42)}=5.346$, $p=0.026$, $\eta^2=0.113$) and the CC/extinction task (CS-type x CS+ category interaction: $F_{(1,42)}=8.706$, $p=0.005$, $\eta^2=0.172$, stronger stimulus-type effect for tools as CS+), but this effect was comparable between groups.

Valence-specific response characterization

Characterizing aversive and appetitive responses

To investigate to what extent SCRs can be used to disentangle anticipation of shock and reward, participants underwent a simplified version of the main experimental task in which category exemplars were replaced by coloured squares, at the end of the experiment. During the valence-specific response characterization task, we observed habituation in SCRs over the course of the task (rmANOVA, CS-type (CS+ S, CS+ R, CS-) x Phase (early, late) x Group (CC, Ext), main effect phase: $F_{(1,76)}=78.460$, $p<0.001$, $\eta^2=0.674$) and different SCR magnitudes for the three different stimulus types (main effect CS-type (CS+S, CS+S, CS-): $F_{(1,76)}=78.460$, $p<0.001$, $\eta^2=0.674$). In addition, habituation depended on stimulus type (CS-type*phase interaction: $F_{(1,76)}=6.825$, $p=0.002$, $\eta^2=0.152$). During the early phase, SCRs in response to the rewarded CS+ and the CS- were not distinguishable ($t(40)=0.115$, $p=0.909$, CS+R: 0.32±0.03, CS-:0.32 ±0.03), while during the late phase SCRs to the rewarded CS+ were larger than the CS-($t(40)=4.993$, $p<0.001$, CS+R: 0.29±0.03, CS-:0.19±0.02). SCRs to the shock reinforced CS+S were consistently larger than SCRs to the CS+R (early: $t(41)=9.345$, CS+S: 0.62±0.04, $p<0.001$, late: $t(40)=5.952$, $p<0.001$, CS+S: 0.56±0.04) and the CS- (early: $t(40)=10.020$, $p<0.001$, late: $t(4)=10.122$, $p<0.001$). Thus, anticipation of aversive reinforcement (CS+S) led to increased SCRs compared to anticipation of reward (CS+R) and CS- presentation throughout the task. Although the CS+R and the CS- elicited comparable SCRs during the early phase, the CS+R elicited stronger SCRs during the late phase.

After the acquisition task, participants in the CC group underwent appetitive CC and were able to obtain monetary rewards in a task similar to the monetary incentive delay task (Knutson et al., 2000). To facilitate the interpretation of the CC/extinction tasks, we included an additional task at the end of the experiment to characterize aversive and appetitive responses (**Figure 1E**). The valence-specific response characterization served to evaluate to what extent PDRs can be used to disentangle anticipation of aversive and appetitive reinforcement (i.e. shock and monetary reward) and to verify that responses to a target in the tasks did not obstruct physiological measures of differential conditioned threat responding. Participants viewed three different coloured squares and learned that



Supplementary Figure 4.10. PDRs, explicit arousal and valence rating for the different CSs presented during the valence-specific response characterisation task. (A) PDRs to the shock reinforced (CS+S), reward reinforced (CS+R) and CS- stimuli, averaged across the task and all participants. PDRs were increased for the CS+S and CS+R as compare to the CS- (B) Explicit ratings of arousal and (C) valence provided immediately after the task. Explicit ratings of arousal for the CS+S exceeded ratings for the CS-, and the CS+R was rated higher in arousal than the CS+S. Valence ratings for the CS+R were more positive than ratings for the CS-, while ratings for the CS+S were more negative than the CS- and CS+R. Error bars represent ± standard error of the mean *=p<0.05, ***=p<0.001

one colour was associated with shocks (CS+S), one colour with rewards (CS+R) and one colour served as CS-. The trial structure was otherwise identical to comparable trials from the acquisition and CC phases. At the end of the task, participants were asked to rate the three stimuli on valence and arousal self-assessment manikin scales (Bradley & Lang, 1994).

While both shock anticipation and reward anticipation led to similar increases in PDRs as compared to the neutral condition, valence and arousal ratings indicated that participants experienced shock and reward trials differently. In comparison to the neutral CS-, both the shock-reinforced CS+ (CS+S) and reward-reinforced CS+ (CS+R) evoked larger PDRs (**Figure 4.10A**, $t(36)=7.071$, $p<0.001$ and $t(26)=4.900$, $p<0.001$ respectively, CS+S: 1.05±0.03, CS+R: 1.04±0.04, CS-: 1.01±0.02). However, reward- and shock-induced PDRs did not differ statistically ($t(36)=1.146$, $p=0.259$). Explicit ratings of valence confirmed that the CS+R was rated more positive than the CS- ($t(47)=9.046$, $p<0.001$, CS+R: 7.79±0.14, CS-: 5.96±0.16, **Supplementary Figure 4.10C**) while the CS+S was rated less positive than the CS- ($t(47)=-10.337$, $p<0.001$, CS+S: 2.96±0.25). Participants reported increased arousal to both the CS+S and CS+R as compared to the CS- ($t(47)=4.666$, $p<0.001$ and $t(47)=8.897$, $p<0.001$ respectively, CS+S: 5.42±0.35, CS+R: 6.31±0.21, CS-: 3.33±0.30, **Figure 4.10B**). While it was not possible to distinguish PDRs to the CS+S and CS+R, explicit ratings of arousal were marginally increased for the CS+R as compared to the CS+S ($t(47)=-2.100$, $p=0.041$). In conclusion, the response characterization shows that while anticipation of reward and shock both generate increased PDRs as compared to the CS-, they nevertheless show distinct retrospective valence ratings in the expected directions.

Supplementary methods

fMRIPrep preprocessing details

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.0.6 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.2.0 (Avants et al. 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, RRID:SCR_001847, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al. 2017). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym], ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

Functional data preprocessing

For each of the 6 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A B0-nonuniformity map (or fieldmap) was estimated based on a phase-difference map calculated with a dual-echo GRE (gradient-recall echo) sequence, processed with a custom workflow of SDCFlows inspired by the epidewarp.fsl script and further improvements in HCP Pipelines (Glasser et al. 2013). The fieldmap was then co-registered to the target EPI (echo-planar

imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other SDCflows tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin6Asym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.6.2 (Abraham et al. 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

Copyright Waiver

Chapter 5. Online survey study: Public attitudes towards Memory Modification Techniques

Maxime C. Houtekamer, Lisa Wirz, Judith Homberg, Marloes J.A.G. Henckens, Pim Haselager, Erno J. Hermans

Abstract

Neuroscientific advances in the field of memory may soon lead to the introduction of novel Memory Modification Techniques (MMTs) that could improve treatment efficacy for disorders such as post-traumatic stress disorder (PTSD). In response to the development of MMTs, bioethicists have raised theoretical concerns regarding their ethical, legal, and societal implications. It is unclear, however, whether similar caution exists in the general public. Here, we report results of an online survey exploring public attitudes towards the use of reconsolidation-based MMTs for the treatment of PTSD. Attitudes towards MMTs were somewhat positive and dependent on the scenario in which they are used. Safety beliefs were strongly predictive of attitudes towards MMTs, while demographic factors and moral intuitions contributed minimally. We did not find evidence that more extensive information about the scientific foundation of MMTs modulated attitudes. This study provides preliminary evidence that MMT-based treatment options may be well-received by the public

Introduction

Emotional memories are typically strengthened compared to memories for neutral events. In some cases, strong emotional memories can become maladaptive and contribute to mental disorders, such as trauma- and stress-related disorders like Posttraumatic Stress Disorder (PTSD). Problematically, these memories are rather resistant to change, rendering the associated disorders resistant to treatment and patients prone to relapse. Recent advances in the understanding of memory manipulations, however, may pave the way towards novel treatments that selectively modify or inhibit maladaptive memories (for a review, see Phelps and Hofmann 2019). At the same time, memories are tightly connected to our sense of who we are (Prebble et al., 2013) and popular culture warns us that tinkering with memories may have unwanted consequences. The movie "The eternal sunshine of the spotless mind" for example, tells the story of a separated couple that undergoes a medical procedure to erase each other from their memory, and asks us whether forgetting painful memories is the right thing to do. Hence, while the possibility to edit troubling memories may hold great clinical promise, it also offers opportunities for abuse, and we need to carefully consider societal expectations, hopes, as well as concerns regarding Memory Modification Techniques (MMTs). This raises the question: to what extent would MMTs be accepted in society?

While the notion of erasing all memories of a relationship remains science-fiction, MMTs may soon have applications for traumatic memories (Astill Wright et al., 2021) and the treatment of addiction (Chen et al., 2019). Here, we focus on the application of MMTs for the treatment of Posttraumatic Stress Disorder (PTSD) as a result of traumatic experiences and ask whether, and under what conditions, the public would approve of manipulating traumatic memories. Most people experience at least one traumatic event during their lifetime (Kilpatrick et al., 2013; Knipscheer et al., 2020), and while the majority recovers from the initial stress, traumatic experiences can lead to the development of PTSD. Patients affected by PTSD suffer from intrusions and try to avoid thoughts or external reminders that are associated with the trauma. In addition, they experience negative changes in cognition and mood, and may experience symptoms of hyper-arousal (American Psychiatric Association, 2013). The primary treatment for PTSD is exposure therapy (Vervliet et al., 2013), during which people are typically guided to re-imagine the traumatic experience vividly while re-evaluating and reinterpreting aspects of the situation. This allows their emotional responses to decrease and feeling of control to increase (Lang, 1977). Whereas exposure therapy may initially improve PTSD symptoms, relapse is common (Vervliet et al., 2013). It is thought that the extinction process during exposure therapy creates a novel safety memory that inhibits the expression of the traumatic experience, while it leaves the original trauma memory intact, allowing the traumatic memory to resurface over time (Bouton, 2002). MMTs aim to overcome this issue by directly modifying the original

traumatic memory or artificially strengthening the extinction learning that creates a novel safety memory. Several types of interventions have been proposed as MMTs for traumatic memories (see Table 5.1, for a review see e.g. Parsons & Ressler, 2013 or Phelps & Hofmann, 2019).

Table 5.1. Examples of different types of proposed MMTs

Type of intervention	Name
Pharmacological manipulation of extinction	D-cycloserine (DCS) (Inslicht et al., 2021; Ressler et al., 2004; Walker et al., 2002)
	Selective Serotonin-reuptake-inhibitors (SSRIs) (Bui et al., 2013; C. hao Yang et al., 2012)
	Glucocorticoids (De Quervain et al., 2011; Inslicht et al., 2021; Surís et al., 2010; Yang et al., 2006; Yehuda & LeDoux, 2007)
	Cannabinoids (Chhatwal et al., 2005; Chhatwal & Ressler, 2007; Das et al., 2013; Rabinak et al., 2013)
	Oxytocin (Acheson et al., 2013)
Manipulation of reconsolidation	Reactivation + betablocker (Brunet et al., 2008; Kindt et al., 2009)
	Reactivation + electroconvulsive shock (ECS) (Kroes et al., 2014)
	Reactivation + anesthetic (Vallejo et al., 2019)
Transcranial Magnetic Stimulation (TMS) during extinction	Deep Frontal TMS (Isserles et al., 2013)
	TMS of the prefrontal cortex during extinction (Raij et al., 2018)
Transcranial Direct-Current Stimulation (tDCS) during extinction	tDCS of the vmPFC during extinction (Dittert et al., 2018)

Several MMTs have been tested in clinical trials (Astill Wright et al., 2021), and it seems likely that MMTs could become available in the clinic soon. Among bioethicists, three general sets of concerns have been raised regarding MMTs: a) safety and social justice concerns, b) concerns about threats to authenticity and identity, and c) the possible legal and moral duties to retain certain memories. As such, MMTs have sparked a debate about ethical, legal and societal implications (Cabrera & Elger, 2016; Elsey & Kindt, 2016; Erler, 2011; Henry et al., 2007; Kass, 2003; Kroes & Liivoja, 2018; Lavazza, 2015; Liao & Sandberg, 2008; Liao & Wasserman, 2007; Parens, 2010). Yet, it is not clear to what extent there is currently a public demand for MMTs, or whether the resistance of bioethicists against manipulating memories is publicly shared. A previous study demonstrated a negative disposition in the general public towards prophylactic administration of memory dampening drugs after exposure to trauma, an approach that could be used to prevent PTSD (Newman et al., 2011). However, the authors suggested that this negative disposition may be directed at the prophylactic nature of the treatment instead at the MMT itself. People may be reluctant to undergo a preventive treatment for PTSD when they feel they are unlikely to develop the disorder (Newman et al., 2011). In this study, we focus on reconsolidation-based interventions that directly modify selective memories after they have been

rendered labile through reactivation (for a review, see Haubrich and Nader 2018), and probe attitudes towards the use of reconsolidation-based MMTs to treat PTSD.

If MMTs are to be successfully implemented as treatment for PTSD, it is important to not only understand how the public views MMTs, but also to understand the factors that may shape public attitudes towards MMTs. While it could be that the public supports an across-the-board use of MMTs, attitudes towards MMTs more likely depend on the precise context in which they are used and may vary between countries or different demographic groups. Additionally, if there is hesitancy regarding MMTs, the availability of scientifically justified information about PTSD and the specific therapeutic benefits of MMTs may modulate public attitudes towards them.

In this pre-registered study (https://osf.io/ztg7u), we varied the background information that participants received about MMTs and PTSD, and subsequently asked them to complete a survey to probe how morally acceptable they find MMTs. In addition, we asked participants to rate the moral acceptability of treatment with MMTs in specific scenarios. In these scenarios, we systematically varied four factors: 1) the professional background of the subject (varying from civilian to military, indicating the degree of *militarization* in the scenario) 2) the *agency* of the subject in the traumatic experience (varying from observer to intentional murder), 3) whether the subject develops PTSD or has an unpleasant but healthy response to the traumatic experience (indicating the degree of *medicalization*) and 4) whether there were any *stakeholders* that could benefit from memory retention (i.e., whether there may be a collective interest in a vivid recollection of this event or not, e.g. a moral duty for remembrance). We expected that more extensive information would render participants' judgment more positive towards MMTs, and that attitudes towards MMTs in all participants would depend on the specific situation in which they are used. In an attempt to identify predictors for more positive or negative attitudes, we furthermore asked participants to answer several question about safety beliefs, as well as the Moral Foundations Questionnaire (Graham et al., 2012) and a set of demographic questions.

Methods

Ethics statement

The study design and the materials used were reviewed and approved by the Ethics Committee of the Faculty of Social Sciences (ECSS) of Radboud University. All participants read and signed an informed consent form prior to participation, informing them of their right to opt out and withdraw their submission without penalties.

Participants

A total of 881 participants completed the survey via Prolific, a web-based recruitment platform for research (https://www.prolific.co/, Palan & Schitter, 2018). According to pre-registered criteria, submissions from 46 participants were rejected through Prolific based on failure to correctly answer more than one explicit attention check. In addition, 38 participants were excluded based on mismatches between demographic information entered on Prolific and in our survey, 61 for failing more than one comprehension check, 8 for spending less than 12 minutes on the survey and 57 for a median response time of less than 8 seconds in the sub-section of the study that presented scenarios. The final sample contained 716 participants.

Design

Participants were randomly assigned to a brief or extensive introduction about MMTs and PTSD. The brief introduction introduced MMTs as techniques that can enhance or weaken memories and specified that this survey exclusively discussed MMTs that weaken memories. The following description was given: "*MMTs can be drugs or newly developed forms of psychotherapy that change specific memories. Depending on the technique used, the targeted memory can become less vivid and have diminished factual content. Specifically, MMTs can be used to attenuate traumatic memories. As an example, through remembering, and thereby reactivating, a traumatic memory, followed by oral administration of a non-invasive drug, the traumatic quality of a specific memory may be removed. As a result, when remembered, the memory will feel less invasive, and no longer evoke an emotional response. Undergoing MMTs is safe and only affects the specific memories that are reactivated.*" At the end of the introduction we provided an example of an MMT-based treatment. In this example, the traumatic memory was modulated by means of reactivation of the memory followed by oral administration of propranolol. The extensive introduction contained the same information, but was preceded by a scientifically-informed description of traumatic memories, the prevalence and symptomatology of PTSD, currently available treatments and their limitations, and a basic description of the seminal work by Nader et al (2000), demonstrating that fear-conditioned rats no longer expressed fear after administration of a protein-synthesis inhibitor following memory reactivation. The complete introductions can be found on the OSF storage (https://osf.io/286yk/).

In line with a previous study by Newman et al. (2011), we presented participants with short scenarios. To probe what aspects of a situation influence attitudes towards MMTs, we used a 3 (Militarization: civilian, firefighter, military) x 3 (Agency: Observer, Accidental actor, Intentional actor) x 2 (Medicalization: no PTSD, PTSD) x 2 (Stakeholders: no collective interest, collective interest) within-subjects design, modelled after Young & Saxe, 2008. The scenarios describe John, in either a civilian, firefighter or military capacity (varying levels of militarization). In the scenarios, John either witnesses,

accidentally commits, or intentionally commits, a murder (varying levels of agency). In the non-PTSD variation, John is coping well but feels uncomfortable thinking about the event and feels that he would perform even better in his personal life and work without the memory of this event, while in the PTSD-variation, John develops PTSD: he experiences flashbacks of the event multiple times a day, feels stress, and experiences significant distress and impairment as a result of the disorder (varying levels medicalization). Finally, it is either stated that it is in no one's interest for John to retain a vivid recollection of the event, or that it may be important for our collective memory for him to retain a vivid recollection (stakeholder variation). The complete thirty-six scenario texts are available on the OSF storage (https://osf.io/xsjc5/).

Measured variables

Comprehension of introductory tekst

To verify comprehension of the introductory text, both groups were presented with five yes/no questions about the information provided: 'Do MMTs weaken memory in general?', 'Do MMTs target specific memories?', 'When we refer to MMTs in this survey, do we refer to techniques that enhance memories (i.e. increase vividness)?', 'Is oral administration of a drug that reduces the vividness of specific memories and example of MMTs?', 'Are MMTs dangerous?'.

General approval of MMTs

To measure the general attitude towards MMTs, participants were asked to respond to four questions on a 7-point anchored Likert-scale: 'How moral do you find the use of MTMs? (Completely immoral (1) – Completely moral (7))', 'To what extent do you agree with the following statement: There is nothing wrong with the use of MMTs (Completely agree (1) – Completely disagree (7))', 'How appealing do you find the use of MMTs? (Very unappealing (1) – Very Appealing (7))', 'To what extent do you agree with the following statement: Thinking about the use of MMTs makes me angry (Completely agree (1) – Completely disagree (7))'. Given that the interrelatedness of the four questions was acceptable (Cronbach's alpha of 0.745 in the brief introduction group and 0.751 in the extensive introduction group), a mean approval score was calculated by reversing the answer to the second item and subsequently calculating the mean of all statement answers for each participant.

Moral acceptability of scenarios

In response to each of the scenarios, participants responded on a 7-point Likert-scale to indicate their agreement to the following: 'I find the use of MMTs morally acceptable in this situation (Completely disagree (1) – Completely agree (7)).

Safety and efficacy beliefs

To measure to what extent participants believed MMTs are safe and effective, we asked participants to respond to four 7-point Likert-scale items anchored to 'Completely agree (1)' and 'Completely disagree' (7): 'MMTs may have serious side effects', 'MMTs are an effective treatment for PTSD', 'MMTs are safe' and 'Medical professionals in charge of MMTs would have the patient's best interest at heart'. Responses to these four items showed acceptable levels of interrelation (Cronbach's alpha of 0.759 brief introduction and 0.706 in the extensive introduction group) to combine the response to the first item and the reversed response to the last three items into a mean safety score. In addition, participants responded to two items regarding PTSD and treatment with MMTs: 'PTSD strongly decreases the quality of life' and 'If I had PTSD, I would undergo MMTs'. These are not included in mean scores and treated as separate measures.

Moral foundations questionnaire

Participants completed the 30-item Moral Foundations Questionnaire (MFQ) (Graham et al., 2012), which measures the extent to which they rely on each of five moral foundations: care/harm (related to the ability to feel and dislike the pain of others), fairness/cheating (related to reciprocal altruism), loyalty/betrayal (related to the ability to form coalitions and value self-sacrifice for a group), authority/subversion (shaped by our long history of hierarchical social interactions, legitimizes authority and respect for traditions), and purity/sanctity (shaped by the psychology of disgust and contamination, underlies the religious notions of striving to live in a more noble way). The MFQ has been shown to be a reliable instrument and to predict a variety of moral and political attitudes, while being independent of political ideology (Graham et al., 2012).

Demographics

Demographic information was collected for all participants, including age, gender, ethnicity, educational level, employment status, household income, marital status, parental status, political ideology, country of residence, religion, (close) personal experience with trauma and (close) personal experience working in the military.

Data analysis

Cluster analysis

Scores from the MFQ were submitted to k-means clustering using Python (Jupyter 3.0). Using the elbow method for visual inspection of the average of the squared distances from the cluster centers of the respective clusters, an optimal number of 2 clusters was determined.

Stepwise linear regression on demographic factors

To test whether demographic information predicted attitudes towards MMTs, we built a stepwise regression model with age, gender, ethnicity, education level, employment status, income, marital

status, parental status, political identity, country of residence, religious affiliation, personal experience with traumatic events and personal experience with working in the military. Individual response categories that contained less than 5% of the responses were discarded or merged if appropriate (e.g. in case of adjacent income ranges). Please refer to the supplementary information for the included response options and baseline response for each factor.

Hypothesis testing

Statistical analyses (ANOVAs, t-tests and regression analysis) were performed in SPSS (IBM SPSS Statistics Inc.). Dependent measures were submitted to repeated measures ANOVAs, paired-samples t-tests or independent-samples t-tests and statistics were Greenhouse–Geisser or Huyn-Feldt corrected for non-sphericity when appropriate (i.e., if sphericity assumptions were violated and epsilon was smaller or greater than 0.75, respectively). Significant findings from ANOVAs were followed-up by paired- and independent samples t-tests. We report partial eta-squared as a measure of effect size. Means ± standard error of the mean (s.e.m.) are provided where relevant unless otherwise indicated.

Deviations from the pre-registration

We did not deviate from the pre-registration in any meaningful way. While we stated that we would test for differences between MFQ clusters with an ANOVA, we used a t-test instead because there were only 2 clusters. In addition to the pre-registered analysis, we explored correlations between mean attitudes and safety beliefs, and mean attitudes and MFQ foundations. To explore potential associations between attitudes towards MMTs in specific scenarios and moral foundations, we added the moral foundation scores as covariates in a separate analysis of the scenarios. To assess to what extent demographic variables, safety beliefs and moral foundations explain unique variance, we explored whether stepwise regression retains all previously identified predictors in a final model.

To verify that participants who read the extensive introduction retained additional information compared to participants who read the brief introduction, we carried out a follow-up study in another sample. In this follow-up study, we repeated the introduction, comprehension checks and general attitude questions, and finally added 16 questions about the information in the introduction. This follow-up study was pre-registered separately (https://osf.io/e98c7).

Results

The final sample contained 716 participants (366 male, 339 female, 5 other gender, mean age 30±0.40), of which 254 resided in Mexico, 241 in the USA and 221 in the Netherlands (NL). Participants assigned to the two information conditions were equally able to pass the five comprehension checks about the definitions of MMTs used in this study ($t(74)=-1.075$, $p=0.284$, Brief Introduction: 4.78±0.02,

Extensive Introduction: 4.81±0.02). The two groups did not differ significantly on any of the measured demographic variables (all p's>0.07 before correction for multiple comparisons, see Supplementary Table 5.5).

Background information does not modulate general attitudes towards MMTs

The overall attitude towards MMTs was somewhat positive, indicated by an overall mean attitude of 5.25 out of 7 (that was significantly more positive than a neutral answer of 4, t(715)=29.742, p<0.001), and on average, participants indicating that they 'somewhat agreed' that they would undergo MMTs if they suffered from PTSD. We expected participants who received more extensive information to have more positive attitudes towards MMTs compared to participants that read the brief introduction. However, the mean attitude was comparable between groups (t(714)=0.050, p=0.960, Brief Introduction: 5.25±0.06, Extensive Introduction: 5.25±0.06, Figure 5.1A). While we expected the extensive introduction to increase beliefs that MMTs are safe and effective, we also did not find evidence for different safety beliefs between groups (t(714)=-0.376, p=0.707, Brief Introduction: 5.09±0.05, Extensive Introduction: 5.12±0.05, Figure 5.1B). We further did not find any effect of the information on the reported likelihood of undergoing MMTs when suffering from PTSD (t(714)=0.030, p=0.976, Brief Introduction: 5.43±0.07, Extensive Introduction: 5.43±0.07, Figure 5.1C) or on the belief that PTSD decreases quality of life (t(714)=0.465, p=0.642, Brief Introduction: 6.18±0.08, Extensive Introduction: 6.13±0.07, Figure 5.1D).

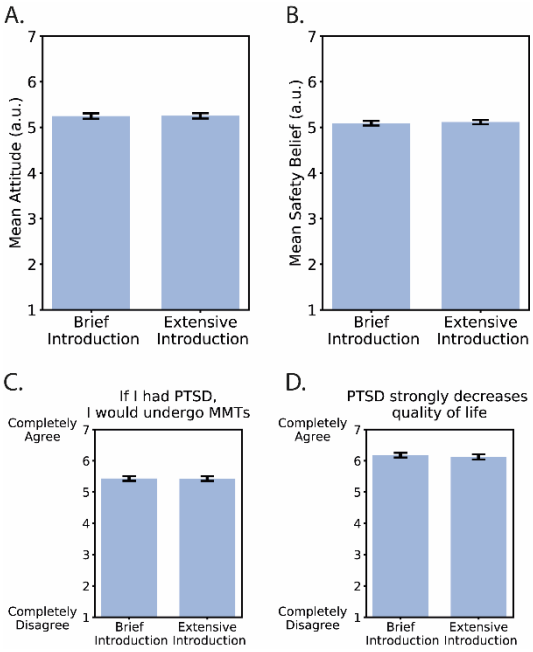


Figure 5.1. General attitudes towards MMTS did not differ in participants that read a brief or extensive introduction. *Mean attitudes towards MMTs (A) and safety beliefs (B). The most positive approval score was 7, while the most negative possible score was 1. (C) Reported likelihood of undergoing MMT treatment for PTSD. (D) Reported beliefs about the impact of PTSD on quality of life. Separated bars are displayed for the Brief Introduction and Extensive Introduction groups. Error bars represent ± standard error of the mean (S.E.M.).*

As we did not find any effect of the additional information on any of the attitudes towards MMTs or beliefs regarding MMTS and PTSD, we ran a follow-up study to investigate whether participants who read the extensive introduction were actually able to reproduce more of the additional information

presented in the extensive introduction compared to participants who read the brief introduction. In this follow-up study with 106 participants, participants who read the extensive introduction answered significantly more questions about the content of the introduction correctly than participants who read the brief introduction ($t(104)=7.511$, $p<0.001$, Brief Introduction: 12.0±0.23, Extensive Introduction: 14.5±0.24). Mean attitudes towards MMTs did not differ between the brief and extensive introduction groups ($t(104)=0.600$, $p=0.550$) and did not correlate with the number of questions answered correctly ($r(104)=0.078$, $p=0.426$). Moreover, mean attitude scores in the follow-up sample were comparable to mean attitude scores in the main study ($t(820)=0.039$, $p=0.968$). Thus, the follow-up study demonstrated that participants who read the extensive introduction did retain additional information compared to participants who read the brief introduction, confirming our initial finding that the additional information contained in the extensive introduction did not improve attitudes towards MMTs.

Attitudes towards MMTs are associated with safety beliefs

We expected that presenting participants with scientifically grounded information about MMTs would improve attitudes towards MMTs by reducing potential safety concerns. Although we did not find any effect of the information that was presented to participants on their safety perceptions or attitudes towards MMTs we nevertheless explored whether there was an association between safety beliefs and MMT approval. Mean MMT approval scores were positively correlated with safety beliefs ($r(713)=0.530$, $p<0.001$, partial correlation controlling for brief/extensive introduction, Figure 5.2). Thus, while we did not successfully modulate safety beliefs, safety beliefs are positively associated with MMT approval.

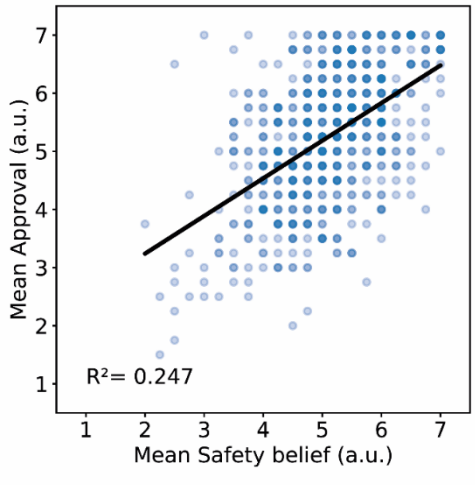


Figure 5.2. Mean attitudes towards MMTs correlate positively with safety beliefs. *The most positive approval score was 7, while the most negative possible score was 1. Individual data points are displayed in light blue. Darker hues of blue indicate increasing numbers of overlapping datapoints*

Attitudes towards MMTs differ between countries

Mean attitudes towards MMTs differed between countries ($F_{(2,713)}$=7.030, p=0.001, Figure 5.3A). Specifically, attitudes were more positive in Mexico compared to the USA (t(493)=3.230, p=0.001, Mexico: 5.46±0.07, USA: 5.13±0.08) and the Netherlands (t(473)=3.257, p=0.002, the Netherlands: 5.14±0.07), while they did not differ between the USA and the Netherlands (t(460)=0.131, p=0.896). Reported willingness to undergo MMT treatment for PTSD also differed between countries ($F_{(2,713)}$=14.576, p<0.001, Figure 5.3B). Willingness was lower in the USA compared to Mexico (t(493)=-5.152, p<0.001, Mexico: 5.71±0.08, USA: 5.05±0.1) and the Netherlands (t(460)=-3.500, p<0.001, the Netherlands: 5.52±0.09), whereas there was no difference between Mexico and the Netherlands (t(473)=1.625, p=0.105).

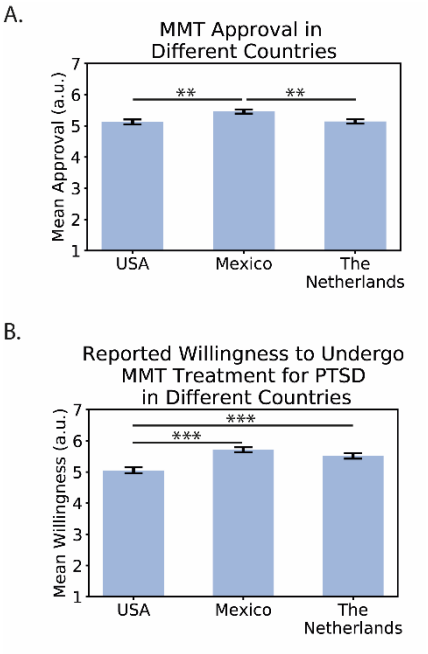


Figure 5.3. Attitudes towards MMTs differ between countries. *A. Mean MMT approval ratings in the USA, Mexico and the Netherlands. B. Reported willingness to undergo MMT-based treatment when suffering from PTSD in the USA, Mexico and the Netherlands. The most positive approval score was 7, while the most negative possible score was 1. Error bars represent ± S.E.M., ** p<0.01, *** p<0.001*

Attitudes towards MMTs are context-dependent

We hypothesized that attitudes towards MMTs would differ depending on the specific scenario in which MMTs were applied. Attitudes towards MMTs were influenced by the combination of all four scenario variations (Figure 5.4, Agency x Medicalization x Stakeholder x Militarization Interaction effect: F(4,2772)=4.589, p<0.001, η^2=0.007). Thus, acceptability of MMTs in specific scenarios depended on the unique combination of the degree of militarization, agency, medicalization and the presence or absence of external stakeholders.

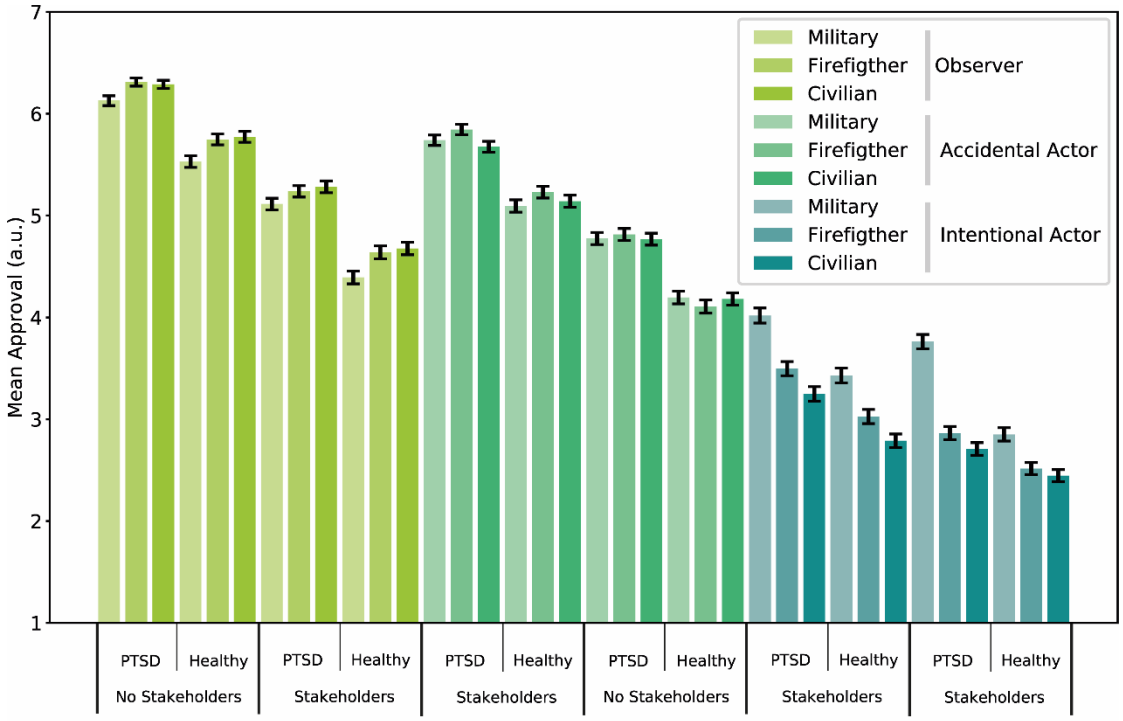


Figure 5.4. Mean approval of MMTs differed between scenarios with varying levels of militarization, agency, medicalization, and stakeholders. *Each bar represents mean responses to a unique scenario across participants. A value of 7 indicates that the participant found the use of MMTs in this scenario completely acceptable, while a value of 1 indicates that the participant found the use of MMTs completely unacceptable. Responses are averaged across country of residence and information group. Error bars represent ± S.E.M.*

Attitudes towards the acceptability of MMTs were most strongly modulated by the agency of the actor in the scenario (main effect of Agency, $F(_{2,988})=164.151$, $p<0.001$, $\eta^2=0.191$, Figure 5.5A). Specifically, attitudes were less positive for intentional agents as opposed to accidental agents ($t(716)=29.285$, $p<0.001$) and passive observers ($t(716)=41.329$, $p<0.001$), and less positive for accidental agents compared to passive observers ($t(716)=15.856$, $p<0.001$). Participants were more positive towards application of MMT-based treatment when the agent of the scenario suffered from PTSD compared to someone who was undergoing an unpleasant but healthy stress response (Main effect Medicalization, $F(_{1,693})=92.905$, $p<0.001$, $\eta^2=0.030$, Figure 5.5B). Attitudes were also clearly influenced by whether or not there was a collective interest in retaining the traumatic memory (main effect of stakeholder ($F(_{1,693})=112.120$, $p<0.001$, $\eta^2=0.139$, Figure 5.5C), as shown by more negative attitudes when there was a collective interest in retaining the memory.

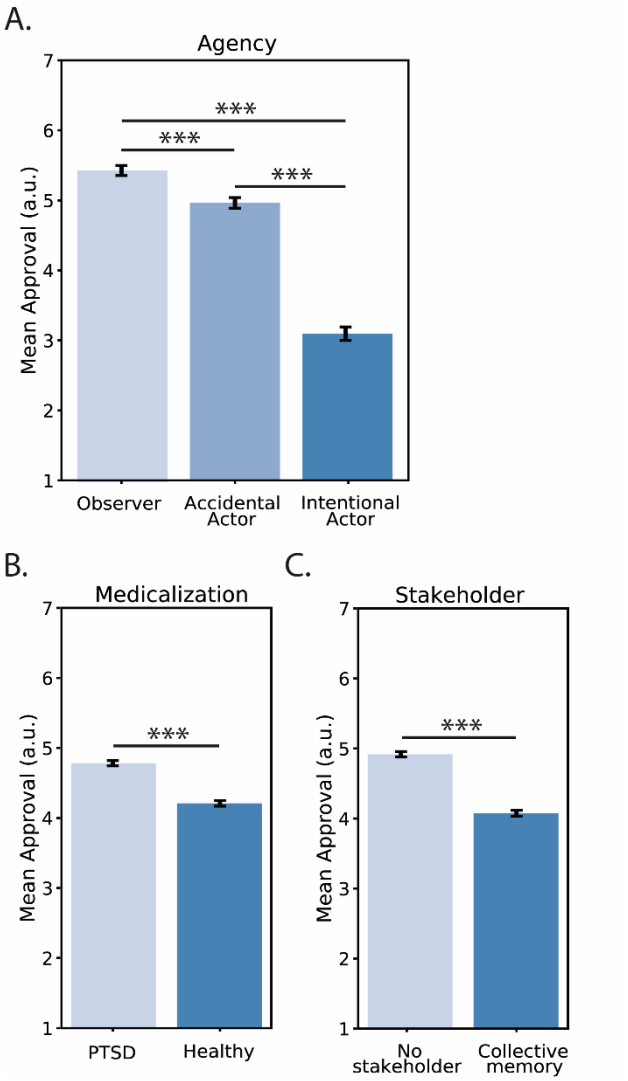


Figure 5.5. Attitudes towards MMTs depend on the degree of agency during a traumatic experience, medical diagnosis and the presence of an explicit stakeholder that would benefit from memory retention. *A. Attitudes towards MMT-based treatment were less positive when the to-be-treated individual had increased agency in the traumatic experience. Attitudes were most positive for observers of a traumatic event (murder), less positive for those accidentally contributing to a traumatic event and the least positive for those that intentionally were involved in the event. B. Attitudes towards MMT-based treatment were more positive in scenarios that described treatment of a person with PTSD compared to a healthy person. C. Attitudes towards MMT-based treatment were more positive when it was explicitly mentioned that there was no collective interest in retaining a memory as compared to when there was a collective interest in remembrance. A value of 7 indicates that the participant found the use of MMTs in this scenario completely acceptable, while a value of 1 indicates that the participant found the use of MMTs completely unacceptable. Error bars represent ± S.E.M, *** p<0.001*

Attitudes towards MMT-based treatment in scenarios with different degrees of agency differed depending on whether an explicit stakeholder was mentioned who would benefit from retention of the memory (Agency x Stakeholder interaction effect $F_{(1.597,1106.844)}=35.148$, $p<0.001$, $\eta^2=0.048$, Figure 5.6A). Mean approval was lower at each level of agency when there was an explicit collective interest in retaining the memory compared to when there was no benefit in retaining the memory, and the effect of agency was stronger in the absence of a stakeholder ($F_{(2,1430)}=1442.517$, $p<0.001$, $\eta^2=0.669$) than in the presence of a stakeholder ($F_{(2,1430)}=1058.611$, $p<0.001$, $\eta^2=0.597$).

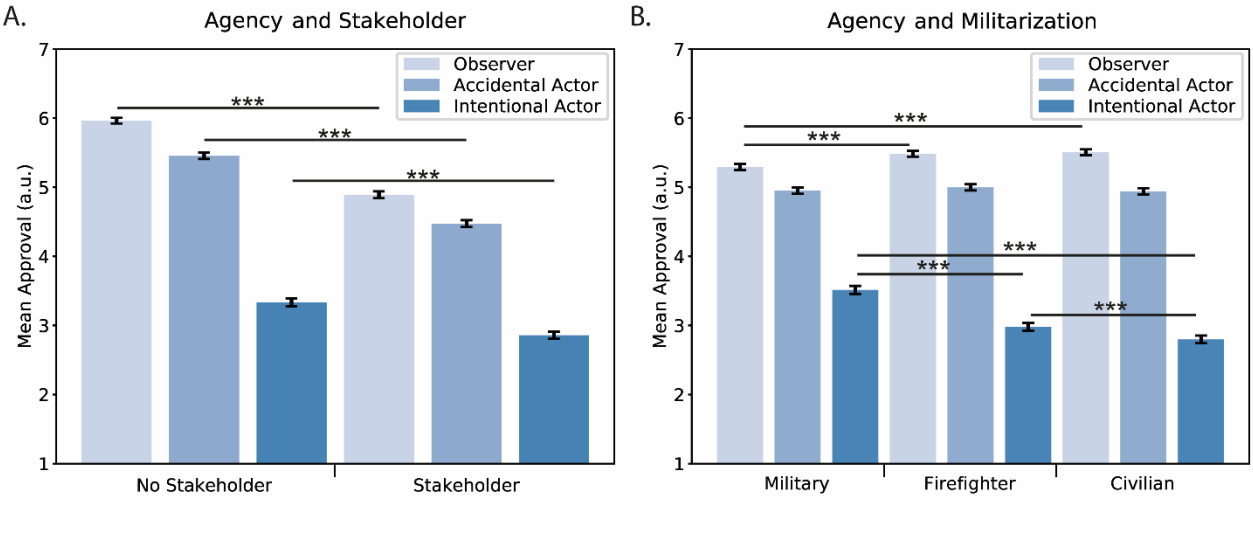


Figure 5.6. Mean approval of MMTs in scenarios with different levels of agency depends on the presence of external stakeholders and militarization. A*. Attitudes towards MMTs differed in scenarios with varying levels of agency depending on whether or not there was a collective stakeholder that benefited from memory retention.* ***B.*** *Attitudes towards MMTs-based treatment for people that acted with varying levels of agency during a traumatic event depended on whether they were on duty as a soldier, firefighter or civilian. A value of 7 indicates that the participant found the use of MMTs in this scenario completely acceptable, while a value of 1 indicates that the participant found the use of MMTs completely unacceptable. Error bars represent ± S.E.M, *** p<0.001*

Attitudes towards the acceptability of MMTs for subjects with a professional background in the military, as a firefighter or a civilian interacted with their degree of agency in the described scenario (Agency x Militarization interaction effect, $F_{(3.156,2169.896)}=17.201$, $p<0.001$, $\eta^2=0.024$, Figure 5.6B). Attitudes were less positive for MMT treatment of military professionals who observed a traumatic event compared to firefighters ($t(715)=-6.635$, $p<0.001$, military: 5.30±0.04, firefighter: 5.49±0.04) and civilians ($t(715)=-7.119$, $p<0.001$, civilians: 5.51±0.04). For intentional killing, attitudes towards MMTs were more positive for military personnel compared to firefighters ($t(715)=13.001$, $p<0.001$, military: 3.52±0.06, firefighter: 2.98±0.06) and civilians ($t(715)=16.372$, $p<0.001$, civilians: 2.81±0.06) and for firefighters compared to civilians ($t(715)=6.507$, $p<0.001$). For accidental acts, attitudes were comparable for all professional backgrounds (all p's>0.1). Attitudes towards the acceptability of MMTs in specific scenarios further showed an interaction between aspects of the scenario, country, and information group (see Supplementary information for a full description).

Attitudes towards MMTs do not differ between clusters based on moral intuitions

We expected that attitudes towards MMTs may be shaped by moral intuitions. Using K-means clustering for 2 clusters based on MFQ scores, we obtained two clusters that differed mainly on the loyalty/betrayal, authority/subversion and sanctity/degradation indices (Table 5.2). To verify that the two clusters based on MFQ scores were distinct sub-groups, we examined demographic characteristics for both groups. Comparing demographic information indicated significant differences between clusters (see Table 5.2). For example, cluster 1 contained a higher percentage of right-

wing/conservative (t(688)=6.104, p<0.001, cluster 1: 15%, cluster 2: 3%) and Christian (t(688)=8.338, p<0.001, cluster 1: 52%, cluster 2: 23%) participants, while cluster 2 contained more left-wing/liberal participants (t(688)=7.349, p<0.001, cluster 1: 33%, cluster 2: 60%, Table 5.2). However, the two groups did not differ significantly on general attitudes towards MMTs (p=0.173) or reported likelihood of undergoing MMT-based treatment (p=0.318). Thus, we did not find different attitudes towards MMTs in clusters of participants with different moral intuitions.

Table 5.2. Descriptive statistics for the two clusters based on MFQ scores. *Values indicated mean ± S.E.M.*

	B (95% confidence interval)	**SE B**	**β**	**t**	***p***
Step 1 (R^2=0.019)					
Constant	5.134 (5.032, 5.236)	0.052		99.022	p<0.001
Country=Mexico	0.326 (0.156, 0.497)	0.087	0.139	3.750	p<0.001
Step 2 (R^2=0.031)					
Constant	0.089 (4.983, 5.194)	0.054		94,654	p<0.001
Country=Mexico	0.372 (0.199, 0,544)	0.088	0.158	4,230	p<0.001
Black/African Ethnicity	0.567 (0.194, 0,940)	0.190	0.112	2,982	p=0.003
Step 3 (R^2=0.041)					
Constant	5.167 (5.047, 5.287)	0.061		84,488	p<0.001
Country=Mexico	0.403 (0.230, 0.577)	0.088	0.172	4,564	p<0.001
Black/African Ethnicity	0.592 (0.220, 0.963)	0.189	0.117	3,123	p=0.002
Trauma in family	-0.224 (-0.391, -0.058)	0.085	-0.098	-2,641	p=0.008
Step 4 (R^2=0.047)					
Constant	5.226 (5.094, 5.358)	0.067		77,656	p<0.001
Country=Mexico	0.390 (0.217, 0.564)	0.088	0.166	4,417	p<0.001
Black/African Ethnicity	0.568 (0.196,0.939)	0.189	0.112	2,998	p=0.003
Trauma in family	-0.231 (-0.398, -0.065)	0.085	-0.101	-2,729	p=0.007
Full-time student	-0.198 (-0.384, -0.012)	0.095	-0.077	-2,089	p=0.037

The two clusters of participants showed significantly different scores on all five MFQ foundations (Table 5.2), with most prominent differences in loyalty/betrayal, authority/subversion and sanctity/degradation. To explore whether individual indices of the MFQ could be associated with attitudes towards MMTs, we correlated the individual indices with general attitudes towards MMTs (Figure 5.7). While care/harm (r(690)=0.149, p<0.001, Figure 5.7A), fairness/cheating (r(690)=0.147, p<0.001, Figure 5.7B) and authority/subversion (r(690)=0.088, p=0.020, Figure 5.7C) showed a minimal but positive correlation with mean attitudes, there was no correlation between loyalty/betrayal (p=0.133) or purity/sanctity (p=0.829) and attitudes towards MMTs.

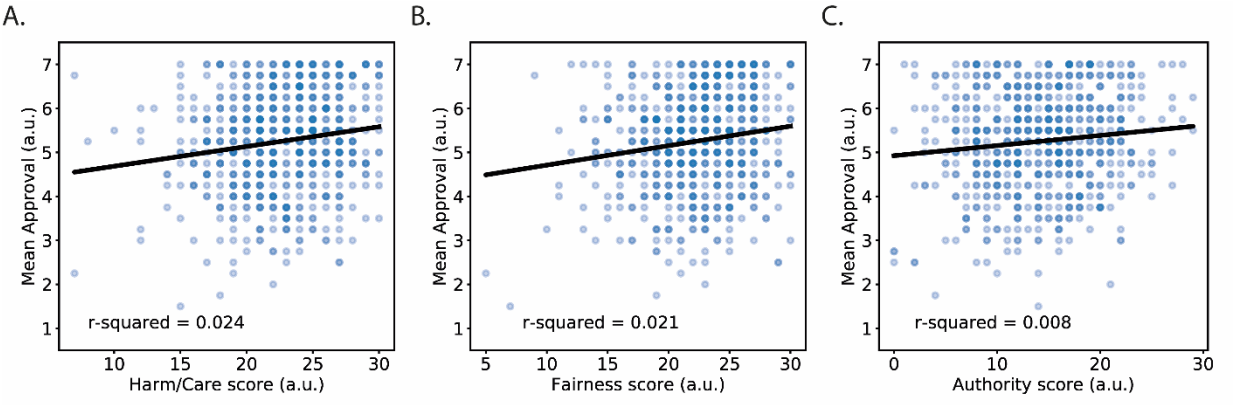


Figure 5.7. Mean attitudes towards MMTs positively correlated with moral foundations. *Individual data points are displayed in light blue. Darker hues of blue indicate increasing numbers of overlapping datapoints*

Similarly, we explored whether the individual indices of the MFQ were associated with responses to the distinct scenarios. We added the five MFQ indices as covariates to a 3 (Militarization) x 3 (Agency) x 2 (Medicalization) x 2 (Stakeholders) rmANOVA containing all 36 scenarios. To determine the direction of any effects, we compared the two groups of participants with the 20% highest and lowest scores on the relevant foundation, and visually inspected their responses at different levels of the interacting factor. Responses to scenarios with different degrees of agency varied between participants with different scores on the care/harm ($F_{(2,1342)=}$79.285, $p<0.001$, $\eta^2= 0.013$) and fairness/cheating ($F_{(2,1342)}$=46.796, p=0.005, η^2=0.008) foundations. Participants with high scores on the care/harm foundation and/or the fairness/cheating foundation were more willing to treat cases with lower levels of agency (observers and accidental actors). Responses to scenario with different levels of medicalization (PTSD vs. a healthy stress response) varied along the loyalty/betrayal ($F_{(1,667)}$=26.130, p=0.016, η^2=0.009), authority/subversion ($F_{(1,668)}$=6.860, p=0.009, η^2=0.010) and purity/sanctity ($F_{(1,664)}$=7.580, p=0.006, η^2=0.011) foundations. Participants who scored high on these foundations were more willing to accept MMTs in scenarios that described a healthy stress response. The loyalty/betrayal foundation was further associated with responses to scenarios describing agents employed as soldiers, firefighters or civilians (militarization, $F_{(2,1342)}$=4.236, p=0.015, η^2=0.006), where participants scoring higher on the loyalty foundation were more accepting of MMTs as treatment for soldiers and firefighters.

Demographic information minimally predicts attitudes towards MMTs

We expected that attitudes towards MMTs would differ between respondents in different demographic groups. The results of our stepwise regression model suggested that country of residence, ethnicity, personal experience with trauma and student status play a role in understanding attitudes towards MMTs (Table 5.3). Specifically, residing in Mexico and being of Black/African ethnicity were associated with more positive attitudes towards MMTs, while being a full-time student and, interestingly, having a family member that experienced a traumatic event were associated with

less positive attitudes towards MMTs. Age, gender, level of education, political affiliation, religious affiliation, marital status, being a parent and having experience with military employment were not associated with attitudes towards MMTs.

Table 5.3. Linear model of demographic predictors of mean attitudes towards MMTs. *Confidence intervals and standard errors are based on 1000 bootstrap samples.*

	B (95% confidence interval)	**SE B**	**β**	**t**	***p***
Step 1 (R^2=0.019)					
Constant	5.134 (5.032, 5.236)	0.052		99.022	p<0.001
Country=Mexico	0.326 (0.156, 0.497)	0.087	0.139	3.750	p<0.001
Step 2 (R^2=0.031)					
Constant	0.089 (4.983, 5.194)	0.054		94.654	p<0.001
Country=Mexico	0.372 (0.199, 0,544)	0.088	0.158	4.230	p<0.001
Black/African Ethnicity	0.567 (0.194, 0,940)	0.190	0.112	2.982	p=0.003
Step 3 (R^2=0.041)					
Constant	5.167 (5.047, 5.287)	0.061		84.488	p<0.001
Country=Mexico	0.403 (0.230, 0.577)	0.088	0.172	4.564	p<0.001
Black/African Ethnicity	0.592 (0.220, 0.963)	0.189	0.117	3.123	p=0.002
Trauma in family	-0.224 (-0.391, -0.058)	0.085	-0.098	-2.641	p=0.008
Step 4 (R^2=0.047)					
Constant	5.226 (5.094, 5.358)	0.067		77.656	p<0.001
Country=Mexico	0.390 (0.217, 0.564)	0.088	0.166	4.417	p<0.001
Black/African Ethnicity	0.568 (0.196,0.939)	0.189	0.112	2.998	p=0.003
Trauma in family	-0.231 (-0.398, -0.065)	0.085	-0.101	-2.729	p=0.007
Full-time student	-0.198 (-0.384, -0.012)	0.095	-0.077	-2.089	p=0.037

Since safety beliefs and three of the moral foundations (care/harm, fairness/reciprocity, and authority/subversion) were positively associated with attitudes towards MMTs, we next explored to what extent these factors could improve predictions of attitudes to MMTs. Therefore, we added these factors and the demographic predictors to a stepwise linear regression model. The final model contained the same demographic predictors as identified previously, as well as mean safety beliefs and fairness/reciprocity indices as positive predictors (See Table 5.4). Since adding care/harm and authority/subversion indices as predictors did not significantly improve the model, they were not retained in the final model. Adding safety beliefs and fairness/reciprocity greatly improved the model (R^2=0.577, demographics only: R^2=0.047). Thus, while we found several demographic factors that were predictive of attitudes towards MMTs, their relative contribution was minimal compared to the combined predictive value of safety beliefs and fairness/reciprocity.

Table 5.4. Linear model of predictors of mean attitudes towards MMTs. *Confidence intervals and standard errors are based on 1000 bootstrap samples.*

	B (95% confidence interval)	SE B	β	t	*p*
Final Model (R^2=0.577)					
Constant	1.410 (0.874, 1.947)	0.273		5.160	p<0.001
Mean safety belief	0.639 (0.564, 0.715)	0.038	0.522	16.643	p<0.001
Country=Mexico	0.326 (0.135, 478)	0.078	0.138	4.199	p<0.001
Black/African Ethnicity	0.454 (0.135, 0.774)	0.163	0.089	2.790	p=0.005
Trauma in family	-0.238 (-0.381, -0.095)	0.073	-0.103	-3.267	p=0.001
Full-time student	-0.251 (-0.410, -0.0910	0.081	-0.097	-3.083	p=0.002
Fairness/Reciprocity	0.027 (0.009, 0.045)	0.009	0.095	2.955	p=0.003

Discussion

In this study, we set out to explore attitudes towards MMTs in a sample of the public in Mexico, the USA and the Netherlands. Overall, attitudes were somewhat positive, and varied between countries. Contrary to our hypothesis, providing extended information did not alter attitudes towards MMTs or safety beliefs, nor did it increase the belief that PTSD decreases quality of life, nor did it change the reported likelihood of undergoing MMT-based treatment for PTSD. Attitudes towards MMTs strongly depended on the specific scenario in which MMTs were applied and showed minor associations with moral intuitions and demographic variables.

We expected that highlighting the symptoms of PTSD, shortcomings of current treatments, and specific advantages of MMTs, would positively modulate safety beliefs and increase the perceived severity of PTSD, improving attitudes towards MMTs. However, safety beliefs and perceived severity of PTSD were comparable between information groups and somewhat positive across both groups. This could indicate that the public is positively disposed towards MMTs and their attitude is not strongly dependent on details of the technique itself. However, it may also be that we reached a ceiling effect in the brief introduction by referring to PTSD and stating that MMTs are safe, after which the extensive introduction did not have any additional effect. This is supported by the finding that general attitudes towards MMTs were somewhat positive, independent of the extensiveness of the information that was given. Thus, while extensive background information did not have any additional influence on MMTs compared to a brief introduction, both introductions could have positively and equally modulated attitudes towards MMTs.

While approval rates for MMTs in the most accepted were high (as indicated by a mean acceptance score above 6, out of 7), responses to all scenarios showed considerable variation and were affected by the unique combination of military background, personal agency in the traumatic experience, degree of medicalization and the presence of external stakeholders. We did not find support for an across the board acceptance of MMTs, because mean moral acceptability of MMTs decreased to "somewhat unacceptable" in the least accepted scenarios. Specifically, attitudes were more positive when the agency of the described subject was lower, when applied for treatment of PTSD as compared to use in subjects who remained healthy, and when there were no stakeholders that benefited from retaining the memory. An interaction effect of militarization and agency further showed that while acceptance was comparable for accidental actors across professional backgrounds, MMT-based treatment was more acceptable for military personnel that intentionally killed, and less acceptable for observers of lethal situations, compared to firefighters or civilians. Potentially, these low acceptance rates reflect participants making judgements of fault/blame and withholding MMTs out of retribution (Cabrera & Elger, 2016; Gerber & Jackson, 2013). As such, the effects of agency on moral acceptability of MMTs might be similar for interventions that have no relation with memory at all, e.g. medical treatment for physical injuries might also be rated less acceptable for murderers compared to those that observed murder (Nagelsen & Huckelbury, 1969).

Newman et al. (2011) previously investigated attitudes towards prophylactic use of memory dampening drugs after a traumatic experience to prevent development of PTSD, and suggested that people are negatively disposed to memory dampening drugs (Newman et al., 2011). As the authors mentioned, interventions that 'repair' and move performance towards normative functioning are viewed as more acceptable than interventions that enhance capacities (Cabrera et al., 2015), which is in line with our current finding that attitudes towards the use of MMTs for treatment of PTSD are more positive than attitudes towards MMT-based treatment in healthy subjects.

While the MFQ revealed two distinct clusters of participants, the two groups did not differ on mean attitudes towards MMTs and did not report different likelihoods of undergoing MMTs when suffering from PTSD. Nevertheless, there was a weak association between individual foundations as measured by MFQ and general attitudes to MMTs. Furthermore, the moral foundations were associated with distinct attitudes towards MMTs in specific scenarios, suggesting that moral intuitions do shape attitudes towards MMTs, while these may not be detectable in more general clusters. MFQs may become more relevant in situations that include a wider variety of morally questionable aspects, including for example a soldier that intentionally kills an explicitly innocent civilian to cover up war crimes.

Using stepwise linear regression, we identified four demographic variables that were predictive of attitudes towards MMTs: country of residence, ethnicity, employment status and personal experience with traumatic events. Participants in Mexico reported more positive attitudes towards MMTs. Previous work on the use of memory dampening drugs to prevent PTSD similarly reported that attitudes differed between countries (Newman et al., 2011). Here, we sampled two western and one southern country, and found that while participants in the USA and the Netherlands responded similarly, Mexican participants generally displayed more positive attitudes. Differences between moral attitudes in southern and western countries have been reported previously (Awad et al., 2018) and may be related to differences in values between individualistic cultures (USA, NL) and collectivistic cultures (Mexico, Hofstede, 2008). Participants of Black/African ethnicity reported more positive attitudes towards MMTs compared to participants of other ethnicities. Given that African-American race-ethnicity has previously been identified as a predictor of greater reported willingness to seek treatment for mental health disorders (Shim et al., 2009), it could be that more positive attitudes in participants of Black/African ethnicity are mediated by a broader increased acceptance of psychiatric treatment. Being a full-time student or having a family member who experienced a traumatic event were associated with a more negative attitude towards MMTs. While we may speculate that full-time students could be generally more sceptic towards novel technology, and participants that witnessed a family member undergoing trauma-therapy may be more careful in selecting treatment options, it is hard to yield a meaningful interpretation of the identified demographic predictors in the current study. Future research could employ focus group interviews to give more insight in the structure of public moral reasoning on MMTs and may identify specific patterns of reasoning in these demographic sub-groups. However, compared to a regression model that included safety beliefs, the explanatory value of these demographic variables was minimal. Nevertheless, it appears that attitudes towards MMTs may vary between demographic groups, which could potentially lead to varying treatment uptake when MMT-based treatments become available.

Several aspects of our study design likely influenced the measured attitudes towards MMTs and should be noted as limitations. First, emphasizing the negative effect of PTSD on quality of life may have biased attitudes, towards the positive side. By selectively highlighting the potential burden of PTSD without discussing potential undesirable effects of MMTs, we may have influenced participants to develop attitudes only based on the emotional pain evoked by PTSD. Second, as we were interested in fundamental attitudes towards the idea that novel biomedical techniques could be used to modify memories, we emphasized that the treatment is safe and does not have any side-effects. It may be hard for participants to think of indirect consequences of treatments with MMTs, e.g. personality changes or shifts in identity, especially given that the novel treatment may have remained abstract,

and we emphasized that there would not be any side-effects. Third, the scenarios we presented were not set up to evoke strong moral convictions. Although we included scenarios that described intentional murder, none of the scenarios described subjects who e.g., cheated the system or escaped justice. Fourth, we explicitly discussed application of MMTs to a single, burdensome memory, potentially limiting any concerns about MMTs as a relatively minor threat to authenticity (low-stake, high gain). In a future study, instead of varying the factors implemented here, it would be worthwhile to investigate whether providing information from a completely different angle, e.g. not mentioning PTSD and compromised quality of life, would result in comparable attitudes towards MMTs.

In the current study, we did not find evidence for a strong discomfort around MMTs. While extensive background information on the scientific foundation of MMTs did not modulate attitudes towards MMTs compared to a brief introduction, the general attitudes towards MMTs were somewhat positive. Attitudes towards MMTs were most clearly associated with safety beliefs and varied strongly depending on the scenario in which they would be used. In other words, the current sample seems mostly concerned about practical aspects of MMTs (safety and situation) rather than expressing a general discomfort around modifying memories or fundamental objections based on authenticity as raised before by bioethicists (Cabrera & Elger, 2016; Erler, 2011; Henry et al., 2007; Hui & Fisher, 2015; Lavazza, 2015; Liao & Sandberg, 2008; Liao & Wasserman, 2007; Parens, 2010a). As such, when MMT-based treatments become available for PTSD in the near future, and their safety has been adequately demonstrated, it seems likely that these treatments for cases such as studied here will be well-received by the general public.

Supplementary information – Chapter 5

Demographic information

Supplementary Table 5.5. Descriptive statistics for the Brief and Extensive introduction groups. *Note that response options in some categories add up to more than 100% as response options were not exclusive and participants were asked to select all that applied.*

Variable		Brief introduction Group	Extensive introduction Group
Sample size (*n*)		359	357
Mean Age ± S.E.M		30.41±0.56	30.43±0.56
Gender	Male (%)	52%	50%
	Female (%)	47%	48%
	Other (%)	0%	1%
Ethnicity	Caucasian	49%	50%
	Hispanic/Latinx	37%	38%
	Asian	9%	9%
	Black/African	5%	5%
	Other	0%	0%
Education level	High school GED	13%	15%
	College (no degree)	26%	23%
	Associate degree	4%	6%
	Bachelor's degree	40%	41%
	Master's degree	12%	10%
	Other	5%	5%
Employment status (non-exclusive, select all that apply)	Full-time job	41%	40%
	Part-time job	25%	24%
	Unemployed	13%	12%
	Full-time student	26%	25%
	Part-time student	7%	6%
	Homemaker	5%	6%
	other	4%	3%
Annual Income	$0-$10.000	13%	16%
	$10.001-$25.000	21%	21%
	$25.001-$50.000	24%	22%
	$50.001-$75.000	17%	17%
	$75.001-$100.000	10%	9%
	$100.001-$150.000	6%	6%
	$150.000+	4%	4%
Marital status	Single, never married	63%	61%
	Married or domestic partnership	34%	32%
	Separated	1%	2%
	other	2%	5%
Parental status	Parent	21%	23%
	Not a parent	79%	77%
Political identification	Right-wing/Conversative	11%	6%
	Moderate	39%	35%
	Left-wing/Liberal	45%	51%
Country of residence	USA	35%	32%
	Mexico	35%	36%
	The Netherlands	30%	32%

Religious affiliation	No religion	53%	59%
	Christian	39%	33%
	Other	7%	6%
Experience with trauma (select all that apply)	Self	43%	50%
	Partner	19%	12%
	Member of the household	15%	18%
	Family member	38%	43%
	Close friend	34%	35%
	None	25%	30%

Attitudes towards MMTs in specific scenarios differ between countries and information groups

Two factors in the scenarios showed an interaction with the country of residence of the participants. Participants residing in Mexico showed a more positive attitude towards application of MMTs in healthy subjects (Medicalization x Country interaction effect $F_{(1,693)}=14.773$, $p<0.001$, $\eta^2=14.773$, Mexico: 4.39±0.06, Figure S1A) as compared to participants in the USA ($t(493)=2.231$, $p=0.026$, USA: 4.16±0.08) and the Netherlands ($t(473)=-3.236$, $p=0.001$, the Netherlands: 4.08±0.07). Attitudes were similar towards scenarios in which the subject developed PTSD. Participants residing in different countries also responded differentially to actors with varying levels of agency (Agency x Country interaction effect, $F_{(1.326,988.422)}=4.180$, $p=0.027$, $\eta^2=0.006$, see S1B). For scenarios that described accidental actors, participants in Mexico had a more positive attitude towards MMT compared to participants in the USA ($t(493)=2.990$, $p=0.003$, Mexico: 5.16±0.06, USA: 4.85±0.08) and the Netherlands ($t(473)=2.893$, $p=0.004$, the Netherlands: 4.87±0.07). For scenarios that described observers, participants in the USA showed less positive attitudes compared to participants in Mexico ($t(493)=-4.170$, $p<0.001$, USA: 5.22±0.07, Mexico: 5.62±0.06) and the Netherlands ($t(460)=-2.375$, $p=0.018$, the Netherlands: 5.45±0.07). Attitudes were comparable for scenarios that described intentional murder.

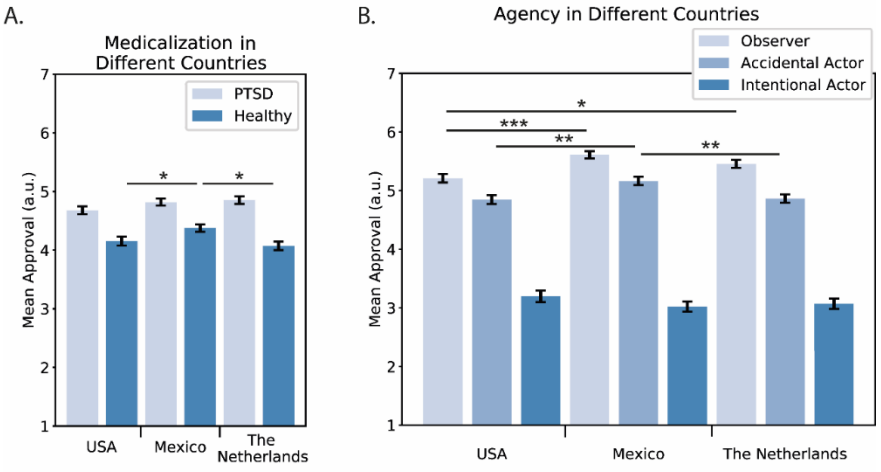


Figure 5.8. Attitudes towards scenarios describing varying levels of agency and medicalization differ between countries. *Error bars represent ± S.E.M., * p<0.05, ** p<0.01, *** p<0.001*

The brief and extensive introduction groups had different attitudes towards scenarios with different interactions between agency and stakeholder (Agency x Stakeholder x Introduction Group interaction effect, $F_{(1.597,1106.844)}=4.164$, $p=0.023$, $\eta^2=0.006$) and agency and medicalization (Agency x Medicalization x Introduction Group interaction effect, $F_{(1.740,\ 1199.889)}=6.316$, $p=0.003$, $\eta^2=0.009$). However, post-hoc contrasts did not reveal any specific significant differences.

Stepwise regression model

We built a stepwise regression model with age, gender (male as baseline, female), ethnicity (Caucasian as baseline, Asian, Black/African, Hispanic, Latinx), education level (bachelor's degree as baseline, high-school, some college but no degree, associate degree, master's degree), employment status (full-time employment as baseline, part-time employment, not employed, homemaker, full-time student, part-time student), income (25-50k as baseline, <10k, 10k-25k, 50-75k, 75-100k, 100-150k), marital status (single as baseline, married, separated), parental status (not a parent as baseline, parent), political identity (moderate as baseline, right-wing/conservative, left-wing/liberal), country of residence (USA as baseline, Mexico, the Netherlands), religious affiliation (no religion as baseline, Christian), personal experience with traumatic events (no experience as baseline, personal experience, partner, other member of the household, family member, close friend) and personal experience with working in the military (no experience as baseline, family member, close friend).

Chapter 6: General Discussion

Given that current treatments for maladaptive threat memories leave patients susceptible to relapse, the identification of mechanisms that can persistently attenuate threat responses is a key step in the advancement of treatment for anxiety-, trauma- and stressor-related disorders. In this thesis, I set out to identify mechanisms of safety learning that can prevent the recovery of previously acquired threat responses. To this end, we conducted a behavioural study in rodents and experiments involving behavioural, physiological and fMRI measurements in humans. Threat memories were established through repeated pairings of the conditioned stimuli and mild electric shocks. Their retention was inferred from behavioural freezing responses in rodents, and physiological responses, including fear-potentiated startle, skin conductance and pupil dilation, in humans. Declarative memory was measured as the ability to remember the specific location of items within a context, or the ability to correctly identify previously presented stimuli. Neural mechanisms were investigated using univariate analyses on BOLD-fMRI data acquired during experimental tasks. In addition, we used an online study to explore public attitudes towards novel memory modification techniques.

In **chapter 2**, we investigated whether presentation of an isolated reminder before extinction, aimed to render the memory labile and sensitive to updating through reconsolidation processes, can persistently attenuate contextual threat memories in humans. We found that post-retrieval extinction (PRE) did not prevent the recovery of fear-potentiated startle responses. We also did not find any evidence for enhanced attenuation of (retrospective) shock estimations, or decreased avoidance of the threat-conditioned context in the PRE compared to regular extinction group. Participants that received a reminder before extinction were also equally able to remember the location of items in the conditioned and unconditioned contexts compared to participants who did not receive a reminder. These data show that presentation of an isolated reminder before extinction neither enhanced nor attenuated the extinction of the conditioned contextual threat memory. While previous studies showed disrupted threat responses after PRE for cue-conditioned threat memoires, we failed to extend these findings to contextually conditioned threat memories, suggesting that contextual threat memories may be resistant to disruption through PRE. This could indicate either that the reminder did not successfully destabilize the contextual threat memory or that subsequent extinction failed to disrupt the labile memory. However, given the increasing number of studies that fail to replicate disruption of conditioned threat responses after an isolated reminder before extinction (in particular see Chalkia et al., 2020; Luyten & Beckers, 2017), it is perhaps more likely that our current understanding of the reconsolidation process is too limited to produce reliable effects.

In response to the increasing number of null findings in the field of reconsolidation, it has been suggested that the efficacy of reconsolidation-based interventions may depend on specific experimental parameters or boundary conditions. In **Chapter 3,** we investigated whether the intensity of aversive Pavlovian threat conditioning could be a boundary condition that limits the effectiveness of PRE at higher intensities of the aversive unconditioned stimulus. Rats conditioned using aversive electric shocks at increasing intensities showed increased conditioned freezing responses during the acquisition phase, the reminder trial, and the extinction phase. However, all groups showed comparable reinstatement of conditioned freezing responses, which was unaffected by the presentation of an isolated reminder before extinction. Thus, our data do not indicate that the "strength" of conditioned threat memories affects the efficacy of the PRE.

Given that our current understanding of PRE appears to be too limited to consistently reproduce disruption of threat responses, we next turned to a second strategy for attenuation of threat responses that aims to enhance novel safety learning. In **Chapter 4**, we investigated whether the neural mechanisms underlying counterconditioning (CC) are distinct from the mechanisms underlying extinction. Compared to regular extinction, CC was able to prevent the spontaneous recovery of differential conditioned PDRs. At a neural level, we found that CC engages a different network compared to extinction, including enhanced recruitment of the nucleus accumbens, and increased suppression of activity in the ventromedial prefrontal cortex and the hippocampus. In addition, CC resulted in retrospective enhancement of item recognition of stimuli from the conditioned category presented during the acquisition phase and enhanced recognition of items from the conditioned category presented during the counterconditioning phase. These findings indicate that CC recruits a network distinct from the network recruited by classic extinction, resulting in a strengthened inhibition of the conditioned threat memory.

In addition to behavioural interventions, threat memories may also be disrupted using more artificial Memory Modification Techniques (MMTs), for instance through pharmacological approaches, although this type of intervention may be received with caution by the public. In **Chapter 5**, we explored public attitudes towards MMTs. We found that general attitudes towards MMTs were relatively positive and dependent on the specific context in which they are used. The belief that MMTs are safe was found to be an important predictor of positive attitudes towards MMTs, while demographic predictors and moral intuitions had a minimal contribution. These results indicate that, contrary to views reflected in the initial debate among bioethicists, the public does not express strong moral reservations with respect to the modification of memories through novel techniques, but rather safety concerns.

Integration of the findings and open questions

In this thesis, we investigated two conceptually different interventions, a reconsolidation-based intervention and an intervention directed at enhancing novel safety learning, both aiming to attenuate conditioned threat responses in a more persistent manner compared to classic extinction learning. While a reconsolidation-based intervention theoretically appears to be the most promising approach as it targets the problem of maladaptive memories at the root by directly modifying the threat memories themselves, we did not find any evidence that post-retrieval extinction was able to modify contextual threat memories in humans (**chapter 2**) or cue-conditioned memories of different strengths in rodents (**chapter 3**). These findings fit to a larger trend of mixed findings within the field of reconsolidation (for a meta-analysis see Kredlow et al., 2016), and are especially unfortunate given that public attitudes towards reconsolidation based MMTs were positive (**chapter 5**). Yet, while the direct modification of maladaptive memories appeared to be challenging, CC proved to be a promising alternative approach to reduce the recovery of threat responses (**chapter 4**). In combination with reduced threat recovery, retrospectively enhanced item memory for the acquisition phase even suggests that within-session CC can engage processes that influence the consolidation of episodic memory for the previous acquisition phase.

These findings raise new questions. But first, the null findings in **chapter 2** and **3** leave us with several unanswered questions. Could the lack of an effect of presenting an isolated reminder before extinction for contextual conditioned threat memories indicate that hippocampus-dependent memories are less susceptible to modification through reconsolidation-based interventions? Or should the lack be interpreted more generally as a failure to conceptually replicate previous findings of persistent attenuation after post-retrieval extinction? Could a different reactivation procedure be more effective in destabilizing the threat memory, and might a different intervention be better able to disrupt labile memories? After addressing these questions, I will discuss the strengths and limitations of the work investigating CC (**chapter 4**) and will make suggestions for interesting follow-up studies. In addition, I will discuss to what extent classic extinction could engage mechanisms of CC less potently, and evaluate to what extent there is evidence to support the view that the two types of interventions, reconsolidation- and enhanced extinction-based interventions, rely on completely dissociable mechanisms. Finally, I will discuss the potential clinical implications of the work in this thesis, including a note on how public attitudes towards MMTs may inform policies regarding treatment with MMTs.

1. Reconsidering reconsolidation

While we hoped to find that PRE would persistently reduce conditioned contextual threat responses in humans, we did not find any additional effect after PRE compared to regular extinction (**chapter 2**).

Given the mixed findings in the field, it seems unlikely that our failure to extend the PRE effect to contextually conditioned threat memories should be interpreted as a specific failure for contextual conditioning, especially in light of the recently published failure (Chalkia et al., 2020) to verify the conclusions of the original publication that demonstrated the effectiveness of PRE in humans (Daniela Schiller et al., 2010). Indeed, to facilitate the interpretation of negative findings, any attempt to extend the original findings should include a positive control condition that demonstrates the efficacy of PRE to persistently attenuate threat responses. However, while we are generally unable to consistently reproduce the initial retrieval-extinction effect in cued fear conditioning, further attempts to extend the original findings may be futile. It has been suggested that a better understanding of boundary conditions or moderators of the effect is needed to increase reproducibility (Auber et al., 2013; Nader, 2003; Zuccolo & Hunziker, 2019). Yet, our attempt to investigate whether memory strength could be a boundary condition was also unsuccessful (**chapter 3**) as we did not find an effect of a reminder before extinction for any level of memory strength. Indeed, if even direct replication attempts, that adhere to the original procedures as closely as possible, fail (Luyten & Beckers, 2017), it becomes challenging to find evidence for or against potential boundary conditions. In spite of this apparent impasse in the field, I will first discuss how the negative findings in **chapter 2** and **3** may fit within the broader field of reconsolidation research. Then, I will discuss how a quest for the identification of boundary conditions may be fundamentally flawed, and how open-science practices can increase reproducibility.

Are hippocampus-dependent memories susceptible to disruption through the post-retrieval extinction?

Our failure to persistently attenuate contextual conditioned threat memories using PRE (**chapter 2**) could indicate that in humans, contextual threat memories are not susceptible to updating through PRE. PRE was first demonstrated to reduce recovery of threat responses for cue-conditioned threat memories in rodents (Monfils et al., 2009), a findings that was shortly afterwards translated to humans (Schiller et al., 2010, but see Luyten & Beckers, 2017). Cue-conditioned threat memories can be disrupted by blocking novel protein synthesis in the amygdala after presentation of an isolated reminder (Nader et al., 2000) and PRE also appears to target processes in the amygdala, evidenced by a selective, local increase in markers of synaptic plasticity after presentation of an isolated reminder (Monfils et al., 2009). Indeed, cue-conditioned threat memories are thought to rely on 'lower-level' rapid, automatic learning of an implicit association between the cue and threat that largely relies on processing within the amygdala and can be expressed in the absence of the cortex (Romanski & LeDoux, 1992) through a lower-level thalamo-amygdala pathway (Grillon, 2009; Luyten & Beckers, 2017; Phelps et al., 2005). Because other forms of threat conditioning have different neural

underpinnings, relying on higher-order mechanisms that are relatively slow and deliberate and that require hippocampal processing, they could be less susceptible to disruption through PRE.

To evaluate whether persistent disruption of threat responses after PRE is a phenomenon that is specific to the amygdala, or whether it extends to the hippocampus, we can compare its effect for cued and contextual threat conditioning, as the hippocampus is required for contextual but not cued threat conditioning (Phillips & LeDoux, 1992). Initial studies in rodents indicated that PRE is also effective for persistent attenuation of contextual threat memories in mice (Rao-Ruiz et al., 2011) and rats (Flavell et al., 2011; Liu et al., 2014; Monti et al., 2017; Piñeyro et al., 2014), although others failed to find such an effect (Chan, 2014; Costanzi et al., 2011). In humans, a study using cue-in-context conditioning has suggested that human contextual threat memories are not sensitive to disruption through PRE (Meir Drexler et al., 2014). In this variant of the contextual conditioning paradigm, the context acts as an occasion setter that modulates cue-related conditioned responses, as opposed to foreground context conditioning where the context itself elicits the conditioned responses (Andreatta et al., 2015). These two versions of contextual conditioning could rely on two distinct neural systems. Foreground context conditioning appears to require a map-like representation of a context containing the relative locations of cues, relying on a conjunctive representation that is thought to be mediated by the hippocampal formation (Nadel & Willner, 1980; Rudy, 2009). Cue-in-context conditioning, on the other hand may only requires representations of the individual cues, and this feature recognition could be mediated by the neocortex in the absence of the hippocampal formation (Nadel & Willner, 1980; Rudy, 2009), although others have argued that context-dependent expression of fear to cues does require hippocampal processing (see e.g. Maren et al., 2013). Nevertheless, a failure to find an effect in cue-in-context conditioning may leave open the possibility that PRE could also lead to a persistent attenuation of contextual threat memories in humans in foreground context conditioning. However, using virtual reality to directly translate the contextual threat conditioning paradigm used in rodents, we still did not find evidence that PRE can prevent the recovery of threat responses (**chapter 2**). Therefore, it seems likely that in contrast to studies in rodents, PRE does not lead to a persistent attenuation of contextual conditioned threat memories in humans.

How can we explain this failure to translate successful disruption of contextual conditioned threat memories using PRE in rodents to humans? It has previously been suggested that explicit contingency awareness in humans may reduce the effectiveness of PRE by increasing the relative dominance of higher-order, hippocampus-dependent processing (Bechara et al., 1995; Grillon, 2009; Kredlow et al., 2015; Weike et al., 2007). Yet, given that contextual threat memories are by themselves hippocampus-dependent in both rodents and humans, it seems unlikely that a difference in sensitivity of rodent and human contextual threat memories to PRE can be explained in terms of hippocampal dependence.

Instead, we may thus speculate that explicit awareness of the relationship between the US and context could occlude a potential effect of PRE on contextual conditioned threat responses in humans, but not in rodents. However, this speculation may be of limited value while it remains unclear whether our null finding is specific to contextual threat memories or may indicate a broader inability to (conceptually) replicate a persistent disruption of threat responses to PRE.

Memory strength as boundary condition in reconsolidation-based interventions

Memory strength has been identified as a boundary condition that limits the effectiveness of pharmacological reconsolidation-based interventions for stronger memories (Finnie & Nader, 2020; Gazarini et al., 2015; Haubrich et al., 2020; Holehonnur et al., 2016; Kwak et al., 2012; Suzuki et al., 2004; Wang et al., 2009). Unlike these findings from pharmacological interventions, a meta-analysis across several studies in rodents that used different US intensities during threat acquisition suggested a trend towards increased efficacy of PRE in preventing the recovery of threat responses for studies using a higher US intensity (Kredlow et al., 2018). In an attempt to resolve this apparent conflict between the effect of memory strength on the efficacy of pharmacological and behavioural interventions, we systematically investigated how the effectiveness of PRE may depend on US intensity during threat acquisition (**chapter 3**). However, we failed to find any evidence for more persistent attenuation of threat responses after PRE compared to regular extinction, irrespective of shock intensity (**chapter 3**). This finding leaves us with two questions. First, how can we explain the lack of an effect of PRE at any shock intensity in **chapter 3**? And second, returning to the suggested conflicting effect of memory strength on pharmacological and behavioural reconsolidation-based interventions, how plausible is it that memory strength could be a boundary condition for pharmacological reconsolidation-based interventions but not for PRE?

Barring the possibility that PRE simply does not prevent the recovery of threat responses (Chalkia et al., 2020; Luyten & Beckers, 2017), it may be that the strong reinstatement procedure used in c**hapter 3** restored access to the original threat memory, obscuring a potential effect of PRE on retrievability. For the reinstatement procedure, we used the same shock intensity for all groups to avoid the possibility that differences in the return of threat could be attributed to differences in shock intensity during reinstatement. We chose a shock intensity for reinstatement that was novel to all animals and higher than any of the shock intensities used during acquisition to assure that all animals experienced the shock during the reinstatement procedure as more intense than previous shocks. In doing so, we hoped to minimize differences in the reinstatement procedure between groups. It is unlikely that freezing during the reinstatement test is driven exclusively by new learning during the reinstatement procedure, as increased freezing levels over the four trials of the reinstatement test in rats conditioned at higher US intensities provided evidence that rats retained a memory representation of the intensity

of the original aversive conditioning experience (**chapter 3**). However, it has been suggested that reconsolidation-based interventions do not entirely remove memories, but rather reduce synaptic strength, so that mere exposure to the CS does not trigger memory retrieval, while potent, artificial stimulation does (Josselyn & Tonegawa, 2020; Roy et al., 2017). Since we did not test for spontaneous recovery, we might have missed out on the behavioural detection of such potential reductions in synaptic strength, and we cannot exclude the possibility that the strong reinstatement procedure restored access to memories that would otherwise be irretrievable.

Setting aside the negative findings in **chapter 3,** it seems unlikely that memory strength could be a boundary condition for pharmacological reconsolidation-based interventions but not for PRE, as molecular evidence for memory destabilization indicates that stronger memories do not destabilize after presentation of a reminder (Gazarini et al., 2015; Haubrich et al., 2020; Holehonnur et al., 2016; Wang et al., 2009). Destabilization of memories appears to require activation of GluN2B-containing NMDA receptors (Mamou et al., 2006; Milton et al., 2013) and endocytosis or downregulation of GluA2-containing AMPA receptors (Clem & Huganir, 2010; Rao-Ruiz et al., 2011). Increasing evidence suggests that the formation of weak memories appears to upregulate GluN2B levels in postsynaptic densities within the BLA, increasing their susceptibility to destabilization, whereas levels of GluN2B are lower after the formation of strong memories (Haubrich et al., 2020; Holehonnur et al., 2016; Wang et al., 2009). Thus, low levels of GluN2B appear to be a marker for resistance to disruption, and this synaptic profile appears to be mediated by noradrenergic projections from the locus coeruleus to the BLA, strengthening memory (Haubrich et al., 2020). In line with findings of memory strengthening by repeated parings, threat memories that are strengthened by administration of the α_2-adrenoceptor antagonist yohimbine, mimicking noradrenergic activation, are initially resistant to disruption but can be rendered labile after activation of the NMDA receptor agonist D-cycloserine (Gazarini et al., 2015). After a reminder, weak memories appear to destabilize, showing reduced expression of GluA2-containg AMPA receptors in the postsynaptic density and increased levels extrasynaptically within the BLA, while strong memories appear to retain GluA2 in the postsynaptic density (Haubrich et al., 2020). In summary, pharmacological studies in rodents suggest that stronger memories appear to have a synaptic profile that renders them more resistant to destabilization after a reminder. Given that memory destabilization is a common requirement for the efficacy of both pharmacological reconsolidation-based interventions and PRE, it thus seems unlikely that strong memories that are resistant to destabilization can be persistently attenuated through PRE. In light of these findings, it seems that reported trend of increased efficacy of PRE in preventing the recovery of threat responses for studies using a higher US intensity reported in a meta-analysis on PRE in humans (Kredlow et al., 2015) may be a trend finding.

However, although strong memories appear to be less sensitive to destabilization, it could be argued that an unreinforced CS presentation generates a larger prediction error for strong memories, rendering strong memories more prone to destabilization. It has been argued that destabilization may only occur when a reminder evokes an optimal degree of prediction error, or mis-match between what has been learned and is expected and what actually occurs (Pedreira, 2004). When the US intensity during threat acquisition is higher, we may expect the mis-match evoked by an unreinforced CS presentation to be larger. This would be in line the trend towards increased efficacy of PRE in preventing the recovery of threat responses for studies using a higher US intensity reported in a meta-analysis on PRE in humans (Kredlow et al., 2018). Yet while several studies provide experimental evidence in support of the hypothesis that prediction errors drive destabilization (see e.g. Cahill et al., 2018; Chen et al., 2021; Díaz-Mataix et al., 2013; Sevenster et al., 2014), it nevertheless remains puzzling that direct attempts to replicate the original PRE findings, using parameters of reactivation and the intervention method as closely as possible, have failed to replicate the findings, given that they should evoke comparable prediction errors (Chalkia et al., 2020; Luyten et al., 2021).

We should note that limited replicability of PRE effects can also be explained by a failure of extinction during the reconsolidation window to persistently disrupt the conditioned threat memory, instead of indicating that the reminder does not reliably destabilize the memory. However, given that other reconsolidation-based interventions also suffer from limited reproducibility (see e.g. Bos et al., 2014; Chalkia et al., 2019; Elahi et al., 2020; Luyten et al., 2021; Schroyens et al., 2017, 2019), it does seem likely that the presentation of an isolated reminder does not reliably induce memory destabilization and reconsolidation. Thus, it appears that the field is currently unable to fully explain replication failures for both memory disruption through PRE and failure to destabilize memories and disrupt reconsolidation more generally.

Post-hoc boundary conditions: a threat to the principle of falsifiability

While a number of proposed boundary conditions are grounded in mechanistic explanations and have been investigated experimentally, including memory age and strength (Debiec et al., 2002; Elsey & Kindt, 2017; Fernández et al., 2016; Haubrich et al., 2020; Milekic & Alberini, 2002), some caution may be warranted when attributing failure to replicate to *post hoc*, hypothetical boundary conditions or critical methodological differences. For example, in absence of a clear explanation for discrepancies in replication attempts, failures to replicate reconsolidation-based interventions have been suggested to be due to small genetic variations between animals obtained from different suppliers (Luyten et al., 2021; Schroyens, Schnell, et al., 2019) or unidentified differences between laboratories and experimenters (Schroyens, Alfei, et al., 2019). A well-known principle articulated in philosophy of

science is the falsification principle proposed by Karl Popper, that suggests that theories can only be considered scientific when we can conceive a test that would prove that it is false (Thornton, 2021). When we allow any failure to replicate the effect of a reconsolidation-based intervention to be attributed to boundary conditions, including potential boundary conditions that we have not yet identified, it no longer meets the criteria of falsifiability.

In practice, some nuance is needed in that science is clearly more complicated than Popper's textbook science and always relies on assumptions regarding the nature of the experiment (Mulkay & Gilbert, 1981), such as the assumption that a replication attempt was carried out exactly as the to-be-replicated experiment. Nevertheless, it seems clear that keeping in mind the principle of falsification can increase the replicability of findings, and help actively resist "undead" theories, such as the idea of unknown boundary conditions, that remain popular but have little scientific basis (Derksen, 2019; Ferguson & Heene, 2012; Wagenmakers et al., 2012).

Open science practices as weapon against the reproducibility crisis

To maximize reproducibility, we should actively strive to minimize sources of bias in the field, including both publication bias and sources of bias within the research process (Rosenthal, 1979). Publication of null-results and carrying out critical meta-analyses may help to limit the effects of publication bias (Ferguson & Heene, 2012). With regards to the research process, it is key to raise awareness that the field of threat conditioning has many 'researcher degrees of freedom' during data-processing and analysis (Simmons et al., 2011), i.e. many individual decisions during analysis that by itself may not have large influences on the outcome of a study, but collectively can form paths that yield considerably different outcomes (Lonsdorf et al., 2019). An open science culture could reduce bias during the research process, for example through a pre-registration, that encourages researchers to make any decisions surrounding data-processing and analysis prior to the collection of the data, minimizing the possibility that the results are highly dependent on the specific 'forking path' (Lonsdorf et al., 2019; Nosek, 2015). Additionally, increasing transparency in methods, analysis and publicly available data, and incentivizing replication and verification could reduce the influence of potential erroneous or chance findings on the field (Chalkia et al., 2020; Nosek, 2015).

2. Strengthening novel safety learning using reward

In **chapter 4**, we showed that within-session CC can prevent the recovery of threat responses, enhance recognition memory for items from the conditioned category presented during CC, as well as retroactively enhances recognition for conditioned category exemplars presented during the

acquisition phase. Here, I will discuss several limitations and strengths of this work and make suggestions for further research.

It is important to note that CC and extinction took place within the consolidation window for the acquisition of conditioned threat memory. Formally, this may constitute a test of post-conditioning effects of CC and extinction, as opposed to a test of the efficacy of CC and extinction-training for the extinction of a consolidated threat memory. As a result, this investigation of immediate CC may have two limitations for the translation of CC as a treatment for stress-related disorders. First, we may overestimate the efficacy of CC compared to extinction because immediate extinction may be more sensitive to spontaneous recovery compared to delayed extinction (Maren, 2014). For extinction learning, it has been shown in rodents that when the interval between acquisition and extinction is less than 6 hours, the extinction training does not result in a long-term loss of fear but instead results in increased spontaneous recovery, a phenomenon called the immediate extinction deficit (Chang & Maren, 2009; Devenport, 1998; Maren & Chang, 2006). Episodic memory research suggests that the drop in memory for the extinction training is mediated by event boundaries that enable prioritization of the emotional information (i.e. memory for threat acquisition) at the expense of neutral information (i.e. extinction memory) presented immediately after (Dunsmoor et al., 2018). However, in rodents, the immediate extinction deficit does not seem to rely on prioritization of segmented emotional episodes (Totty et al., 2019), but might arise from an inefficient activation of the mPFC during immediate extinction as a result of the stress induced during threat acquisition (Chang et al., 2010; Kim et al., 2010; Maren, 2014). Thus, a limitation of the current study may be that by comparing immediate extinction and CC, we overestimate the recovery of threat responses after extinction due to the immediate extinction deficit, also resulting in an overestimation of the extent to which CC reduces threat recovery compared to extinction.

A second limitation of the experimental design, in which the acquisition and CC phases are only separated by a short break, may be that immediate CC may affect the consolidation of threat acquisition. In contrast to studies demonstrating extinction deficits as a result of immediate extinction, others have shown that extinction immediately after acquisition may result in “unlearning” of the conditioned threat memory, thereby prevent recovery of threat responses (Myers et al., 2006). In fact, we observed that CC retroactively strengthened recognition memory for items presented during acquisition (**chapter 4**), which indicates that CC likely affected the consolidation of threat acquisition. It has previously been shown that reward conditioning can retroactively enhance recognition of neutral items presented prior to reward conditioning (Braun et al., 2018; Patil et al., 2017). Thus, immediate CC seems to influence consolidation of prior threat conditioning, and we may ask whether the observed reduction in spontaneous recovery can also be explained by immediate effects of CC on the

consolidation of the memory for conditioned threat. While we showed that immediate CC prevented the recovery of threat responses compared to extinction (**chapter 4**), it could be that delayed CC does not. Therefore, the findings in **chapter 4** could be limited to immediate CC, and a delayed CC manipulation is needed to establish whether CC is equally effective for consolidated threat memories.

A particular strength of the experimental design in **chapter 4** is that partial reinforcement was used during threat acquisition and that the first three trials during the CC phase were always unrewarded, making the transition from threat acquisition to CC more gradual. According to latent cause models for extinction, sudden transitions in reinforcement result in the inference of a novel latent cause and can prevent extinction trials from influencing the original threat learning (for a review, see Dunsmoor, Niv, et al., 2015). Gradual extinction, on the other hand, in which the frequency of reinforced trials is diminished slowly during the course of the extinction session to allow the latent cause, established during threat acquisition, to be updated so that it is no longer associated with the US, has been shown to prevent spontaneous recovery and reinstatement of conditioned threat responses. Thus, the gradual transition from acquisition to CC may have increased the likelihood that trials from the CC phase were attributed to the latent cause inferred during the acquisition of conditioned threat.

In **chapter** 4, we suggested that strengthened memory after CC could be mediated by reward-induced reverse replay and dopaminergic modulation during CC. Here, we would like to make several suggestions on how future studies could investigate these hypothetical mechanisms experimentally. First, to establish whether replay after CC plays a role in strengthening memory for CC, a resting-state functional MRI scan could be carried out immediately after CC to investigate whether neural activity patterns evoked during CC are reactivated during the consolidation window. In line with previous studies investigating the role of spontaneous reactivations in the retention of extinction memory (Gerlicher et al., 2018), finding that the number of spontaneous reactivations correlates with extinction recall or memory strength may support the hypothesis that enhanced memory after CC is mediated by post-learning reactivation (replay). In turn, increased replay after CC could be mediated by enhanced dopaminergic modulation (Ambrose et al., 2018; Singer & Frank, 2009). The role of the dopaminergic system in CC could be investigated experimentally through pharmacological manipulation of the dopaminergic system. Application of dopamine receptor antagonists (such as haloperidol or risperidone) could provide evidence for the necessity of dopaminergic modulation for fear reduction, memory enhancement and/or spontaneous reactivations after CC. In addition, although less powerful, we may compare neural activity and learning during CC in participants that carry different functional polymorphisms in the mesostriatal dopamine transporter gene DAT1, mirroring previous studies that showed involvement of dopamine in prediction errors during extinction (Raczka et al., 2011). Given that stimulus-specific activation of the nucleus accumbens was increased during CC compared to

extinction, we may expect that carriers of the 9-repeat allele, that is associated with enhanced phasic DA release (Raczka et al., 2011), would show higher learning rates and nucleus accumbens activation during CC, resulting in a stronger safety memory and reduced recovery of threat responses.

Role of reward-related circuits in classic extinction

Recent work in rodents has provided compelling evidence that the omission of expected aversive reinforcement engages reward circuits in the BLA and ventral striatum (Correia et al., 2016; Zhang et al., 2020). During the extinction of conditioned threat, a novel extinction memory may be created in a neuronal population in the BLA that drives reward behaviour and inhibits neurons that mediate threat-related responses (Kim et al., 2016; Zhang et al., 2020). In addition, extinction recruits a reward-sensitive BLA-ventral striatum pathway that appears to supress the recovery of fear (Correia et al., 2016). This circuit can be enhanced by pairing extinction with reward, resulting in reduced recovery of threat responses (Correia et al., 2016). Based on these findings in rodents, it seems likely that extinction and CC may engage the same, reward-related pathways with different potencies. However, our findings in **chapter 4** do provide evidence for graded activation of the reward-related activity observed during CC. While BOLD-responses measured with functional magnetic resonance imaging revealed significant stimulus-specific activation in both the amygdala and the nucleus accumbens during CC, neither region showed significant stimulus-specific changes in activity during extinction. How can we explain this apparent discrepancy between rodent and human work? Taken at face value, the lack of evidence for reward-associated pathways during extinction could indicate that the participants did not experience the omission of USs to be rewarding. Threat conditioning could be more aversive in rodents compared to humans, for instance because in line with ethical regulations, human volunteers are given control over the US intensity and are pre-exposed to the US during its calibration before threat acquisition (Haaker et al., 2019). The threat associated with reinforced CS presentations may not be sufficiently intense to warrant the experience of reward when the US is omitted. A second possibility may be that while reward-related processing is involved in extinction learning in humans, neuroimaging experiments may not be able to detect this activity. For example, while loss-of-function experiments in both rodents and humans reveal a clear involvement of the amygdala in threat conditioning (Bechara et al., 1995; Blanchard & Blanchard, 1972; Gentile et al., 1986; Hitchcock & Davis, 1986; Klumpers et al., 2014), human neuroimaging studies are unable to find evidence for stimulus specific changes in amygdala activity during threat conditioning (Fullana et al., 2016; Visser et al., 2021). Similarly, the amygdala might be involved but remain undetected in extinction learning. In addition to differences between experimental paradigms induced by cross-species translation, functional magnetic resonance imaging may have insufficient spatial resolution to

detect signals in structures like the amygdala that contain nuclei and neuronal subpopulation that can have distinct and opposite functions, for instance coding for reward or threat (Visser et al., 2021). Both threat and reward memories appear to be represented in few, sparsely distributed neurons within the amygdala, and their signal may be cancelled out at voxel level (Redondo et al., 2014; Reijmers et al., 2007; Visser et al., 2021).

In conclusion, CC enhances the recruitment of reward-related brain regions that are also recruited during classic extinction training. Discrepancies between animal research and the findings in **chapter 4** could be explained by poor translation of the experimental paradigm to human participants, limited spatial resolution in neuroimaging or insufficient specificity in the modelled processes in the analysis of neuroimaging data.

3. Dissociating interventions that target reconsolidation and extinction

According to the dominant views on extinction and reconsolidation of conditioned threat memories, extinction and reconsolidation are two distinct processes. Whereas during extinction, a new memory trace that temporarily inhibits the original memory is formed, reconsolidation renders an old memory trace sensitive to change and persistently updates the original memory (Bouton, 2002; Nader et al., 2000). Yet, given the striking similarity between protocols that induce PRE and protocols for regular extinction, differing from regular extinction in the amount of time (e.g. ten additional minutes) between the first and second unreinforced CS presentation, we may ask to what extent extinction and reconsolidation are fully dissociable. Whether a protocol consisting of unreinforced CS presentations engages reconsolidation or extinction has been shown to depend on prediction error and the duration or amount of re-exposure. With regards to prediction errors, the absence of a mismatch between what is expected and observed leads to extinction learning (see e.g. Cahill et al., 2019; Exton-McGuinness et al., 2015; Junjiao et al., 2019; Sevenster et al., 2014). The amount of re-exposure reveals an interesting pattern where brief re-exposure engages reconsolidation and extensive exposure engages extinction, but intermediate degrees of exposure engage neither process (Cassini et al., 2017; Merlo et al., 2014). At a molecular level, it appears that increasing the extent of exposure drives the synthesis of enzymes required for extinction but not reconsolidation, suggesting that the two are dissociable and mutually exclusive processes (Merlo et al., 2014; Su et al., 2021). Yet, while this evidence supports the view that extinction and reconsolidation are distinct processes, it does not imply that the former exclusively creates a novel memory while the latter only modifies an existing memory.

The view that extinction learning only leads to the formation of a novel, temporary safety memory, and reconsolidation-based interventions only make lasting alternations to threat memories, may be too restrictive. Indeed, whereas analysis of mRNA expression in the BLA during acquisition, extinction

and PRE of conditioned threat memories confirmed that memory extinction and PRE are two independent processes for reducing fear memory, it also suggested that activation of pathways for the formation of novel memories are only a minor part of the pathways activated during extinction (Su et al., 2021). Thus, extinction may not exclusively be mediated by the formation of a novel safety memory. At the same time, PRE and other reconsolidation-based interventions may not rely on putative processes of memory destabilization and result in persistent modification of original threat memories. In fact, PRE has been shown to occur in the presence of pharmacological interventions that prevent destabilization of memories consolidated in the BLA (Cahill et al., 2019). At a behavioural level, the view that application of the protein synthesis inhibitor anisomycin after a reminder 'erases' conditioned threat memories, has also been challenged by the finding that conditioned freezing returns after another administration of anisomycin, suggesting that postretrieval anisomycin may only render the expression of the threat memory dependent on the internal state created by the drug (Gisquet-Verrier et al., 2015). Work in the field of memory engrams that identifies and manipulates the specific neurons involved in the formation, expression and retrievability of memories, suggests that amnestic treatments in reconsolidation-based interventions reduce the synaptic efficacy within memory engrams (Roy et al., 2017). As a result, the engram becomes 'silent' but not lost, in that mere exposure to the CS does not trigger memory retrieval while artificial stimulation does (Josselyn & Tonegawa, 2020; Roy et al., 2017). Thus, based on the currently available evidence, we may need to adopt a more nuanced view about the mechanisms underlying interventions that target reconsolidation and extinction, where neither appears to consist exclusively of memory updating or formation of novel memories.

4. Clinical perspective and implications

Lastly, I will discuss the clinical implications of the findings covered by this thesis. **chapter 4** suggests that CC can attenuate threat memories in a persistent manner, preventing the recovery of threat responses, and that this is mediated by neural circuits related to reward learning and motivated behaviour instead of classic extinction networks. The clinical implications of this are two-fold. First, given that CC can prevent the recovery of threat responses compared to classic extinction, its application in the treatment of trauma- and stress-related disorders may reduce relapse rates (Vervliet et al., 2013). Second, CC could be an especially useful alternative to classic extinction in populations of patients that show impairments in extinction learning because CC appears to rely less on the vmPFC, which appears to play a key role in regular extinction training. For example, attenuation of fear through CC and reward-related networks could be key in overcoming treatment resistance in PTSD patients, given that PTSD patients experience difficulty maintaining extinguished responding, which is related

to decreased functionality in hippocampo-prefrontal-amgydala circuits (Milad et al., 2008, 2009). Similarly, CC may be a suitable alternative for several other populations with extinction resistance, including adolescents, who may be impaired due to slow maturation of the prefrontal cortex (Pattwell et al., 2012) and patients that suffered from early-life or chronic stress that led to extinction impairments (Maren & Holmes, 2016). In addition, recently developed forms of implicit CC using fMRI neurofeedback do not require any exposure to trauma-related materials and could be used in patients that do not tolerate the direct exposure during classic extinction-based treatments (Taschereau-Dumouchel et al., 2018).

While we did not find evidence that PRE can prevent the recovery of fear, and fundamental work on reconsolidation-based interventions generally suffers from limited replicability, it is worth noting that in a meta-analysis of clinical trials, the recently emerged 'reconsolidation therapies' that are informed by reconsolidation theory, have nevertheless been found to be effective in the treatment of PTSD (Astill Wright et al., 2021). While there was no meta-analytic evidence that classic pharmacological interventions, in which potentially amnestic pharmacological agents such as propranolol are administered combined with memory reactivation, are effective in the treatment of PTSD, a psychological reconsolidation-based intervention was found to have a large effect (Astill Wright et al., 2021). The intervention, termed Reconsolidation of Traumatic Memories, consists of an incomplete trauma narrative as a reminder, where the narrative is interrupted as soon as signs of physiological arousal are observed, followed by classic cognitive exercises that aim to transform the traumatic memory into a complete narrative in a more distant, third-person status (Gray et al., 2019). Thus, while a direct translation of the original reconsolidation protocols has proven to be challenging (**chapter 2, chapter 3)**, some aspects of reconsolidation research have nevertheless proved useful, although the underlying mechanism, i.e. to what extent the efficacy relies on the putative mechanisms of reconsolidation, remains unclear.

'Memory editing' may become a standard part of clinical practice as we develop an increasing understanding of the mechanisms underlying the reconsolidation of threat memories and the consolidation of extinction memories (Phelps & Hofmann, 2019). However, the use of novel Memory Modification Techniques (MMTs) that target these processes in specific threat memories, has raised concerns among bioethicists (for a review, see Kroes & Liivoja, 2018). Yet, in contrast to the dominant view expressed by bioethicists, the public does not appear to have strong fundamental concerns about MMTs (**chapter 5**). Instead, public attitudes towards MMTs are positive when they are believed to be safe and constitute an effective treatment for PTSD without unwanted side effects (**chapter 5**). It could be that the introduction of techniques that persistently alter traumatic memories, instead of

temporarily inhibiting them, may especially be perceived as a large shift among scientists, while the general public may not feel that MMTs are radically different from existing forms of psychotherapy (Elsey & Kindt, 2016). Although there may be a public demand for regulation of the use of MMTs, as treatment approval varied depending on the specific case in which MMTs were used (**chapter 5**), it seems likely that the these novel treatment options will be well received.

5. Conclusion

In this thesis, we set out to identify mechanisms of safety learning that can prevent the recovery of threat responses after initial safety learning. To this end, we compared classic extinction training to two novel interventions, each using a different hypothetical strategy to reduce the recovery of threat responses. We did not find evidence for enhanced attenuation of threat responses after PRE that aims to persistently update the original threat memory during reconsolidation. However, using a second strategy aimed at enhancing extinction learning instead of updating the threat memory, we found that CC can prevent the recovery of threat responses through the engagement of reward circuits. Together, these studies increase our understanding of potential mechanisms for persistent attenuation of threat responses and may pave the way towards treatments that allow specific memories to be edited. From a scientific perspective, the development of techniques that alter specific memories would profoundly change the nature of treatment for stress-related disorders and this has sparked ethical debate. Yet the public appears to be positively disposed to the introduction of novel techniques to modify memories and may welcome the introduction of novel, safe and effective treatments for PTSD, even when their mechanisms are fundamentally different from existing treatment options.

References

Acheson, D., Feifel, D., De Wilde, S., McKinney, R., Lohr, J., & Risbrough, V. (2013). The effect of intranasal oxytocin treatment on conditioned fear extinction and recall in a healthy human sample. *Psychopharmacology*, *229*(1), 199–208. https://doi.org/10.1007/s00213-013-3099-4

Agren, T., Engman, J., Frick, A., Björkstrand, J., Larsson, E.-M., Furmark, T., & Fredrikson, M. (2012). Disruption of Reconsolidation Erases a Fear Memory Trace in the Human Amygdala. *Science*, *337*. https://doi.org/10.1371/journal.pone.0129393

Alberini, C. M. (2005). Mechanisms of memory stabilization: Are consolidation and reconsolidation similar or distinct processes? *Trends in Neurosciences*, *28*(1), 51–56. https://doi.org/10.1016/j.tins.2004.11.001

Alberini, C. M. (2011). The Role of Reconsolidation and the Dynamic Process of Long-Term Memory Formation and Storage. *Frontiers in Behavioral Neuroscience*, *5*. https://doi.org/10.3389/fnbeh.2011.00012

Alberini, C. M., & Ledoux, J. E. (2013). Memory reconsolidation. *Current Biology*, *23*(17). https://doi.org/10.1007/7854_2016_463

Alexandra Kredlow, M., Unger, L. D., & Otto, M. W. (2016). Harnessing reconsolidation to weaken fear and appetitive memories: A meta-analysis of post-retrieval extinction effects. *Psychological Bulletin*, *142*(3), 314–336. https://doi.org/10.1037/bul0000034

Alfei, J. M., Ferrer Monti, R. I., Molina, V. A., De Bundel, D., Luyten, L., & Beckers, T. (2020). Generalization and Recovery of Post-Retrieval Amnesia. *Journal of Experimental Psychology: General*, *149*(11), 2063–2083. https://doi.org/10.1037/xge0000765

Alfei, J. M., Monti, R. I. F., Molina, V. A., Bueno, A. M., & Urcelay, G. P. (2015). Prediction error and trace dominance determine the fate of fear memories after post-training manipulations. *Learning and Memory*, *22*(8), 385–400. https://doi.org/10.1101/lm.038513.115

Alvarez, R. P., Biggs, A., Chen, G., Pine, D. S., & Grillon, C. (2008). Contextual Fear Conditioning in Humans: Cortical-Hippocampal and Amygdala Contributions. *Journal of Neuroscience*, *28*(24), 6211–6219. https://doi.org/10.1523/JNEUROSCI.1246-08.2008

Ambrose, R. E., Pfeiffer, B. E., & Foster, D. J. (2018). Reverse Replay of Hippocampal Cells is Uniquely Modulated By Changing Reward. *Neuron*, *91*(5), 1124–1136. https://doi.org/10.1016/j.neuron.2016.07.047.Reverse

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-V*.

Andreatta, M., Leombruni, E., Glotzbach-Schoon, E., Pauli, P., & Mühlberger, A. (2015). Generalization of Contextual Fear in Humans. *Behavior Therapy*, *46*(5), 583–596. https://doi.org/10.1016/j.beth.2014.12.008

Astill Wright, L., Horstmann, L., Holmes, E. A., & Bisson, J. I. (2021). Consolidation/reconsolidation therapies for the prevention and treatment of PTSD and re-experiencing: a systematic review and meta-analysis. *Translational Psychiatry*, *11*(1), 1–14. https://doi.org/10.1038/s41398-021-01570-w

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Atherton, L. A., Dupret, D., & Mellor, J. R. (2015). Memory trace replay: The shaping of memory consolidation by neuromodulation. *Trends in Neurosciences*, *38*(9), 560–570. https://doi.org/10.1016/j.tins.2015.07.004

Auber, A., Tedesco, V., Jones, C. E., Monfils, M.-H., & Chiamulera, C. (2013). Post-retrieval extinction as reconsolidation interference: methodological issues or boundary conditions? *Psychopharmacology*, *226*(4), 631–647. https://doi.org/:10.1007/s00213-013-3004-1

Auchter, A., Cormack, L. K., Niv, Y., Gonzalez-Lima, F., & Monfils, M. H. (2017). Reconsolidation-extinction interactions in fear memory attenuation: The role of inter-trial interval variability. *Frontiers in Behavioral Neuroscience*, *11*(January), 1–9. https://doi.org/10.3389/fnbeh.2017.00002

Auchter, A. M., Shumake, J., Gonzalez-Lima, F., & Monfils, M. H. (2017). Preventing the return of fear using reconsolidation updating and methylene blue is differentially dependent on extinction learning. *Scientific Reports*, *7*(April), 1–13. https://doi.org/10.1038/srep46071

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baker, K. D., McNally, G. P., & Richardson, R. (2013). Memory retrieval before or after extinction reduces recovery of fear in adolescent rats. *Learning and Memory*, *20*(9), 467–473. https://doi.org/10.1101/lm.031989.113

Ballarini, F., Moncada, D., Martinez, M. C., Alen, N., & Viola, H. (2009). Behavioral tagging is a general mechanism of long-term memory formation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(34), 14599–14604. https://doi.org/10.1073/pnas.0907078106

Barto, A. G. (2018). Adaptive Critics and the Basal Ganglia. *Models of Information Processing in the Basal Ganglia*. https://doi.org/10.7551/mitpress/4708.003.0018

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., R, A., & Damasio, A. R. (1995). Double Dissociation of Conditioning and Declarative Knowledge Relative to the Amygdala and Hippocampus in Humans. *Science*, *269*(5227), 1115–1118.

Beckers, T., & Kindt, M. (2017). Memory Reconsolidation Interference as an Emerging Treatment for Emotional Disorders: Strengths, Limitations, Challenges, and Opportunities. *Annual Review of Clinical Psychology*, *13*, 99–121. https://doi.org/10.1146/annurev-clinpsy-032816-045209

Berkman, L. F., & Syme, S. L. (1994). Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. In A. Steptoe & J. Wardle (Eds.), *Psychosocial Processes and Health: A Reader* (pp. 43–67). Cambridge University Press. https://doi.org/10.1017/CBO9780511759048.005

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., Stokes, J., Handelsman, L., Medrano, M., Desmond, D., & Zule, W. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse and Neglect*, *27*(2), 169–190. https://doi.org/10.1016/S0145-2134(02)00541-0

Björkstrand, J., Agren, T., Frick, A., Engman, J., Larsson, E. M., Furmark, T., & Fredrikson, M. (2015). Disruption of memory reconsolidation erases a fear memory trace in the human amygdala: An 18-month follow-up. *PLoS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0129393

Blanchard, D. C., & Blanchard, R. J. (1972). Innate and conditioned reactions to threat in rats with amygdaloid lesions. *Journal of Comparative and Physiological Psychology*, *81*(2), 281–290.

Blanchard, R., & Blanchard, D. (1969). Crouching as an index of fear. *Journal of Comparative and Physiological Psychology*, *67*(3), 370–375.

Blanchard, R. J., Dielman, T. E., & Blanchard, D. C. (1968). Postshock crouching: Familiarity with the shock situation. *Psychonomic Science*, *10*(11), 371–372. https://doi.org/10.3758/BF03331566

Boccia, M. M., Acosta, G. B., Blake, M. G., & Baratti, C. M. (2004). Memory consolidation and reconsolidation of an inhibitory avoidance response in mice: Effects of i.c.v. injections of hemicholinium-3. *Neuroscience*, *124*(4), 735–741. https://doi.org/10.1016/j.neuroscience.2004.01.001

Boeke, E. A., Moscarello, J. M., LeDoux, J. E., Phelps, E. A., & Hartley, C. A. (2017). Active avoidance: Neural mechanisms and attenuation of pavlovian conditioned responding. *Journal of Neuroscience*, *37*(18), 4808–4818. https://doi.org/10.1523/JNEUROSCI.3261-16.2017

Bos, M. G. N., Beckers, T., & Kindt, M. (2014). Noradrenergic blockade of memory reconsolidation: A failure to reduce conditioned fear responding. *Frontiers in Behavioral Neuroscience*, *8*(NOV), 1–8. https://doi.org/10.3389/fnbeh.2014.00412

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry*, *52*(10), 976–986. https://doi.org/10.1016/s0006-3223(02)01546-9

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory*, *11*(5), 485–494. https://doi.org/10.1101/lm.78804

Bouton, M. E., & Moody, E. W. (2004). Memory processes in classical conditioning. *Neuroscience and Biobehavioral Reviews*, *28*(7), 663–674. https://doi.org/10.1016/j.neubiorev.2004.09.001

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Braun, E. K., Wimmer, G. E., & Shohamy, D. (2018). Retroactive and graded prioritization of memory by reward. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-07280-0

Brewin, C. R. (2006). Understanding cognitive behaviour therapy: A retrieval competition account. *Behaviour Research and Therapy*, *44*(6), 765–784. https://doi.org/10.1016/j.brat.2006.02.005

Brooks, D. C., Hale, B., Nelson, J. B., & Bouton, M. E. (1995). Reinstatement after counterconditioning. *Animal Learning & Behavior*, *23*(4), 383–390. https://doi.org/10.3758/BF03198938

Brown, S., & Shafer, E. (1888). An investigation into the functions of the occipital and temporal lobes of themonkey's brain. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *179*, 303–327.

Brunet, A., Orr, S. P., Tremblay, J., Robertson, K., Nader, K., & Pitman, R. K. (2008). Effect of post-retrieval propranolol on psychophysiologic responding during subsequent script-driven traumatic imagery in post-traumatic stress disorder. *Journal of Psychiatric Research*, *42*(6), 503–506. https://doi.org/10.1016/j.jpsychires.2007.05.006

Brzosko, Z., Schultz, W., & Paulsen, O. (2015). Retroactive modulation of spike timingdependent

plasticity by dopamine. *ELife*, *4*(OCTOBER2015), 1–13. https://doi.org/10.7554/eLife.09685

Buckholtz, J. W., Treadway, M. T., Cowan, R. L., Woodward, N. D., Benning, S. D., Li, R., Ansari, M. S., Baldwin, R. M., Schwartzman, A. N., Shelby, E. S., Smith, C. E., Cole, D., Kessler, R. M., & Zald, D. H. (2010). Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits. *Nature Neuroscience*, *13*(4), 419–421. https://doi.org/10.1038/nn.2510

Bui, E., Orr, S. P., Jacoby, R. J., Keshaviah, A., LeBlanc, N. J., Milad, M. R., Pollack, M. H., & Simon, N. M. (2013). Two weeks of pretreatment with escitalopram facilitates extinction learning in healthy individuals. *Human Psychopharmacology*, *28*, 447–456. https://doi.org/10.1002/hup

Bulganin, L., Bach, D. R., & Wittmann, B. C. (2014). Prior fear conditioning and reward learning interact in fear and reward networks. *Frontiers in Behavioral Neuroscience*, *8*(March), 1–11. https://doi.org/10.3389/fnbeh.2014.00067

Cabrera, L. Y., & Elger, B. S. (2016). Memory Interventions in the Criminal Justice System: Some Practical Ethical Considerations. *Journal of Bioethical Inquiry*, *13*(1), 95–103. https://doi.org/10.1007/s11673-015-9680-2

Cabrera, L. Y., Fitz, N. S., & Reiner, P. B. (2015). Empirical Support for the Moral Salience of the Therapy-Enhancement Distinction in the Debate Over Cognitive, Affective and Social Enhancement. *Neuroethics*, *8*(3), 243–256. https://doi.org/10.1007/s12152-014-9223-2

Cahill, E. N., & Milton, A. L. (2019). Neurochemical and molecular mechanisms underlying the retrieval-extinction effect. *Psychopharmacology*, *236*(1), 111–132. https://doi.org/10.1007/s00213-018-5121-3

Cahill, E. N., Wood, M. A., Everitt, B. J., & Milton, A. L. (2019). The role of prediction error and memory destabilization in extinction of cued-fear within the reconsolidation window. *Neuropsychopharmacology*, *44*(10), 1762–1768. https://doi.org/10.1038/s41386-018-0299-y

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, *21*(7), 294–299. https://doi.org/10.1016/S0166-2236(97)01214-9

Cannon, B. (1929). Organization for Physiological Homeostasis. *Physiological Reviews*, *IX*(3), 399–431. https://doi.org/10.1152/physrev.1929.9.3.399

Carleton, R. N., Norton, M. A. P. J., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders*, *21*(1), 105–117. https://doi.org/10.1016/j.janxdis.2006.03.014

Cassini, L. F., Flavell, C. R., Amaral, O. B., & Lee, J. L. C. (2017). On the transition from reconsolidation to extinction of contextual fear memories. *Learning and Memory*, *24*(9), 392–399. https://doi.org/10.1101/lm.045724.117

Chalkia, A., Van Oudenhove, L., & Beckers, T. (2020). Preventing the return of fear in humans using reconsolidation update mechanisms: A verification report of Schiller et al. (2010). *Cortex*, *129*, 510–525. https://doi.org/10.1016/j.cortex.2020.03.031

Chalkia, A., Weermeijer, J., Van Oudenhove, L., & Beckers, T. (2019). Acute but not permanent effects of propranolol on fear memory expression in humans. *Frontiers in Human Neuroscience*, *13*(February), 1–14. https://doi.org/10.3389/fnhum.2019.00051

Chan, W. Y. M. (2014). *The effects of retrieval - extinction training on the restoration of Pavlovian conditioned fear (Unpublished PhD thesis)* (Issue January). University of New South Wales, New South Wales, Australia.

Chan, W. Y. M., Leung, H. T., Westbrook, R. F., & McNally, G. P. (2010). Effects of recent exposure to a

conditioned stimulus on extinction of Pavlovian fear conditioning. *Learning & Memory*, *7*, 5120521. http://www.learnmem.org/cgi/doi/10.1101/lm.1912510

Chang, C. hui, Berke, J. D., & Maren, S. (2010). Single-unit activity in the medial prefrontal cortex during immediate and delayed extinction of fear in rats. *PLoS ONE*, *5*(8). https://doi.org/10.1371/journal.pone.0011971

Chang, C. hui, & Maren, S. (2009). Early extinction after fear conditioning yields a context-independent and short-term suppression of conditional freezing in rats. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *16*(1), 62–68. https://doi.org/10.1101/lm.1085009

Chen, W., Li, J., Xu, L., Zhao, S., Fan, M., & Zheng, X. (2021). Destabilizing Different Strengths of Fear Memories Requires Different Degrees of Prediction Error During Retrieval. *Frontiers in Behavioral Neuroscience*, *14*(January), 1–18. https://doi.org/10.3389/fnbeh.2020.598924

Chen, Y. Y., Zhang, L. B., Li, Y., Meng, S. Q., Gong, Y. M., Lu, L., Xue, Y. X., & Shi, J. (2019). Post-retrieval extinction prevents reconsolidation of methamphetamine memory traces and subsequent reinstatement of methamphetamine seeking. *Frontiers in Molecular Neuroscience*, *12*(July), 1–11. https://doi.org/10.3389/fnmol.2019.00157

Chhatwal, J. P., Davis, M., Maguschak, K. A., & Ressler, K. J. (2005). Enhancing cannabinoid neurotransmission augments the extinction of conditioned fear. *Neuropsychopharmacology*, *30*(3), 516–524. https://doi.org/10.1038/sj.npp.1300655

Chhatwal, J. P., & Ressler, K. J. (2007). Modulation of fear and anxiety by the endogenous cannabinoid system. *CNS Spectrums*, *12*(3), 211–220. https://doi.org/10.1017/S1092852900020939

Christianson, S., & Loftus, E. F. (1987). Memory for traumatic events. *Applied Cognitive Psychology*, *1*(4), 225–239. https://doi.org/10.1002/acp.2350010402

Clem, R. L., & Huganir, R. L. (2010). Calcium-Permeable AMPA Receptor Dynamics Mediate Fear Memory Erasure. *Science*, *330*(6007), 1108–1113. https://doi.org/10.1126/science.1195298

Correia, S. S., McGrath, A. G., Lee, A., Graybiel, A. M., & Goosens, K. A. (2016). Amygdala-ventral striatum circuit activation decreases long-term fear. *ELife*, *5*(September), 1–25. https://doi.org/10.7554/eLife.12669

Costanzi, M., Cannas, S., Saraulli, D., Rossi-Arnaud, C., & Cestari, V. (2011). Extinction after retrieval: Effects on the associative and nonassociative components of remote contextual fear memory. *Learning and Memory*, *18*(8), 508–518. https://doi.org/10.1101/lm.2175811

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. *Advances in Neural Information Processing Systems*.

Cristea, I. A., & Naudet, F. (2019). Increase value and reduce waste in research on psychological therapies. *Preprint*, *March*, 1–35. https://doi.org/10.31219/osf.io/ps7x2

Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *Neuroscientist*, *8*(2), 132–142. https://doi.org/10.1177/107385840200800209

Das, R. K., Kamboj, S. K., Ramadas, M., Yogan, K., Gupta, V., Redman, E., Curran, H. V., & Morgan, C. J. A. (2013). Cannabidiol enhances consolidation of explicit fear extinction in humans. *Psychopharmacology*, *226*(4), 781–792. https://doi.org/10.1007/s00213-012-2955-y

Davis, M., & Astrachan, D. I. (1978). Conditioned fear and startle magnitude: Effects of different footshock or backshock intensities used in training. *Journal of Experimental Psychology: Animal Behavior Processes*, *4*(2), 95–103.

De Bitencourt, R. M., Pamplona, F. A., & Takahashi, R. N. (2013). A current overview of cannabinoids and glucocorticoids in facilitating extinction of aversive memories: Potential extinction enhancers. *Neuropharmacology*, *64*, 389–395. https://doi.org/10.1016/j.neuropharm.2012.05.039

de Haan, M. I. C., van Well, S., Visser, R. M., Scholte, H. S., van Wingen, G. A., & Kindt, M. (2018). The influence of acoustic startle probes on fear learning in humans. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-32646-1

De Kloet, E. R., Joëls, M., & Holsboer, F. (2005). Stress and the brain: From adaptation to disease. *Nature Reviews Neuroscience*, *6*(6), 463–475. https://doi.org/10.1038/nrn1683

De Quervain, D. J. F., Bentz, D., Michael, T., Bolt, O. C., Wiederhold, B. K., Margraf, J., & Wilhelm, F. H. (2011). Glucocorticoids enhance extinction-based psychotherapy. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(16), 6621–6625. https://doi.org/10.1073/pnas.1018214108

de Voogd, L. D., Fernández, G., & Hermans, E. J. (2016a). Awake reactivation of emotional memory traces through hippocampal-neocortical interactions. *NeuroImage*, *134*, 563–572. https://doi.org/10.1016/j.neuroimage.2016.04.026

de Voogd, L. D., Fernández, G., & Hermans, E. J. (2016b). Disentangling the roles of arousal and amygdala activation in emotional declarative memory. *Social Cognitive and Affective Neuroscience*, *11*(9), 1471–1480. https://doi.org/10.1093/scan/nsw055

de Voogd, L. D., Kanen, J. W., Neville, D. A., Roelofs, K., Fernández, G., & Hermans, E. J. (2018). Eye-movement intervention enhances extinction via amygdala deactivation. *Journal of Neuroscience*, *38*(40), 8694–8706. https://doi.org/10.1523/JNEUROSCI.0703-18.2018

Dębiec, J., & Ledoux, J. E. (2004). Disruption of reconsolidation but not consolidation of auditory fear conditioning by noradrenergic blockade in the amygdala. *Neuroscience*, *129*(2), 267–272. https://doi.org/10.1016/j.neuroscience.2004.08.018

Debiec, J., LeDoux, J. E., & Nader, K. (2002). Cellular and systems reconsolidation in the hippocampus. *Neuron*, *36*(3), 527–538. https://doi.org/10.1016/S0896-6273(02)01001-2

Delgado, M. R., Jou, R. L., LeDoux, J. E., & Phelps, E. A. (2009). Avoiding negative outcomes: Tracking the mechanisms of avoidance learning in humans during fear conditioning. *Frontiers in Behavioral Neuroscience*, *3*(NOV), 1–9. https://doi.org/10.3389/neuro.08.033.2009

Denniston, J. C., Chang, R. C., & Miller, R. R. (2003). Massive extinction treatment attenuates the renewal effect. *Learning and Motivation*, *34*(1), 68–86. https://doi.org/10.1016/S0023-9690(02)00508-8

Derksen, M. (2019). Putting Popper to work. *Theory and Psychology*, *29*(4), 449–465. https://doi.org/10.1177/0959354319838343

Devenport, L. D. (1998). Spontaneous recovery without interference: Why remembering is adaptive. *Animal Learning and Behavior*, *26*(2), 172–181. https://doi.org/10.3758/BF03199210

Díaz-Mataix, L., Ruiz Martinez, R. C., Schafe, G. E., Ledoux, J. E., & Doyère, V. (2013). Detection of a temporal error triggers reconsolidation of amygdala-dependent memories. *Current Biology*, *23*(6), 467–472. https://doi.org/10.1016/j.cub.2013.01.053

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, *10*(10), 1241–1242. https://doi.org/10.1038/nn1961

Dickinson, A., & Pearce, J. M. (1977). Inhibitory interactions between appetitive and aversive stimuli.

Psychological Bulletin, *84*(4), 690–711.

Diekhof, E. K., Kaps, L., Falkai, P., & Gruber, O. (2012). The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude - An activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and outcome processing. *Neuropsychologia*, *50*(7), 1252–1266. https://doi.org/10.1016/j.neuropsychologia.2012.02.007

Dittert, N., Hüttner, S., Polak, T., & Herrmann, M. J. (2018). Augmentation of fear extinction by transcranial direct current stimulation (tDCS). *Frontiers in Behavioral Neuroscience*, *12*(April), 1–16. https://doi.org/10.3389/fnbeh.2018.00076

Drexler, S. M., Merz, C. J., Hamacher-Dang, T. C., Marquardt, V., Fritsch, N., Otto, T., & Wolf, O. T. (2014). Effects of postretrieval-extinction learning on return of contextually controlled cued fear. *Behavioral Neuroscience*, *128*(4), 474–481. https://doi.org/10.1037/a0036688

Duncan, C. P. (1949). The retroactive effect of electroconvulsive shock. *Journal of Comparative and Physiological Psychology1*, *42*, 32–44.

Dunsmoor, J. E., Campese, V. D., Ceceli, A. O., LeDoux, J. E., & Phelps, E. A. (2015). Novelty-Facilitated Extinction: Providing a Novel Outcome in Place of an Expected Threat Diminshes Recovery of Defensive Responses. *Biological Psychiatry*, *78*(3), 203–209. https://doi.org/10.1016/j.biopsych.2014.12.008

Dunsmoor, J. E., Kragel, P. A., Martin, A., & La Bar, K. S. (2014). Aversive learning modulates cortical representations of object categories. *Cerebral Cortex*, *24*(11), 2859–2872. https://doi.org/10.1093/cercor/bht138

Dunsmoor, J. E., & Kroes, M. C. (2019). Episodic memory and Pavlovian conditioning: ships passing in the night. *Current Opinion in Behavioral Sciences*, *26*, 32–39. https://doi.org/10.1016/j.cobeha.2018.09.019

Dunsmoor, J. E., Kroes, M. C. W., Li, J., Daw, N. D., Simpson, H. B., & Phelps, E. A. (2019). Role of human ventromedial prefrontal cortex in learning and recall of enhanced extinction. *The Journal of Neuroscience*, *39*(17), 2713–2718. https://doi.org/10.1523/JNEUROSCI.2713-18.2019

Dunsmoor, J. E., Kroes, M. C. W., Moscatelli, C. M., Evans, M. D., Davachi, L., & Phelps, E. A. (2018). Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour*.

Dunsmoor, J. E., Kroes, M. C. W., Murty, V. P., Braren, S. H., & Phelps, E. A. (2019). Emotional enhancement of memory for neutral information: The complex interplay between arousal, attention, and anticipation. *Biological Psychology*, *145*, 134–141. https://doi.org/10.1016/j.biopsycho.2019.05.001

Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology*, *89*(2), 300–305. https://doi.org/10.1016/j.biopsycho.2011.11.002

Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, *520*(7547), 345–348. https://doi.org/10.1038/nature14106

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking Extinction. *Neuron*, *88*(1), 47–63. https://doi.org/10.1016/j.neuron.2015.09.028

Duvarci, S., & Nader, K. (2004). Characterization of fear memory reconsolidation. *Journal of*

Neuroscience, *24*(42), 9269–9275. https://doi.org/10.1523/JNEUROSCI.2971-04.2004

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Elahi, H., Hong, V., & Ploski, J. E. (2020). Electroconvulsive shock does not impair the reconsolidation of cued and contextual pavlovian threat memory. *International Journal of Molecular Sciences*, *21*(19), 1–14. https://doi.org/10.3390/ijms21197072

Elsey, J., & Kindt, M. (2016). Manipulating Human Memory Through Reconsolidation: Ethical Implications of a New Therapeutic Approach. *AJOB Neuroscience*, *7*(4), 225–236. https://doi.org/10.1080/21507740.2016.1218377

Elsey, J. W. B., & Kindt, M. (2017). Breaking boundaries: Optimizing reconsolidation-based interventions for strong and old memories. *Learning and Memory*, *24*(9), 472–479. https://doi.org/10.1101/lm.044156.116

Erler, A. (2011). Does memory modification threaten our authenticity? *Neuroethics*, *4*(3), 235–249. https://doi.org/10.1007/s12152-010-9090-4

Esser, R., Korn, C. W., Ganzer, F., & Haaker, J. (2021). L-DOPA modulates activity in the vmPFC , nucleus accumbens , and VTA during threat extinction learning in humans. *ELife*, 1–21.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Exton-McGuinness, M. T. J., Lee, J. L. C., & Reichelt, A. C. (2015). Updating memories-The role of prediction errors in memory reconsolidation. *Behavioural Brain Research*, *278*, 375–384. https://doi.org/10.1016/j.bbr.2014.10.011

Exton-McGuinness, M. T. J., Patton, R. C., Sacco, L. B., & Lee, J. L. C. (2014). Reconsolidation of a well-learned instrumental memory. *Learning and Memory*, *21*(9), 468–477. https://doi.org/10.1101/lm.035543.114

Fanselow, M. S. (1994). Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin & Review*, *1*(4), 429–438. https://doi.org/10.3758/BF03210947

Feng, P., Zheng, Y., & Feng, T. (2015). Spontaneous brain activity following fear reminder of fear conditioning by using resting-state functional MRI. *Scientific Reports*, *5*(November), 1–11. https://doi.org/10.1038/srep16701

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fernández, R. S., Boccia, M. M., & Pedreira, M. E. (2016). The fate of memory: Reconsolidation and the case of Prediction Error. *Neuroscience and Biobehavioral Reviews*, *68*, 423–441. https://doi.org/10.1016/j.neubiorev.2016.06.004

Finnie, P. S. B., & Nader, K. (2020). Amyloid Beta Secreted during Consolidation Prevents Memory Malleability. *Current Biology*, *30*(10), 1934-1940.e4. https://doi.org/10.1016/j.cub.2020.02.083

Flavell, C. R., Barber, D. J., & Lee, J. L. C. (2011). Behavioural memory reconsolidation of food and fear memories. *Nature Communications*, *2*(504). https://doi.org/10.1038/ncomms1515

Flexner, L. B., Flexner, J. . B., & Stellar, E. (1965). Memory and cerebral protein synthesis in mice as affected by graded amounts of puromycin. *Experimental Neurology*, *13*, 264–272.

Foa, E. B., Tolin, D. F., Ehlers, A., Clark, D. M., & Orsillo, S. M. (1999). The Posttraumatic Cognitions Inventory (PTCI): Development and validation. *Psychological Assessment*, *11*(3), 303–314. https://doi.org/10.1037/1040-3590.11.3.303

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683. https://doi.org/10.1038/nature04587

Frey, U., & Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, *385*, 533–536.

Fricchione, J., Greenberg, M. S., Spring, J., Wood, N., Mueller-Pfeiffer, C., Milad, M. R., Pitman, R. K., & Orr, S. P. (2016). Delayed extinction fails to reduce skin conductance reactivity to fear-conditioned stimuli. *Psychophysiology*, *53*(9), 1343–1351. https://doi.org/10.1111/psyp.12687

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

Fullana, Miquel A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., & Harrison, B. J. (2018). Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. *Neuroscience and Biobehavioral Reviews*, *88*(February), 16–25. https://doi.org/10.1016/j.neubiorev.2018.03.002

Galarza Vallejo, A., Kroes, M. C. W., Rey, E., Acedo, M. V., Moratti, S., Fernández, G., & Strange, B. A. (2019). Propofol-induced deep sedation reduces emotional episodic memory reconsolidation in humans. *Science Advances*, *5*(3). https://doi.org/10.1126/sciadv.aav3801

Gamache, K., Pitman, R. K., & Nader, K. (2012). Preclinical evaluation of reconsolidation blockade by clonidine as a potential novel treatment for posttraumatic stress disorder. *Neuropsychopharmacology*, *37*(13), 2789–2796. https://doi.org/10.1038/npp.2012.145

Gazarini, L., Stern, C. A. J., Piornedo, R. R., Takahashi, R. N., & Bertoglio, L. J. (2015). PTSD-like memory generated through enhanced noradrenergic activity is mitigated by a dual step pharmacological intervention targeting its reconsolidation. *International Journal of Neuropsychopharmacology*, *18*(1), 1–9. https://doi.org/10.1093/ijnp/pyu026

Gentile, C. G., Jarrell, T. W., Teich, A., McCabe, P. M., & Schneiderman, N. (1986). The role of amygdaloid central nucleus in the retention of differential Palovian conditioning of bradycardia in rabbits. *Behavioral Brain Reserch*, *20*, 263–273.

Gerber, M. M., & Jackson, J. (2013). Retribution as Revenge and Retribution as Just Deserts. *Social Justice Research*, *26*(1), 61–80. https://doi.org/10.1007/s11211-012-0174-7

Gerlicher, A. M. V., Tüscher, O., & Kalisch, R. (2018). Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06785-y

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, Learning, and Extinction. *Psychological Review*, *117*(1), 197–209. https://doi.org/10.1037/a0017808

Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning and Behavior*, *43*(3), 243–250. https://doi.org/10.3758/s13420-015-0176-z

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M.-H., & Niv, Y. (2013). Gradual extinction

prevents the return of fear: implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, *7*(November), 1–6. https://doi.org/10.3389/fnbeh.2013.00164

Gisquet-Verrier, P., Lynch, J. F., Cutolo, P., Toledano, D., Ulmen, A., Jasnow, A. M., & Riccio, D. C. (2015). Integration of new information with active memory accounts for retrograde amnesia: A challenge to the consolidation/reconsolidation hypothesis? *Journal of Neuroscience*, *35*(33), 11623–11633. https://doi.org/10.1523/JNEUROSCI.1386-15.2015

Gisquet-Verrier, P., & Riccio, D. C. (2018). Memory integration: An alternative to the consolidation/reconsolidation hypothesis. *Progress in Neurobiology*, *171*(September), 15–31. https://doi.org/10.1016/j.pneurobio.2018.10.002

Glover, G. H., Li, T. Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine*, *44*(1), 162–167. https://doi.org/10.1002/1522-2594(200007)44:1<162::AID-MRM23>3.0.CO;2-E

Gold, P. E., & Van Buskirk, R. B. (1975). Facilitation of time-dependent memory processes with posttrial epinephrine injections. *Behavioral Biology*, *13*(2), 145–153. https://doi.org/10.1016/S0091-6773(75)91784-8

Golkar, A., Bellander, M., Olsson, A., & Öhman, A. (2012). Are fear memories erasable? - reconsolidation of learned fear with fear relevant and fear-irrelevant stimuli. *Frontiers in Behavioral Neuroscience*, *6*(80). https://doi.org/10.3389/fnbeh.2012.00080

Gomperts, S. N., Kloosterman, F., & Wilson, M. A. (2015). VTA neurons coordinate with the hippocampal reactivation of spatial experience. *ELife*, *4*, 1–22. https://doi.org/10.7554/elife.05360

Goode, T. D., Holloway-Erickson, C. M., & Maren, S. (2017). Extinction after fear memory reactivation fails to eliminate renewal in rats. *Neurobiology of Learning and Memory*, *142*, 41–47. https://doi.org/10.1016/j.nlm.2017.03.001

Gordon, W. C., & Spear, N. E. (1973). Effect of reactivation of a previously acquired memory on the ineraction between memories in the rat. *Journal of Experimental Psychology*, *99*, 349–355.

Gräff, J., Joseph, N. F., Horn, M. E., Samiei, A., Meng, J., Seo, J., Rei, D., Bero, A. W., Phan, T. X., Wagner, F., Holson, E., Xu, J., Sun, J., Neve, R. L., Mach, R. H., Haggarty, S. J., & Tsai, L. H. (2014). Epigenetic priming of memory updating during reconsolidation to attenuate remote fear memories. *Cell*, *156*(1–2), 261–276. https://doi.org/10.1016/j.cell.2013.12.020

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2012). Mapping the Moral Domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847.Mapping

Gray, R., Budden-Potts, D., & Bourke, F. (2019). Reconsolidation of Traumatic Memories for PTSD: A randomized controlled trial of 74 male veterans. *Psychotherapy Research*, *29*(5), 621–639. https://doi.org/10.1080/10503307.2017.1408973

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology*, *91*(3), 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015

Grillon, C. (2009). D-Cycloserine Facilitation of Fear Extinction and Exposure-Based Therapy Might Rely on Lower-Level, Automatic Mechanisms. *Biological Psychiatry*, *66*(7), 636–641. https://doi.org/10.1016/j.biopsych.2009.04.017

Haaker, J., Golkar, A., Hermans, D., & Lonsdorf, T. B. (2014). A review on human reinstatement studies: An overview and methodological challenges. *Learning and Memory*, *21*(9), 424–440. https://doi.org/10.1101/lm.036053.114

Haaker, J., Maren, S., Andreatta, M., Merz, C. J., Richter, J., Richter, S. H., Meir Drexler, S., Lange, M. D., Jüngling, K., Nees, F., Seidenbecher, T., Fullana, M. A., Wotjak, C. T., & Lonsdorf, T. B. (2019). Making translation work: Harmonizing cross-species methodology in the behavioural neuroscience of Pavlovian fear conditioning. *Neuroscience and Biobehavioral Reviews*, *107*(July), 329–345. https://doi.org/10.1016/j.neubiorev.2019.09.020

Hagenaars, M. A., Oitzl, M., & Roelofs, K. (2014). Updating freeze: Aligning animal and human research. *Neuroscience and Biobehavioral Reviews*, *47*, 165–176. https://doi.org/10.1016/j.neubiorev.2014.07.021

Hardwicke, T. E., Taqi, M., & Shanks, D. R. (2016). Postretrieval new learning does not reliably induce human memory updating via reconsolidation. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(19), 5206–5211. https://doi.org/10.1073/pnas.1601440113

Haubrich, J., Bernabo, M., & Nader, K. (2020). Noradrenergic projections from the locus coeruleus to the amygdala constrain auditory fear memory reconsolidation. *ELife*, *9*(March), 1–29. https://doi.org/10.7554/eLife.57010

Haubrich, J., & Nader, K. (2018). Memory reconsolidation. *Current Topics in Behavioral Neurosciences*, *37*(17), 151–176. https://doi.org/10.1007/7854_2016_463

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience*, *11*(7), 523–532. https://doi.org/10.1038/nrn2850

Henry, M., Fishman, J. R., & Youngner, S. J. (2007). Propranolol and the prevention of post-traumatic stress disorder: Is it wrong to erase the "sting" of bad memories? *American Journal of Bioethics*, *7*(9), 12–20. https://doi.org/10.1080/15265160701518474

Hermans, E. J., Henckens, M. J. A. G., Roelofs, K., & Fernández, G. (2013). Fear bradycardia and activation of the human periaqueductal grey. *NeuroImage*, *66*, 278–287. https://doi.org/10.1016/j.neuroimage.2012.10.063

Hitchcock, J., & Davis, M. (1986). Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behavioral Neuroscience*, *100*(1), 11–22.

Hofstede, G. (2008). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage, Thousand Oaks.

Holehonnur, R., Phensy, A. J., Kim, L. J., Milivojevic, M., Vuong, D., Daison, D. K., Alex, S., Tiner, M., Jones, L. E., Kroener, S., & Ploski, J. E. (2016). Increasing the GluN2A/GluN2B ratio in neurons of the mouse basal and lateral amygdala inhibits the modification of an existing fear memory trace. *Journal of Neuroscience*, *36*(36), 9490–9504. https://doi.org/10.1523/JNEUROSCI.1743-16.2016

Holmes, E. A., James, E. L., Coode-Bate, T., & Deeprose, C. (2009). Can playing the computer game "tetris" reduce the build-up of flashbacks for trauma? A proposal from cognitive science. In *PLoS ONE* (Vol. 41, Issue 1). https://doi.org/10.1371/journal.pone.0004153

Houtekamer, M. C., Henckens, M. J. A. G., van den Berg, K. P., Homberg, J., & Kroes, M. C. W. (2021). *The impact of the intensity of aversive Pavlovian conditioning on the return of threat responses following a reminder-extinction procedure.* osf.io/n3bp4/

Houtekamer, Maxime C., Henckens, M. J. A. G., Mackey, W. E., Dunsmoor, J. E., Homberg, J. R., & Kroes, M. C. W. (2020). Investigating the efficacy of the reminder-extinction procedure to disrupt contextual threat memories in humans using immersive Virtual Reality. *Scientific Reports*, *10*(1), 1–18. https://doi.org/10.1038/s41598-020-73139-4

Hu, J., Wang, W., Homan, P., Wang, P., Zheng, X., & Schiller, D. (2018). Reminder duration determines threat memory modification in humans. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-27252-0

Hui, K., & Fisher, C. E. (2015). The ethics of molecular memory modification. *Journal of Medical Ethics*, *41*(7), 515–520. https://doi.org/10.1136/medethics-2013-101891

Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learning & Memory*, *14*, 47–53. https://doi.org/10.1101/lm.365707

Inda, M. C., Muravieva, E. V., & Alberini, C. M. (2011). *Memory Retrieval and the passage of time: from reconsolidation and Strengthening To Extinction*. *31*(5), 1635–1643. https://doi.org/10.1523/JNEUROSCI.4736-10.2011

Inslicht, S. S., Niles, A. N., Metzler, T. J., Lipshitz, S. L., Otte, C., Milad, M. R., Orr, S. P., Marmar, C. R., & Neylan, T. C. (2021). Randomized controlled experimental study of hydrocortisone and D-cycloserine effects on fear extinction in PTSD. *Neuropsychopharmacology*, *September*, 1–8. https://doi.org/10.1038/s41386-021-01222-z

Ishii, D., Matsuzawa, D., Matsuda, S., Tomizawa, H., Sutoh, C., & Shimizu, E. (2012). No erasure effect of retrieval-extinction trial on fear memory in the hippocampus-independent and dependent paradigms. *Neuroscience Letters*, *523*(1), 76–81. https://doi.org/10.1016/j.neulet.2012.06.048

Ishii, D., Matsuzawa, D., Matsuda, S., Tomizawa, H., Sutoh, C., & Shimizu, E. (2015). An isolated retrieval trial before extinction session does not prevent the return of fear. *Behavioural Brain Research*, *287*, 139–145. https://doi.org/10.1016/j.bbr.2015.03.052

Isserles, M., Shalev, A. Y., Roth, Y., Peri, T., Kutz, I., Zlotnick, E., & Zangen, A. (2013). Effectiveness of deep transcranial magnetic stimulation combined with a brief exposure procedure in post-traumatic stress disorder-a pilot study. *Brain Stimulation*, *6*(3), 377–383. https://doi.org/10.1016/j.brs.2012.07.008

Iwata, J., Chida, K., & LeDoux, J. E. (1987). Cardiovascular responses elicited by stimulation of neurons in the central amygdaloid nucleus in awake but not anesthetized rats resemble conditioned emotional responses. *Brain Research*, *418*(1), 183–188. https://doi.org/10.1016/0006-8993(87)90978-4

James, E. L., Bonsall, M. B., Hoppitt, L., Tunbridge, E. M., Geddes, J. R., Milton, A. L., & Holmes, E. A. (2015). Computer Game Play Reduces Intrusive Memories of Experimental Trauma via Reconsolidation-Update Mechanisms. *Psychological Science*, *26*(8), 1201–1215. https://doi.org/10.1177/0956797615583071

Johansen, J. P., Cain, C. K., Ostroff, L. E., & Ledoux, J. E. (2011). Molecular mechanisms of fear learning and memory. *Cell*, *147*(3), 509–524. https://doi.org/10.1016/j.cell.2011.10.009

Johnson, D. C., & Casey, B. J. (2015a). Extinction during memory reconsolidation blocks recovery of fear in adolescents. *Scientific Reports*, *5*, 1–5. https://doi.org/10.1038/srep08863

Johnson, D. C., & Casey, B. J. (2015b). Extinction during memory reconsolidation blocks recovery of fear in adolescents. *Scientific Reports*, *5*, 1–5. https://doi.org/10.1038/srep08863

Jones, C. E., & Monfils, M. H. (2016). Post-retrieval extinction in adolescence prevents return of juvenile fear. *Learning and Memory*, *23*(10), 567–575. https://doi.org/10.1101/lm.043281.116

Jones, C. E., Ringuet, S., & Monfils, M. H. (2013). Learned together, extinguished apart: Reducing fear to complex stimuli. *Learning and Memory*, *20*(12), 674–685. https://doi.org/10.1101/lm.031740.113

Josselyn, S. A., & Tonegawa, S. (2020). Memory engrams: Recalling the past and imagining the future. *Science*, *367*(6473). https://doi.org/10.1126/science.aaw4325

Jozefowiez, J., Berruti, A. S., Moshchenko, Y., Peña, T., Polack, C. W., & Miller, R. R. (2020). Retroactive interference: Counterconditioning and extinction with and without biologically significant outcomes. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(4), 443–459. https://doi.org/10.1037/xan0000272

Junjiao, L., Wei, C., Jingwen, C., Yanjian, H., Yong, Y., Liang, X., Jing, J., & Xifu, Z. (2019). Role of prediction error in destabilizing fear memories in retrieval extinction and its neural mechanisms. *Cortex*, *121*, 292–307. https://doi.org/10.1016/j.cortex.2019.09.003

Kang, S., Vervliet, B., Engelhard, I. M., van Dis, E. A. M., & Hagenaars, M. A. (2018). Reduced return of threat expectancy after counterconditioning verus extinction. *Behaviour Research and Therapy*, *108*(January), 78–84. https://doi.org/10.1016/j.brat.2018.06.009

Kass, L. R. (2003). *Beyond Therapy: Biotechnology and the Pursuit of Human Improvement*. The President's Council on Bioethics. https://bioethicsarchive.georgetown.edu/pcbe/background/kasspaper.html

Keller, N. E., & Dunsmoor, J. E. (2020). The effects of aversive-to-appetitive counterconditioning on implicit and explicit fear memory. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *27*(1), 12–19. https://doi.org/10.1101/lm.050740.119

Keller, N. E., Hennings, A. C., & Dunsmoor, J. E. (2020). Behavioral and neural processes in counterconditioning: Past and future directions. *Behaviour Research and Therapy*, *125*(December 2019). https://doi.org/10.1016/j.brat.2019.103532

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime Prevalence and Age-of-Onset Distributions of. *Arch Gen Psychiatry*, *62*(June), 593–602. http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archpsyc.62.6.593

Kilpatrick, D. G., Resnick, H. S., Milanak, M. E., Miller, M. W., Keyes, katherine M., & Friedman, M. J. (2013). National Estimates of Exposure to Traumatic Events and PTSD Prevalence Using DSM-IV and DSM-5 Criteria. *J Trauma Stress*, *26*(5), 537–547. https://doi.org/10.1002/jts.21848.National

Kim, J., Pignatelli, M., Xu, S., Itohara, S., & Tonegawa, S. (2016). Antagonistic negative and positive neurons of the basolateral amygdala. *Nature Neuroscience*, *19*(12), 1636–1646. https://doi.org/10.1038/nn.4414

Kim, R., Moki, R., & Kida, S. (2011). Molecular mechanisms for the destabilization and restabilization of reactivated spatial memory in the Morris water maze. *Molecular Brain*, *4*(9). https://doi.org/10.1186/1756-6606-4-9

Kim, S. C., Jo, Y. S., Kim, I. H., Kim, H., & Choi, J. S. (2010). Lack of medial prefrontal cortex activation underlies the immediate extinction deficit. *Journal of Neuroscience*, *30*(3), 832–837. https://doi.org/10.1523/JNEUROSCI.4145-09.2010

Kindt, M., & Soeter, M. (2013). Reconsolidation in a human fear conditioning study: A test of

extinction as updating mechanism. *Biological Psychology*, *92*(1), 43–50. https://doi.org/10.1016/j.biopsycho.2011.09.016

Kindt, M., Soeter, M., & Vervliet, B. (2009). Beyond extinction: erasing human fear responses and preventing the return of fear. *Nature Neuroscience*, *12*(3), 256–258. https://doi.org/10.1038/nn.2271

Kleinsmith, L. J., & Kaplan, S. (1963). Paired-associate learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology*, *65*, 190–193.

Klucken, T., Kruse, O., Schweckendiek, J., Kuepper, Y., Mueller, E. M., Hennig, J., & Stark, R. (2016). No evidence for blocking the return of fear by disrupting reconsolidation prior to extinction learning. *Cortex*, *79*, 112–122. https://doi.org/10.1016/j.cortex.2016.03.015

Klumpers, F., Kroes, M. C., Heitland, I., Everaerd, D., Akkermans, S. E. A., Oosting, R. S., Wingen, G. Van, Franke, B., Kenemans, J. L., Fernández, G., & Baas, J. M. P. (2015). Dorsomedial Prefrontal Cortex Mediates the Impact of Serotonin Transporter Linked Polymorphic Region Genotype on Anticipatory Threat Reactions. *Biological Psychiatry*, *78*(8), 582–589. https://doi.org/10.1016/j.biopsych.2014.07.034

Klumpers, F., Kroes, M. C. W., Baas, J., & Fernández, G. (2017). How human amygdala and bed nucleus of the stria terminalis may drive distinct defensive responses. *The Journal of Neuroscience*, *37*(40), 9645–9656. https://doi.org/10.1523/JNEUROSCI.3830-16.2017

Klumpers, F., Morgan, B., Terburg, D., Stein, D. J., & van Honk, J. (2014). Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. *Social Cognitive and Affective Neuroscience*, *10*(9), 1161–1168. https://doi.org/10.1093/scan/nsu164

Klumpers, F., Morgan, B., Terburg, D., Stein, D. J., & van Honk, J. (2015). Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. *Social Cognitive and Affective Neuroscience*, *10*(9), 1161–1168. https://doi.org/10.1093/scan/nsu164

Klüver, H., & Bucy, P. C. (1937). “Psychic blindness” and other symptoms following bilateral temporal lobectomy in Rhesus monkeys. *American Journal of Physiology*, *119*, 352–353.

Knipscheer, J., Sleijpen, M., Frank, L., de Graaf, R., Kleber, R., Ten Have, M., & Dückers, M. (2020). Prevalence of potentially traumatic events, other life events and subsequent reactions indicative for posttraumatic stress disorder in the netherlands: A general population study based on the trauma screening questionnaire. *International Journal of Environmental Research and Public Health*, *17*(5). https://doi.org/10.3390/ijerph17051725

Knutson, B., & Cooper, J. C. (2005). Functional magnetic resonance imaging of reward prediction. *Current Opinion in Neurology*, *18*(4), 411–417. https://doi.org/10.1097/01.wco.0000173463.24758.f6

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. *NeuroImage*, *12*(1), 20–27. https://doi.org/10.1006/nimg.2000.0593

Koizumi, A., Amano, K., Cortese, A., Shibata, K., Yoshida, W., Seymour, B., Kawato, M., & Lau, H. (2017). Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour*, *1*(1), 1–7. https://doi.org/10.1038/s41562-016-0006

Krawczyk, M. C., Fernández, R. S., Pedreira, M. E., & Boccia, M. M. (2017). Toward a better

understanding on the role of prediction error on memory processes: From bench to clinic. *Neurobiology of Learning and Memory*, *142*, 13–20. https://doi.org/10.1016/j.nlm.2016.12.011

Kredlow, M. A., Orr, S. P., & Otto, M. W. (2018). Exploring the boundaries of post-retrieval extinction in healthy and anxious individuals. *Behaviour Research and Therapy*, *108*(June), 45–57. https://doi.org/10.1016/j.brat.2018.06.010

Kredlow, M. A., Unger, L. D., & Otto, M. W. (2015). Harnessing Reconsolidation to weaken fear and appetitive memories: A meta-analysis. *Psychological Bulletin*, *142*(4), 314–336.

Kredlow, M. A., Unger, L. D., & Otto, M. W. (2016). Harnessing Reconsolidation to weaken fear and appetitive memories: A meta-analysis. *Psychological Bulletin*, *142*(3), 314–336. https://doi.org/10.1037/bul0000034

Kroes, M. C., Schiller, D., Ledoux, J. E., & Phelps, E. A. (2016). Translational approaches targeting reconsolidation. *Current Topics in Behavioral Neurosciences*, *28*, 197–230. https://doi.org/10.1007/7854_2015_5008

Kroes, M. C. W., Dunsmoor, J. E., Lin, Q., Evans, M., & Phelps, E. A. (2017). A reminder before extinction strengthens episodic memory via reconsolidation but fails to disrupt generalized threat responses. *Scientific Reports*, *7*(1), 10858. https://doi.org/10.1038/s41598-017-10682-7

Kroes, M. C. W., Dunsmoor, J. E., Mackey, W. E., McClay, M., & Phelps, E. A. (2017). Context conditioning in humans using commercially available immersive Virtual Reality. *Scientific Reports*, *7*(1), 8640. https://doi.org/10.1038/s41598-017-08184-7

Kroes, M. C. W., & Fernández, G. (2012). Dynamic neural systems enable adaptive, flexible memories. *Neuroscience and Biobehavioral Reviews*, *36*(7), 1646–1666. https://doi.org/10.1016/j.neubiorev.2012.02.014

Kroes, M. C. W., & Liivoja, R. (2018). Eradicating war memories: Neuroscientific reality and ethical concerns. *International Review of the Red Cross*, 1–27. https://doi.org/10.1017/s1816383118000437

Kroes, M. C. W., Strange, B. A., & Dolan, R. J. (2010). B-Adrenergic Blockade During Memory Retrieval in Humans Evokes a Sustained Reduction of Declarative Emotional Memory Enhancement. *Journal of Neuroscience*, *30*(11), 3959–3963. https://doi.org/10.1523/JNEUROSCI.5469-09.2010

Kroes, M. C. W., Tendolkar, I., Van Wingen, G. A., Van Waarde, J. A., Strange, B. A., & Fernández, G. (2014). An electroconvulsive therapy procedure impairs reconsolidation of episodic memories in humans. *Nature Neuroscience*, *17*(2), 204–206. https://doi.org/10.1038/nn.3609

Kroes, M. C. W., Tona, K. D., Den Ouden, H. E. M., Vogel, S., Van Wingen, G. A., & Fernández, G. (2016). How administration of the beta-blocker propranolol before extinction can prevent the return of fear. *Neuropsychopharmacology*, *41*(6), 1569–1578. https://doi.org/10.1038/npp.2015.315

Kwak, C., Choi, J. H., Bakes, J. T., Lee, K., & Kaang, B. K. (2012). Effect of intensity of unconditional stimulus on reconsolidation of contextual fear memory. *Korean Journal of Physiology and Pharmacology*, *16*(5), 293–296. https://doi.org/10.4196/kjpp.2012.16.5.293

LaBar, K. S., & Phelps, E. A. (1998). Arousal-mediated memory consolidation: Role of the Medial Temporal Lobe in Humans. *Psychological Science*, *9*(6), 490–493. https://doi.org/10.1111/1467-9280.00090

LaBar, Kevin S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54–64. https://doi.org/10.1038/nrn1825

LaBar, Kevin S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: A mixed-trial fMRI study. *Neuron*, *20*(5), 937–945. https://doi.org/10.1016/S0896-6273(00)80475-4

Lang, P.J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*(3), 377–395.

Lang, P.J., Geenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*(3), 261–273. https://doi.org/10.1111/j.1469-8986.1993.tb03352.x

Lang, Peter J. (1977). Imagery in therapy: an information processing analysis of fear. *Behavior Therapy*, *8*(5), 862–886. https://doi.org/10.1016/S0005-7894(77)80157-3

Lattal, K. M., & Abel, T. (2004). Behavioral impairments caused by injections of the protein synthesis inhibitor anisomycin after contextual retrieval reverse with time. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(13), 4667–4672. https://doi.org/10.1073/pnas.0306546101

Lavazza, A. (2015). Erasing traumatic memories: When context and social interests can outweigh personal autonomy. *Philosophy, Ethics, and Humanities in Medicine*, *10*(1), 1–7. https://doi.org/10.1186/s13010-014-0021-6

Ledoux, J. (2003). The Emotional Brain, Fear, and the Amygdala. *Cellular and Molecular Neurobiology*, *23*(4–5), 727–738. https://doi.org/https://doi.org/10.1023/A:1025048802629

Ledoux, J. E. (1997). Emotional memory and psychopathology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1362), 1719–1726. https://doi.org/10.1098/rstb.1997.0154

LeDoux, J. E. (2009). Emotion Circuits in the Brain. *Focus*, *7*(2), 274–274. https://doi.org/10.1176/foc.7.2.foc274

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(8), 2871–2878. https://doi.org/10.1073/pnas.1400335111

LeDoux, J. E., & Pine, D. S. (2016). Using neuroscience to help understand fear and anxiety: A two-system framework. *American Journal of Psychiatry*, *173*(11), 1083–1093. https://doi.org/10.1176/appi.ajp.2016.16030353

Lee, J. L. C. (2008). Memory reconsolidation mediates the strengthening of memories by additional learning. *Nature Neuroscience*, *11*(11), 1264–1266. https://doi.org/10.1038/nn.2205

Lee, J. L. C., Amorim, F. E., Cassini, L. F., & Amaral, O. B. (2019). Different temporal windows for CB1 receptor involvement in contextual fear memory destabilisation in the amygdala and hippocampus. *PLoS ONE*, *14*(1), 1–16. https://doi.org/10.1371/journal.pone.0205781

Lewis, D. J. (1979). Psychobiology of active and inactive memory. *Psychological Bulletin*, *86*(5), 1054–1083. https://doi.org/10.1037/0033-2909.86.5.1054

Liang, K. C., Juler, R. G., & McGaugh, J. L. (1986). Modulating effects of posttraining epinephrine on memory: involvement of the amygdala noradrenergic system. *Brain Research*, *368*(1), 125–133. https://doi.org/10.1016/0006-8993(86)91049-8

Liang, X., Zou, Q., He, Y., & Yang, Y. (2016). Topologically Reorganized Connectivity Architecture of Default-Mode, Executive-Control, and Salience Networks across Working Memory Task Loads. *Cerebral Cortex*, *26*(4), 1501–1511. https://doi.org/10.1093/cercor/bhu316

Liao, S. M., & Sandberg, A. (2008). The Normativity of Memory Modification. *Neuroethics*, *1*(2), 85–

99. https://doi.org/10.1007/s12152-008-9009-5

Liao, S. M., & Wasserman, D. T. (2007). Neuroethical concerns about moderating traumatic memories. *American Journal of Bioethics*, *7*(9), 38–40. https://doi.org/10.1080/15265160701518623

Liu, J., Zhao, L., Xue, Y., Shi, J., Suo, L., Luo, Y., Chai, B., Yang, C., Fang, Q., Zhang, Y., Bao, Y., Pickens, C. L., & Lu, L. (2014). An unconditioned stimulus retrieval extinction procedure to prevent the return of fear memory. *Biological Psychiatry*, *76*(11), 895–901. https://doi.org/10.1016/j.biopsych.2014.03.027

Livermore, J. J. A., Klaassen, F. H., Bramson, B., Hulsman, A. M., Meijer, S. W., Held, L., Klumpers, F., de Voogd, L. D., & Roelofs, K. (2021). Approach-Avoidance Decisions Under Threat: The Role of Autonomic Psychophysiological States. *Frontiers in Neuroscience*, *15*(March), 1–12. https://doi.org/10.3389/fnins.2021.621517

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

Lonsdorf, T. B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., Jentsch, V. L., Drexler, S. M., Mertens, G., Richter, J., Sjouwerman, R., Wendt, J., & Merz, C. J. (2019). Navigating the garden of forking paths for data exclusions in fear conditioning research. *ELife*, *8*, 1–70. https://doi.org/10.7554/eLife.52465

Luck, C. C., & Lipp, O. V. (2018). Verbal instructions targeting valence alter negative conditional stimulus evaluations (but do not affect reinstatement rates). *Cognition and Emotion*, *32*(1), 61–80. https://doi.org/10.1080/02699931.2017.1280449

Lupien, S. J., & McEwen, B. S. (1997). The acute effects of corticosteroids on cognition: Integration of animal and human model studies. *Brain Research Reviews*, *24*(1), 1–27. https://doi.org/10.1016/S0165-0173(97)00004-0

Luyten, L., & Beckers, T. (2017). A preregistered, direct replication attempt of the retrieval-extinction effect in cued fear conditioning in rats. *Neurobiology of Learning and Memory*, *144*, 208–215. https://doi.org/10.1016/j.nlm.2017.07.014

Luyten, L., Schnell, A. E., Schroyens, N., & Beckers, T. (2021). Lack of drug-induced post-retrieval amnesia for auditory fear memories in rats. *BMC Biology*, *19*(1), 1–15. https://doi.org/10.1186/s12915-021-00957-x

MacPherson, K., Whittle, N., Camp, M., Gunduz-Cinar, O., Singewald, N., & Holmes, A. (2013). Temporal factors in the extinction of fear in inbred mouse strains differing in extinction efficacy. *Biology of Mood & Anxiety Disorders*, *3*(1), 13. https://doi.org/10.1186/2045-5380-3-13

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*(3), 1233–1239. https://doi.org/10.1016/S1053-8119(03)00169-1

Mamou, C. Ben, Gamache, K., & Nader, K. (2006). NMDA receptors are critical for unleashing consolidated auditory fear memories. *Nature Neuroscience*, *9*(10), 1237–1239. https://doi.org/10.1038/nn1778

Maren, S. (2014). Nature and causes of the immediate extinction deficit: A brief review. *Neurobiology of Learning and Memory*, *113*, 19–24. https://doi.org/10.1016/j.nlm.2013.10.012

Maren, S., & Chang, C. H. (2006). Recent fear is resistant to extinction. *Proceedings of the National*

Academy of Sciences of the United States of America, *103*(47), 18020–18025. https://doi.org/10.1073/pnas.0608398103

Maren, S., & Holmes, A. (2016). Stress and fear extinction. *Neuropsychopharmacology*, *41*(1), 58–79. https://doi.org/10.1038/npp.2015.180

Maren, S., Phan, K. L., & Liberzon, I. (2013a). The contextual brain: Implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience*, *14*(6), 417–428. https://doi.org/10.1038/nrn3492

Maren, S., Phan, K. L., & Liberzon, I. (2013b). The contextual brain: Implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience*, *14*(6), 417–428. https://doi.org/10.1038/nrn3492

Marschner, A., Kalisch, R., Vervliet, B., Vansteenwegen, D., & Büchel, C. (2008). Dissociable roles for the hippocampus and the amygdala in human cued versus context fear conditioning. *Journal of Neuroscience*, *28*(36), 9030–9036. https://doi.org/10.1523/JNEUROSCI.1651-08.2008

McFarquhar, M., McKie, S., Emsley, R., Suckling, J., Elliott, R., & Williams, S. (2016). Multivariate and repeated measures (MRM): A new toolbox for dependent and multimodal group-level neuroimaging data. *NeuroImage*, *132*, 373–389. https://doi.org/10.1016/j.neuroimage.2016.02.053

McGaugh, J. L. (2000). Memory - A century of consolidation. *Science*, *287*(5451), 248–251. https://doi.org/10.1126/science.287.5451.248

Meir Drexler, S., Merz, C. J., Hamacher-Dang, T. C., Marquardt, V., Fritsch, N., Otto, T., & Wolf, O. T. (2014). Effects of postretrieval-extinction learning on return of contextually controlled cued fear. *Behavioral Neuroscience*, *128*(4), 474–481. https://doi.org/10.1037/a0036688

Merlo, E., Milton, A. L., Goozee, Z. Y., Theobald, D. E., & Everitt, B. J. (2014). Reconsolidation and Extinction Are Dissociable and Mutually Exclusive Processes: Behavioral and Molecular Evidence. *Journal of Neuroscience*, *34*(7), 2422–2431. https://doi.org/10.1523/JNEUROSCI.4001-13.2014

Milad, M. R., Orr, S. P., Lasko, N. B., Chang, Y., Rauch, S. L., & Pitman, R. K. (2008). Presence and acquired origin of reduced recall for fear extinction in PTSD: Results of a twin study. *Journal of Psychiatric Research*, *42*(7), 515–520. https://doi.org/10.1016/j.jpsychires.2008.01.017

Milad, M. R., Pitman, R. K., Ellis, C. B., Gold, A. L., Shin, L. M., Lasko, N. B., Zeidan, M. a, Orr, S. P., & Rauch, S. L. (2009). Neurobiological Basis of Failure to Recall Extinction Memory in Posttraumatic Stress Disorder. *Bioligical Psychiatry*, *66*(12), 1075–1082. https://doi.org/10.1016/j.biopsych.2009.06.026.Neurobiological

Milad, M. R., & Quirk, G. J. (2012). Fear Extinction as a Model for Translational Neuroscience: Ten Years of Progress. *Ssrn*. https://doi.org/10.1146/annurev.psych.121208.131631

Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., & Rauch, S. L. (2007). Recall of Fear Extinction in Humans Activates the Ventromedial Prefrontal Cortex and Hippocampus in Concert. *Biological Psychiatry*, *62*(5), 446–454. https://doi.org/10.1016/j.biopsych.2006.10.011

Milekic, M. H., & Alberini, C. M. (2002). Temporally graded requirement for protein synthesis following memory reactivation. *Neuron*, *36*(3), 521–525. https://doi.org/10.1016/S0896-6273(02)00976-5

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. In *Psychological Bulletin* (Vol. 117, Issue 3, pp. 363–386). https://doi.org/10.1037//0033-

2909.117.3.363

Milton, A. L., & Everitt, B. J. (2010). The psychological and neurochemical mechanisms of drug memory reconsolidation: Implications for the treatment of addiction. *European Journal of Neuroscience*, *31*(12), 2308–2319. https://doi.org/10.1111/j.1460-9568.2010.07249.x

Milton, A. L., Merlo, E., Ratano, P., Gregory, B. L., Dumbreck, J. K., & Everitt, B. J. (2013). Double dissociation of the requirement for GluN2B-and GluN2A-containing NMDA receptors in the destabilization and restabilization of a reconsolidating memory. *Journal of Neuroscience*, *33*(3), 1109–1115. https://doi.org/10.1523/JNEUROSCI.3273-12.2013

Misanin, J. R., Miller, R. R., & Lewis, D. J. (1968). Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science*, *160*, 203–204.

Monfils, M.-H., Cowansage, K. K., Klann, E., & LeDoux, J. E. (2009). Extinction-Reconsolidation Boundaries: Key to Persistent Attenuation of Fear Memories. *Science*, *324*(5929), 951–955. https://doi.org/10.1126/science.1167975

Monfils, M. H., Cowansage, K. K., Klann, E., & Ledoux, J. E. (2009). Extinction-Reconsolidation boundaries: Key to persistent attenuation of fear memories. *Science*, *324*(5929), 951–955. https://doi.org/10.1126/science.1167975

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947. https://doi.org/10.1523/jneurosci.16-05-01936.1996

Monti, R. I. F., Alfei, J. M., Mugnaini, M., Bueno, A. M., Beckers, T., Urcelay, G. P., & Molina, V. A. (2017). A comparison of behavioral and pharmacological interventions to attenuate reactivated fear memories. *Learning and Memory*, *24*(8), 369–374. https://doi.org/10.1101/lm.045385.117

Morgan, M. A., Romanski, L. M., & LeDoux, J. E. (1993). Extinction of emotional learning: Contribution of medial prefrontal cortex. *Neuroscience Letters*, *163*(1), 109–113. https://doi.org/10.1016/0304-3940(93)90241-C

Morris, R. G. M., Inglis, J., Ainge, J. A., Olverman, H. J., Tulloch, J., Dudai, Y., & Kelly, P. A. T. (2006). Memory Reconsolidation: Sensitivity of Spatial Memory to Inhibition of Protein Synthesis in Dorsal Hippocampus during Encoding and Retrieval. *Neuron*, *50*(3), 479–489. https://doi.org/10.1016/j.neuron.2006.04.012

Mulkay, M., & Gilbert, G. N. (1981). Putting Philosophy to Work: Karl Popper's Influence on Scientific Practice. *Philosophy of the Social Sciences*, *11*(3), 389–407. https://doi.org/10.1177/004839318101100306

Myers, K. M., & Davis, M. (2002). Behavioral and Neural Analysis of Extinction. *Neuron*, *36*(4), 567–584.

Myers, K. M., Ressler, K. J., & Davis, M. (2006). Different mechanisms of fear extinction dependent on length of time since fear acquisition. *Learning and Memory*, *13*(2), 216–223. https://doi.org/10.1101/lm.119806

Nadel, L., & Willner, J. (1980). Context and conditioning: A place for space. *Physiological Psychology*, *8*(2), 218–228. https://doi.org/10.3758/BF03332853

Nader, K. (2003). Memory traces unbound. *Trends in Neurosciences*, *26*(2), 65–72. https://doi.org/10.1016/S0166-2236(02)00042-5

Nader, K., & Hardt, O. (2009). A single standard for memory; the case for reconsolidation. *Nature Neuroscience Reviews*, *10*(3), 224–234. https://doi.org/10.1038/nrn2590

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, *406*(6797), 722–726. https://doi.org/10.1038/35021052

Nagelsen, S., & Huckelbury, C. (1969). Abrogation of the Therapeutic Model in Prison Health Care and the Implications for Public Safety. *Journal of Prisoners on Prisons*, *17*(2), 16–27. https://doi.org/10.18192/jpp.v17i2.5241

Newall, C., Watson, T., Grant, K. A., & Richardson, R. (2017). The relative effectiveness of extinction and counter-conditioning in diminishing children's fear. *Behaviour Research and Therapy*, *95*, 42–49. https://doi.org/10.1016/j.brat.2017.05.006

Newman, E. J., Berkowitz, S. R., Nelson, K. J., Garry, M., & Loftus, E. F. (2011). Attitudes about memory dampening drugs depend on context and country. *Applied Cognitive Psychology*, *25*(5), 675–681. https://doi.org/10.1002/acp.1740

Nosek, B. A. (2015). Promoting an open research culture: The TOP guidelines. *Science*, *348*(6242), 1422–1425.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311–345. https://doi.org/10.1037/0033-295X.108.2.311

Olshavsky, M. E., Jones, C. E., Lee, H. J., & Monfils, M. H. (2013). Appetitive behavioral traits and stimulus intensity influence maintenance of conditioned fear. *Frontiers in Behavioral Neuroscience*, *7*(DEC), 1–7. https://doi.org/10.3389/fnbeh.2013.00179

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*. https://doi.org/10.1155/2011/156869

Ougrin, D. (2011). Efficacy of exposure versus cognitive therapy in anxiety disorders: Systematic review and meta-analysis. *BMC Psychiatry*, *11*(December). https://doi.org/10.1186/1471-244X-11-200

Oyarzún, J. P., Lopez-Barroso, D., Fuentemilla, L., Cucurell, D., Pedraza, C., Rodriguez-Fornells, A., & de Diego-Balaguer, R. (2012). Updating fearful memories with extinction training during reconsolidation: A human study using auditory aversive stimuli. *PLoS ONE*, *7*(6). https://doi.org/10.1371/journal.pone.0038849

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parens, E. (2010a). The ethics of memory blunting: Some initial thoughts. *Frontiers in Behavioral Neuroscience*, *4*(DEC), 190. https://doi.org/10.3389/fnbeh.2010.00190

Parens, E. (2010b). The ethics of memory blunting and the narcissism of small differences. *Neuroethics*, *3*(2), 99–107. https://doi.org/10.1007/s12152-010-9070-8

Parsons, R. G., & Ressler, K. J. (2013). Implications of memory modulation for post-traumatic stress and fear disorders. In *Nature Neuroscience* (Vol. 16, Issue 2, pp. 146–153). https://doi.org/10.1038/nn.3296

Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning and Memory*, *24*(1), 65–69. https://doi.org/10.1101/lm.042978.116

Pattwell, S. S., Duhoux, S., Hartley, C. A., Johnson, D. C., Jing, D., Elliott, M. D., Ruberry, E. J., Powers,

A., Mehta, N., Yang, R. R., Soliman, F., Glatt, C. E., Casey, B. J., Ninan, I., & Lee, F. S. (2012). Altered fear learning across development in both mouse and human. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(40), 16318–16323. https://doi.org/10.1073/pnas.1206834109

Pattwell, S. S., Liston, C., Jing, D., Ninan, I., Yang, R. R., Witztum, J., Murdock, M. H., Dincheva, I., Bath, K. G., Casey, B. J., Deisseroth, K., & Lee, F. S. (2016). Dynamic changes in neural circuitry during adolescence are associated with persistent attenuation of fear memories. *Nature Communications*, *7*(May), 1–9. https://doi.org/10.1038/ncomms11475

Pearce, J. M., & Dickinson, A. (1975). Pavlovian countercondition: Changing the suppressive properties of shock by association with food. *Animal Behavior Processes*, *1*(2), 170–177.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552.

Pedreira, M. E. (2004). Mismatch Between What Is Expected and What Actually Occurs Triggers Memory Reconsolidation or Extinction. *Learning & Memory*, *11*(5), 579–585. https://doi.org/10.1101/lm.76904

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

Phelps, E. A., & Hofmann, S. G. (2019). Memory editing from science fiction to clinical practice. *Nature*, *572*(7767), 43–50. https://doi.org/10.1038/s41586-019-1433-7

Phelps, E. A., Ledoux, J. E., Place, W., & Place, W. (2005). *Contributions of the Amygdala to Emotion Processing : From Animal Models to Human Behavior*. *48*, 175–187. https://doi.org/10.1016/j.neuron.2005.09.025

Phillips, R. G., & LeDoux, J. E. (1992). Differential Contribution of Amygdala and Hippocampus to Cued and Contextual Fear Conditioning. *Behavioral Neuroscience*, *106*(2), 274–285. https://doi.org/10.1037/0735-7044.106.2.274

Piñeyro, M. E., Monti, R. I. F., Alfei, J. M., Bueno, A. M., & Urcelay, G. P. (2014). Memory destabilization is critical for the success of the reactivation-extinction procedure. *Learning and Memory*, *21*(1), 46–54. https://doi.org/10.1101/lm.032714.113

Pitman, R. K., Sanders, K. M., Zusman, R. M., Healy, A. R., Cheema, F., Lasko, N. B., Cahill, L., & Orr, S. P. (2002). Pilot study of secondary prevention of posttraumatic stress disorder with propranolol. *Biological Psychiatry*, *51*(2), 189–192. https://doi.org/10.1016/S0006-3223(01)01279-3

Ponnusamy, R., Zhuravka, I., Poulos, A. M., Shobe, J., Merjanian, M., Huang, J., Wolvek, D., O'Neill, P. K., & Fanselow, M. S. (2016). Retrieval and reconsolidation accounts of fear extinction. *Frontiers in Behavioral Neuroscience*, *10*(MAY), 1–11. https://doi.org/10.3389/fnbeh.2016.00089

Poser, B. A., Versluis, M. J., Hoogduin, J. M., & Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. *Magnetic Resonance in Medicine*, *55*(6), 1227–1235. https://doi.org/10.1002/mrm.20900

Prebble, S. C., Addis, D. R., & Tippett, L. J. (2013). Autobiographical memory and sense of self. *Psychological Bulletin*, *139*(4), 815–840. https://doi.org/10.1037/a0030146

Price, R. B., Paul, B., Schneider, W., & Siegle, G. J. (2013). Neural correlates of three neurocognitive

intervention strategies: A preliminary step towards personalized treatment for psychological disorders. *Cognitive Therapy and Research*, *37*(4), 657–672. https://doi.org/10.1007/s10608-012-9508-x

Przybyslawski, J., Roullet, P., & Sara, S. J. (1999). Attenuation of emotional and nonemotional memories after their reactivation: Role of β adrenergic receptors. *Journal of Neuroscience*, *19*(15), 6623–6628. https://doi.org/10.1523/jneurosci.19-15-06623.1999

Przybyslawski, J., & Sara, S. J. (1997). Reconsolidation of memory after its reactivation. *Behavioural Brain Research*, *84*(1–2), 241–246. https://doi.org/10.1016/S0166-4328(96)00153-2

Qin, S., Hermans, E. J., van Marle, H. J. F., Luo, J., & Fernández, G. (2009). Acute Psychological Stress Reduces Working Memory-Related Activity in the Dorsolateral Prefrontal Cortex. *Biological Psychiatry*, *66*(1), 25–32. https://doi.org/10.1016/j.biopsych.2009.03.006

Quirk, G J, Russo, G. K., Barron, J. L., & Lebron, K. (2000). The role of ventromedial prefrontal cortex in the recovery of extinguished fear. *The Journal of Neuroscience*, *20*(16), 6225–6231. http://www.ncbi.nlm.nih.gov/pubmed/10934272

Quirk, Gregory J., Likhtik, E., Pelletier, J. G., & Paré, D. (2003). Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. *Journal of Neuroscience*, *23*(25), 8800–8807. https://doi.org/10.1523/jneurosci.23-25-08800.2003

Rabinak, C. A., Angstadt, M., Sripada, C. S., Abelson, J. L., Liberzon, I., Milad, M. R., & Phan, K. L. (2013). Cannabinoid facilitation of fear extinction memory recall in humans. *Neuropharmacology*, *64*, 396–402. https://doi.org/10.1016/j.neuropharm.2012.06.063

Raczka, K. A., Mechias, M. L., Gartmann, N., Reif, A., Deckert, J., Pessiglione, M., & Kalisch, R. (2011). Empirical support for an involvement of the mesostriatal dopamine system in human fear extinction. *Translational Psychiatry*, *1*(May), 1–8. https://doi.org/10.1038/tp.2011.10

Raij, T., Nummenmaa, A., Marin, M. F., Porter, D., Furtak, S., Setsompop, K., & Milad, M. R. (2018). Prefrontal Cortex Stimulation Enhances Fear Extinction Memory in Humans. *Biological Psychiatry*, *84*(2), 129–137. https://doi.org/10.1016/j.biopsych.2017.10.022

Rao-Ruiz, P., Rotaru, D. C., Van Der Loo, R. J., Mansvelder, H. D., Stiedl, O., Smit, A. B., & Spijker, S. (2011). Retrieval-specific endocytosis of GluA2-AMPARs underlies adaptive reconsolidation of contextual fear. *Nature Neuroscience*, *14*(10), 1302–1308. https://doi.org/10.1038/nn.2907

Redondo, R. L., Kim, J., Arons, A. L., Ramirez, S., Liu, X., & Tonegawa, S. (2014). Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature*, *513*(7518), 426–430. https://doi.org/10.1038/nature13725

Reijmers, L. G., Perkins, B. L., Matsuo, N., & Mayford, M. (2007). Localization of a stable neural correlate of associative memory. *Science*, *317*(5842), 1230–1233. https://doi.org/10.1126/science.1143839

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review1*, *74*(3), 151–182.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning: Current Research and Theory, Vol. 2*, *January 1972*, 64–99. http://www.researchgate.net/publication/233820243_A_theory_of_Pavlovian_conditioning_Variations_in_the_effectiveness_of_reinforcement_and_nonreinforcement

Ressler, K. J., Rothbaum, B. O., Tannenbaum, L., Anderson, P., Graap, K., Zimand, E., Hodges, L., &

Davis, M. (2004). Cognitive Enhancers as Adjuncts to Psychotherapy. *Archives of General Psychiatry*, *61*(11), 1136. https://doi.org/10.1001/archpsyc.61.11.1136

Reynolds, M., & Brewin, C. R. (1999). Intrusive memories in depression and posttraumatic stress disorder. *Behavior Research and Therapy*, *37*(3), 201–215. https://doi.org/10.1016/S0005-7967(98)00132-6

Robinson, M. J. F., & Franklin, K. B. J. (2010). Reconsolidation of a morphine place preference: Impact of the strength and age of memory on disruption by propranolol and midazolam. *Behavioural Brain Research*, *213*(2), 201–207. https://doi.org/10.1016/j.bbr.2010.04.056

Roelofs, K. (2017). Freeze for action: Neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1718). https://doi.org/10.1098/rstb.2016.0206

Romanski, L. M., & LeDoux, J. E. (1992). Bilateral destruction of neocortical and perirhinal projection targets of the acoustic thalamus does not disrupt auditory fear conditioning. *Neuroscience Letters*, *142*(2), 228–232. https://doi.org/10.1016/0304-3940(92)90379-L

Roozendaal, B., & McGaugh, J. L. (1996). Amygdaloid nuclei lesions differentially affect glucocorticoid-induced memory enhancement in an inhibitory avoidance task. *Neurobiology of Learning and Memory*, *65*(1), 1–8. https://doi.org/10.1006/nlme.1996.0001

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Roy, D. S., Muralidhar, S., Smith, L. M., & Tonegawa, S. (2017). Silent memory engrams as the basis for retrograde amnesia. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(46), E9972–E9979. https://doi.org/10.1073/pnas.1714248114

Rudy, J. W. (2009). Context representations, context functions, and the parahippocampal-hippocampal system. *Learning and Memory*, *16*(10), 573–585. https://doi.org/10.1101/lm.1494409

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, *6*(4), 332–339. www.nature.com/reviews/neuro

Sandi, C., & Rose, S. P. R. (1994). Corticosterone enhances long-term retention in one-day-old chicks trained in a weak passive avoidance learning paradigm. *Brain Research*, *647*(1), 106–112. https://doi.org/10.1016/0006-8993(94)91404-4

Schafe, G. E., & LeDoux, J. E. (2000). Memory consolidation of auditory pavlovian fear conditioning requires protein synthesis and protein kinase A in the amygdala. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(18), 1–5. https://doi.org/10.1523/jneurosci.20-18-j0003.2000

Schelling, G., Kilger, E., Roozendaal, B., De Quervain, D. J. F., Briegel, J., Dagge, A., Rothenhäusler, H. B., Krauseneck, T., Nollert, G., & Kapfhammer, H. P. (2004). Stress doses of hydrocortisone, traumatic memories, and symptoms of posttraumatic stress disorder in patients after cardiac surgery: A randomized study. *Biological Psychiatry*, *55*(6), 627–633. https://doi.org/10.1016/j.biopsych.2003.09.014

Scheveneels, S., Boddez, Y., Vervliet, B., & Hermans, D. (2016). The validity of laboratory-based treatment research: Bridging the gap between fear extinction and exposure treatment. *Behaviour Research and Therapy*, *86*, 87–94. https://doi.org/10.1016/j.brat.2016.08.015

Schields, G. S., Sazma, M. A., McCullough, M., & Yonelinas, A. P. (2017). The Effects of Acute Stress on

Episodic Memory : A Meta-Analysis and Integrative. *Psychological Bulletin*, *143*(6), 636–675. https://doi.org/10.1037/bul0000100.The

Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M.-H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, *110*(50), 20040–20045. https://doi.org/10.1073/pnas.1320322110

Schiller, Daniela, & Delgado, M. R. (2010). Overlapping neural systems mediating extinction, reversal and regulation of fear. *Trends in Cognitive Sciences*, *14*(6), 268–276. https://doi.org/10.1016/j.tics.2010.04.002

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

Schott, B. H., Minuzzi, L., Krebs, R. M., Elmenhorst, D., Lang, M., Winz, O. H., Seidenbecher, C. I., Coenen, H. H., Heinze, H. J., Zilles, K., Düzel, E., & Bauer, A. (2008). Mesolimbic functional magnetic resonance imaging activations during reward anticipation correlate with reward-related ventral striatal dopamine release. *Journal of Neuroscience*, *28*(52), 14311–14319. https://doi.org/10.1523/JNEUROSCI.2058-08.2008

Schroyens, N., Alfei, J. M., Schnell, A. E., Luyten, L., & Beckers, T. (2019). Limited replicability of drug-induced amnesia after contextual fear memory retrieval in rats. *Neurobiology of Learning and Memory*, *166*, 1–22. https://doi.org/10.1016/j.nlm.2019.107105

Schroyens, N., Beckers, T., & Kindt, M. (2017). In search for boundary conditions of reconsolidation: A failure of fear memory interference. *Frontiers in Behavioral Neuroscience*, *11*(April), 1–13. https://doi.org/10.3389/fnbeh.2017.00065

Schroyens, N., Schnell, A. E., & Luyten, L. (2019). *Limited replicability of drug-induced amnesia after contextual fear memory retrieval in rats*. *June*, 1–22. https://doi.org/10.31234/osf.io/akz6b

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwabe, L., & Wolf, O. T. (2009). New episodic learning interferes with the reconsolidation of autobiographical memories. *PLoS ONE*, *4*(10), 4–7. https://doi.org/10.1371/journal.pone.0007519

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*(9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007

Sevenster, D., Beckers, T., & Kindt, M. (2014a). Fear conditioning of SCR but not the startle reflex requires conscious discrimination of threat and safety. *Frontiers in Behavioral Neuroscience*, *8*(February), 1–9. https://doi.org/10.3389/fnbeh.2014.00032

Sevenster, D., Beckers, T., & Kindt, M. (2014b). Prediction error demarcates the transition from retrieval, to reconsolidation, to new learning. *Learning and Memory*, *21*(11), 580–584. https://doi.org/10.1101/lm.035493.114

Sharot, T., & Phelps, E. A. (2004). How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective and Behavioral Neuroscience*, *4*(3), 294–306. https://doi.org/10.3758/CABN.4.3.294

Shiban, Y., Brütting, J., Pauli, P., & Mühlberger, A. (2015). Fear reactivation prior to exposure therapy: Does it facilitate the effects of VR exposure in a randomized clinical sample? *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 133–140. https://doi.org/10.1016/j.jbtep.2014.09.009

Shim, R. S., Compton, M. T., Rust, G., Druss, B. G., & Kaslow, N. J. (2009). Race-Ethnicity as a Predictor of Attitudes Toward Mental Health Treatment Seeking. *Psychiatric Services*, *60*(10), 1336–1341. https://doi.org/10.1176/ps.2009.60.10.1336.Race-Ethnicity

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage*, *20*(1), 114–124. https://doi.org/10.1016/S1053-8119(03)00298-2

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sinclair, A. H., & Barense, M. D. (2019). Prediction Error and Memory Reactivation: How Incomplete Reminders Drive Reconsolidation. *Trends in Neurosciences*, *42*(10), 727–739. https://doi.org/10.1016/j.tins.2019.08.007

Singer, A. C., & Frank, L. M. (2009). Rewarded Outcomes Enhance Reactivation of Experience in the Hippocampus. *Neuron*, *64*(6), 910–921. https://doi.org/10.1016/j.neuron.2009.11.016

Soeter, M., & Kindt, M. (2011). Disrupting reconsolidation: Pharmacological and behavioral manipulations. *Learning and Memory*, *18*(6), 357–366. https://doi.org/10.1101/lm.2148511

Soeter, M., & Kindt, M. (2015). An Abrupt Transformation of Phobic Behavior after a Post-Retrieval Amnesic Agent. *Biological Psychiatry*, *78*(12), 880–886. https://doi.org/10.1016/j.biopsych.2015.04.006

Spielberger, C. D. (1983). *Manual for the State- Trait Anxiety Inventory*. Consulting Psychologists Press.

Squire, L. R. (1992). Memory and the Hippocampus: A Synthesis From Findings With Rats, Monkeys, and Humans. *Psychological Review*, *99*(2), 195–231. https://doi.org/10.1037/0033-295X.99.2.195

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, *5*(2), 169–177. https://doi.org/10.1016/0959-4388(95)80023-9

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13515–13522. https://doi.org/10.1073/pnas.93.24.13515

Stafford, J. M., Maughan, D. A. K., Ilioi, E. C., & Lattal, K. M. (2013). Exposure to a fearful context during periods of memory plasticity impairs extinction via hyperactivation of frontal-amygdalar circuits. *Learning and Memory*, *20*(3), 156–163. https://doi.org/10.1101/lm.029801.112

Steckle, L. C. (1933). A trace conditioning of the galvanic reflex. *Journal of General Psychology*, *9*(2), 475–480. https://doi.org/10.1080/00221309.1933.9920953

Steinfurth, E. C. K., Kanen, J. W., Raio, C. M., Clem, R. L., Huganir, R. L., & Phelps, E. A. (2014). Young and old Pavlovian fear memories can be modified with extinction training during. *Learning and Memory*, *21*(7), 338–341. https://doi.org/10.1101/lm.033589.113

Su, J., Li, P., Zhuang, Q., Chen, X., Zhang, X., Li, X., Wang, J., Yu, X., & Wang, Y. (2021). Identification of the Similarities and Differences of Molecular Networks Associated With Fear Memory

Formation, Extinction, and Updating in the Amygdala. *Frontiers in Molecular Neuroscience*, *14*(December). https://doi.org/10.3389/fnmol.2021.778170

Surís, A., North, C., Adinoff, B., Powell, C. M., & Greene, R. (2010). Effects of exogenous glucocorticoid on combat-related PTSD symptoms. *Annals of Clinical Psychiatry*, *22*(4), 274–279.

Suzuki, A. (2004). Memory Reconsolidation and Extinction Have Distinct Temporal and Biochemical Signatures. *Journal of Neuroscience*, *24*(20), 4787–4795. https://doi.org/10.1523/JNEUROSCI.5491-03.2004

Suzuki, Akinobu, Josselyn, S. A., Frankland, P. W., Masushige, S., Silva, A. J., & Kida, S. (2004). Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *Journal of Neuroscience*, *24*(20), 4787–4795. https://doi.org/10.1523/JNEUROSCI.5491-03.2004

Talmi, D., Anderson, A. K., Riggs, L., Caplan, J. B., & Moscovitch, M. (2008). Immediate memory consequences of the effect of emotion on attention to pictures. *Learning and Memory*, *15*(3), 172–182. https://doi.org/10.1101/lm.722908

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

Taubenfeld, S. M., Milekic, M. H., Monti, B., & Alberini, C. M. (2001). The consolidation of new but not reactivated memory requires hippocampal C/EBPβ. *Nature Neuroscience*, *4*(8), 813–818. https://doi.org/10.1038/90520

Tedeschi, R. G., Calhoun, L. G., Tedeschi, R. G., & Calhoun, L. G. (2016). Posttraumatic Growth : Conceptual Foundations and Empirical Evidence " Posttraumatic Growth : Conceptual Foundations and Empirical Evidence. *Psychological Inquiry*, *7965*(March), 1–18. https://doi.org/10.1207/s15327965pli1501

Thiele, M., Yuen, K. S. L., Gerlicher, A. V. M., & Kalisch, R. (2021). A ventral striatal prediction error signal in human fear extinction learning. *NeuroImage*, *229*(November 2020), 117709. https://doi.org/10.1016/j.neuroimage.2020.117709

Thomas, B. L., Cutler, M., & Novak, C. (2012). A modified counterconditioning procedure prevents the renewal of conditioned fear in rats. *Learning and Motivation*, *43*(1–2), 24–34. https://doi.org/10.1016/j.lmot.2012.01.001

Thompson, A., & Lipp, O. V. (2017). Extinction during reconsolidation eliminates recovery of fear conditioned to fear-irrelevant and fear-relevant stimuli. *Behaviour Research and Therapy*, *92*, 1–10. https://doi.org/10.1016/j.brat.2017.01.017

Thornton, S. (2021). *Karl Popper*. The Stanford Encyclopedia of Philosophy.

Totty, M. S., Payne, M. R., & Maren, S. (2019). Event boundaries do not cause the immediate extinction deficit after Pavlovian fear conditioning in rats. *Scientific Reports*, *9*(1), 1–7. https://doi.org/10.1038/s41598-019-46010-4

Tronson, N. C., Wiseman, S. L., Olausson, P., & Taylor, J. R. (2006). Bidirectional behavioral plasticity of memory reconsolidation depends on amygdalar protein kinase A. *Nature Neuroscience*, *9*(2), 167–169. https://doi.org/10.1038/nn1628

Tulving, E. (1972). Episodic and semantic memory: Where should we go from here? *Behavioral and Brain Sciences*, *9*(3), 573–577. https://doi.org/10.1017/S0140525X00047257

Tulving, E. (2002). Episodic Memory : From Mind to Brain. *Annual Review of Psychology*, *53*, 1–25.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Vallejo, A. G., Kroes, M. C. W., Rey, E., Acedo, M. V., Moratti, S., Fernández, G., & Strange, B. A. (2019). Propofol-induced deep sedation reduces emotional episodic memory reconsolidation in humans. *Science Advances*, *5*(3). https://doi.org/10.1126/sciadv.aav3801

van Dis, E. A. M., Hagenaars, M. A., Bockting, C. L. H., & Engelhard, I. M. (2019). Reducing negative stimulus valence does not attenuate the return of fear: Two counterconditioning experiments. *Behaviour Research and Therapy*, *120*(May), 103416. https://doi.org/10.1016/j.brat.2019.103416

Van Well, S., Visser, R. M., Scholte, H. S., & Kindt, M. (2012). Neural substrates of individual differences in human fear learning: Evidence from concurrent fMRI, fear-potentiated startle, and US-expectancy data. *Cognitive, Affective and Behavioral Neuroscience*, *12*(3), 499–512. https://doi.org/10.3758/s13415-012-0089-7

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear Extinction and Relapse: State of the Art. *Annual Review of Clinical Psychology*, *9*(1), 215–248. https://doi.org/10.1146/annurev-clinpsy-050212-185542

Vieweg, W. V. R., Julius, D. A., Fernandez, A., Beatty-Brooks, M., Hettema, J. M., & Pandurangi, A. K. (2006). Posttraumatic Stress Disorder: Clinical Features, Pathophysiology, and Treatment. *American Journal of Medicine*, *119*(5), 383–390. https://doi.org/10.1016/j.amjmed.2005.09.027

Visser, R. M., Bathelt, J., Scholte, H. S., & Kindt, M. (2021). Robust BOLD Responses to Faces But Not to Conditioned Threat: Challenging the Amygdala's Reputation in Human Fear and Extinction Learning. *The Journal of Neuroscience*, *41*(50), 10278–10292. https://doi.org/10.1523/jneurosci.0857-21.2021

Voogd, L. D. de, Murray, Y. P. J., Barte, R. M., Heide, A. van der, Fernández, G., Doeller, C. F., & Hermans, E. J. (2019). The role of hippocampal spatial representations in contextualization and generalization of fear. *NeuroImage*. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116308

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Walker, D. L., Ressler, K. J., Lu, K. T., & Davis, M. (2002). Facilitation of conditioned fear extinction by systemic administration or intra-amygdala infusions of D-cycloserine as assessed with fear-potentiated startle in rats. *Journal of Neuroscience*, *22*(6), 2343–2351. https://doi.org/10.1523/jneurosci.22-06-02343.2002

Wang, S. H., De Oliveira Alvares, L., & Nader, K. (2009). Cellular and systems mechanisms of memory strength as a constraint on auditory fear reconsolidation. *Nature Neuroscience*, *12*(7), 905–912. https://doi.org/10.1038/nn.2350

Watson, J. B., & Rayner, R. (1920). Conditioned Emotional Reactions. *Journal of Experimental Psychology*, *3*(1), 1–14.

Weike, A. I., Schupp, H. T., & Hamm, A. O. (2007). Fear acquisition requires awareness in trace but not delay conditioning. *Psychophysiology*, *44*(1), 170–180. https://doi.org/10.1111/j.1469-8986.2006.00469.x

Weiland, B. J., Heitzeg, M. M., Zald, D., Cummiford, C., Love, T., Zucker, R. A., & Zubieta, J. K. (2014). Relationship between impulsivity, prefrontal anticipatory activation, and striatal dopamine release during rewarded task performance. *Psychiatry Research - Neuroimaging*, *223*(3), 244–252. https://doi.org/10.1016/j.pscychresns.2014.05.015

Weiland, B. J., Zucker, R. A., Zubieta, J. K., & Heitzeg, M. M. (2017). Striatal dopaminergic reward response relates to age of first drunkenness and feedback response in at-risk youth. *Addiction Biology*, *22*(2), 502–512. https://doi.org/10.1111/adb.12341

Williams, L. M. (2016). Precision Psychiatry: a neural circuit taxonomy for depression and anxiety. *Lancet Psychiatry*, *3*(5), 472–480. https://doi.org/10.1016/S2215-0366(15)00579-9

Xu, J., Zhu, Y., Kraniotis, S., He, Q., Marshall, J. J., Nomura, T., Stauffer, S. R., Lindsley, C. W., Jeffrey, P. C., & Contractor, A. (2013). Potentiating mGluR5 function with a positive allosteric modulator enhances adaptive learning. *Learning and Memory*, *20*(8), 438–445. https://doi.org/10.1101/lm.031666.113

Yang, C. hao, Shi, H. shui, Zhu, W. li, Wu, P., Sun, L. li, Si, J. jian, Liu, M. meng, Zhang, Y., Suo, L., & Yang, J. li. (2012). Venlafaxine facilitates between-session extinction and prevents reinstatement of auditory-cue conditioned fear. *Behavioural Brain Research*, *230*(1), 268–273. https://doi.org/10.1016/j.bbr.2012.02.023

Yang, Y. L., Chao, P. K., & Lu, K. T. (2006). Systemic and intra-amygdala administration of glucocorticoid agonist and antagonist modulate extinction of conditioned fear. *Neuropsychopharmacology*, *31*(5), 912–924. https://doi.org/10.1038/sj.npp.1300899

Yehuda, R., & LeDoux, J. (2007). Response Variation following Trauma: A Translational Neuroscience Approach to Understanding PTSD. *Neuron*, *56*(1), 19–32. https://doi.org/10.1016/j.neuron.2007.09.006

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*(4), 1912–1920. https://doi.org/10.1016/j.neuroimage.2008.01.057

Zhang, X., Kim, J., & Tonegawa, S. (2020). Amygdala Reward Neurons Form and Store Fear Extinction Memory. *Neuron*, *105*(6), 1077-1093.e7. https://doi.org/10.1016/j.neuron.2019.12.025

Zimmermann, J., & Bach, D. R. (2020). Impact of a reminder/extinction procedure on threat-conditioned pupil size and skin conductance responses. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *27*(4), 164–172. https://doi.org/10.1101/lm.050211.119

Zink, C. F., Pagnoni, G., Martin-skurski, M. E., Chappelow, J. C., & Berns, G. S. (2004). Human Striatal Responses to Monetary Reward Depend on Saliency. *Neuron*, *42*, 509–517.

Zuccolo, P. F., & Hunziker, M. H. L. (2019). A review of boundary conditions and variables involved in the prevention of return of fear after post-retrieval extinction. *Behavioural Processes*, *162*, 39–54. https://doi.org/10.1016/j.beproc.2019.01.011