HHS Public Access

Author manuscript

Neuron. Author manuscript; available in PMC 2016 October 07.

Published in final edited form as:

Neuron. 2015 October 7; 88(1): 47–63. doi:10.1016/j.neuron.2015.09.028.

Rethinking Extinction

Joseph E. Dunsmoor[1,*], **Yael Niv**[2], **Nathaniel Daw**[2], and **Elizabeth A. Phelps**[1,3,*]

[1]Department of Psychology and Center for Neural Sciences, New York University, New York, NY, USA

[2]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, New Jersey, USA

[3]Nathan Kline Institute, Orangeburg, NY, USA

Abstract

Extinction serves as the leading theoretical framework and experimental model to describe how learned behaviors diminish through absence of anticipated reinforcement. In the past decade, extinction has moved beyond the realm of associative learning theory and behavioral experimentation in animals and has become a topic of considerable interest in the neuroscience of learning, memory, and emotion. Here, we review research and theories of extinction, both as a learning process and as a behavioral technique, and consider whether traditional understandings warrant a re-examination. We discuss the neurobiology, cognitive factors, and major computational theories, and revisit the predominant view that extinction results in new learning that interferes with expression of the original memory. Additionally, we reconsider the limitations of extinction as a technique to prevent the relapse of maladaptive behavior, and discuss novel approaches, informed by contemporary theoretical advances, that augment traditional extinction methods to target and potentially alter maladaptive memories.

Keywords

extinction; medial prefrontal cortex; emotion; associative learning; latent cause model; fear conditioning; context; amygdala; Rescorla-Wagner

Introduction

Along with the discovery of the conditioned response (CR), one of Pavlov's most significant contributions to physiology and to psychological science was the observation that absence of reinforcement resulted in a weakening or disappearance of acquired behavior. Termed by Pavlov as the *internal inhibition of conditioned reflexes* (Pavlov, 1927), experimental

*Correspondence: Joseph Dunsmoor, Department of Psychology, 6 Washington Place Room 890, New York University, New York, NY 10003 USA. joseph.dunsmoor@nyu.edu. Elizabeth A. Phelps, Department of Psychology, 6 Washington Place Room 890, New York University, New York, NY 10003 USA. liz.phelps@nyu.edu.

extinction generated theoretical and empirical research interest throughout the twentieth century, but research on extinction paled in comparison to studies of conditions that generate acquisition of CRs. In the past decade, however, there has been a surge of interest in experimental extinction for its own sake. The topic spans neurobehavioral studies in laboratory animals and humans, cellular, molecular and genetic research, and computational learning models. Beyond interest in the basic mechanisms of learning and memory, renewed attention to extinction is due in large part to the clinical significance of extinction for the treatment of a variety of psychiatric disorders (Milad and Quirk, 2012; Vervliet et al., 2013). Specifically, extinction serves as the basis for exposure-based therapy, a primary treatment for anxiety disorders, addiction, and trauma and stress related disorders (Powers et al., 2010). Experimental extinction is also considered within the National Institute of Mental Health's Research Domain Criteria as a scientific paradigm to provide objective neurobehavioral measures of mental illness in the domain of Negative Affect. It is hoped that advances in our understanding of extinction across multiple fronts will translate to new, effective treatments for psychiatric conditions characterized by the inability to regulate pathological fear or anxiety.

The purpose of this review is to consider how the view of extinction has changed as new findings have emerged, and to discuss new directions and unanswered questions in this burgeoning field. Notably, research and theory on extinction is immense. This article covers what we believe are significant themes relevant for understanding how the fields of computational learning theory and the neuroscience of learning, memory, and emotion view extinction. Throughout this article, we attempt to delineate between where there is consensus (Box 1: Current status of the field) and where there are theoretical or practical gaps in our understanding (Box 2: Future directions).

Box 1

Current status of the field

- Return of extinguished behavior is common following the passage of time (*spontaneous recovery*), when extinguished cues are encountered outside the extinction context (*contextual renewal*), and following presentation of the unconditioned stimulus (*reinstatement*). These effects provide support for the widely held view that extinction is a new form of learning, and that conditioning and extinction memories may coexist in distinct neural circuits and be reactivated independently based on environmental or situational factors.
- Contemporary computational models have been developed to reflect the understanding that extinction is not simply a change (decrease) in a previously learned value. Accordingly, they augment such learning with the possibility that extinction may also arise when a new "state" (or association) is created, for which a new value is learned.
- Neurobiological models of extinction focus on interactions between and processes within the medial prefrontal cortex, amygdala, and hippocampus. This basic neurocircuitry appears to be conserved across species.

- The principles of extinction serve as the basis for clinical treatments such as exposure-based therapy, which is considered an effective treatment for a host of anxiety disorders, as well as addiction.

Box 2

Future directions

- Under what conditions is a fear memory retrieved and updated, as opposed to a new extinction memory trace being laid down? Computationally, the question is what are the factors that determine when a new state (or latent cause) of the associative learning task will be inferred, versus retrieval and updating of an old state?
- What is the neurobiological signature of updating of a persistent memory, and what are the necessary and sufficient conditions to demonstrate that a memory has been persistently altered.
- Contemporary studies of extinction of instrumental conditioning, including extinction of avoidance behaviors, have received far too little attention, and should be integrated into a general picture of learning and unlearning in the brain.
- What is the role of predisposing genetic and epigenetic variants associated with extinction learning? To what extent do individual differences such as early life stress, trait anxiety, and intolerance of uncertainty moderate extinction and extinction retention in humans?
- Are extinction deficits a diagnostic biomarker of trauma and stressor related disorders like PTSD, and clinical anxiety disorders such as obsessive compulsive, generalized anxiety, and panic disorders?
- How will techniques that appear to persistently alter conditioned threat memories in non-human animals translate to complex fear memories in humans? For instance, invasive techniques like blocking protein synthesis in the amygdala during consolidation or reconsolidation of a threat memory appear effective for simple associative memories like a tone-shock pairing, but under what circumstances will they be effective for traumatic memories such as those implicated in PTSD? Relatedly, do noninvasive behavioral techniques that effectively eliminate the conditioned response translate to more generalized threat memories or human emotional episodic memories, and if so, what are the boundary conditions that define when these techniques will and when they will not be useful?

The first section is composed of a brief background on the theoretical foundation upon which contemporary views of extinction rest, a description of the neurobiology of extinction, psychological factors, and major associative learning models. A primary question is whether the mechanisms supporting extinction involve new learning that inhibits or

interferes with original learning, as is the current mainstay, or also cause erasure of the original learning, as suggested by recent theoretical and experimental work. In particular, we survey a recent framework that reinterprets extinction in terms of sound statistical reasoning about the causes of events in the world, and suggest that this framework can conceptualize the tradeoff between new learning and memory modification. In the second section, we detail the shortfalls of traditional extinction techniques in preventing the return of unwanted behaviors, and discuss novel approaches to augment extinction that compensate for these shortfalls. We attempt to understand the success of these approaches in terms of several distinct theoretical mechanisms, including interference and erasure, which might contribute to extinction. Of note, we focus almost exclusively on extinction in the domain of fear or threat conditioning, as it is in this arena that many of the advances in neuroscience, behavior, learning theory, and clinical translational research have been made.

Foundational research and theories of extinction

The canonical expression of experimental extinction rests on Pavlovian conditioning, in which a conditional stimulus (CS; e.g., a tone or light) is paired with a naturally salient unconditional stimulus (US; e.g., food or an electric shock). Once a relationship between the CS and US is established, presentation of the CS initiates a conditioned response (e.g., increases in salivation). In the domain of fear conditioning, in which the US is naturally unpleasant or painful, the CR often takes the form of defensive behaviors or emotional reactions such as increases in sweating, heart rate, pupil size, freezing, and blood pressure. With continuing presentation of the CS in the absence of the US, the CR gradually diminishes or is eliminated altogether.

Contemporary theoretical views of extinction are in many ways based directly on early formulations by Pavlov (Pavlov, 1927). Pavlov interpreted extinction as a form of ‘internal inhibition’ (as opposed to decreases in the CR resulting from the presence of another stimulus, which he termed ‘external inhibition’). According to Pavlov, extinction disrupts the CR, but does not destroy it. Evidence that the CR is preserved comes from the fact that it tends to return over time, what Pavlov termed *spontaneous recovery* or restoration. Pavlov (1927) considered spontaneous recovery to be a measure of the depth of the extinction process itself: “[Extinction] is measured, other conditions being equal, by the time taken for spontaneous restoration of the extinguished reflex to its original strength” (pp 58). Other evidence for the persistence of the original CS-US association includes *contextual renewal* (the return of the CR if tested in a different context), *reinstatement* (the return of the CR when tested after a reminder US) and *rapid reacquisition* (rapid re-learning of the CS-US association) (Box 1: Current status of the field).

Of theoretical import is the question of what occurs during extinction that reduces the CR. For Pavlov, the central mechanism involved inhibitory properties accruing to the CS over the course of extinction training, a process putatively subserved by inhibitory cells in the cortex (notably, Pavlov’s references to the central nervous system were vague). The notion that the CS acquires inhibitory properties that suppress the CR is still the predominant view of extinction (e.g., Bouton et al., 2006; Larrauri and Schmajuk, 2008), though theories on the nature of inhibitory learning vary, as detailed below.

The obvious alternative formulation to inhibition is that of erasure or modification of the original CS-US associative memory. Erasure seems a less tenable mechanism overall, simply because spontaneous recovery is so common following traditional extinction. However, some early theories proposed that erasure (or, at least, partial erasure) does play a role in the extinction process. For instance, Razran (1956) proposed a two-stage process of extinction in which the early stage consists of partial erasure (or 'de-conditioning') resulting from a loss of feedback, and the later stage consists of new learning that counteracts the residual excitatory CR.

An important consideration is that spontaneous recovery is rarely complete (Delamater and Westbrook, 2014); that is, the CR does not return to its original level, and rapidly re-extinguishes. This may suggest some partial erasure of original learning. However, since affirmative signatures of memory erasure or modification do not currently exist, weakened recovery might in principle reflect strengthened inhibitory learning and not erasure (see Box 2: Future directions). It is also important to consider that a CS-US association likely involves multiple independent components (sensory/perceptual, emotional, temporal, conceptual, etc.) (Brandon et al., 2000; Delamater, 2012a, b). Fear extinction may reduce emotional elements, while leaving other associations (e.g., sensory) intact. Evidence of extinction is therefore sensitive to the specific choice of which behavioral response to assay at the time of test (Delamater and Westbrook, 2014; Lattal and Wood, 2013), and effective extinction may only mimic erasure by eliminating a conditioned fear response, while leaving other elements of the CS-US association intact. In short, it is possible th at extinction simultaneously erases, inhibits, and has no effect on separate aspects of the same memory.

Neurobiology of fear extinction

Studies investigating the neural mechanisms of fear conditioning across species indicate that the amygdala is critical for the acquisition, storage, and expression of conditioned fear (see LeDoux, 2000 for review). The lateral nucleus of the amygdala (LA) is thought to be the site of synaptic plasticity that encodes the association between CS and US sensory inputs. In the presence of the CS, the LA excites the central nucleus (CE), which mediates CR expression through projections to the brainstem and hypothalamus. The LA also indirectly projects to the CE through the basal nucleus and the intercalated (ITC) cell masses (clusters of inhibitory GABAergic neurons). The basal nucleus itself also projects directly to the ITC. These pathways provide multiple potential circuits for gating fear expression in extinction. Research in rodents using lesions, pharmacological manipulations, and electrophysiology provide an increasingly detailed model of the neural circuitry of fear extinction. This research suggests that interactions between of the amygdala, the ventral medial prefrontal cortex (vmPFC), and the hippocampus support the acquisition, storage, retrieval, and contextual modulation of fear extinction (see Milad and Quirk, 2012 for review).

Pharmacological and electrophysiological studies in rodents suggest that the amygdala, in addition to its role in the acquisition and expression of conditioned fear, also plays a role in the acquisition and consolidation of fear extinction. For instance, blockade of NMDA in the LA (Sotres-Bayon et al., 2007) or glutamate (Kim et al., 2007) receptors within the basolateral amygdala complex (BLA) impairs extinction learning, and the blockade of

mitogen-activated protein kinase (MAPk) activity in the BLA entirely prevents the acquisition of extinction (Herry et al., 2006). Furthermore, several studies suggest that the consolidation of extinction learning is supported by morphological changes in synapses of the BLA (Chhatwal et al., 2005). Consistent with the notion that extinction results in new learning, not erasure of the original fear memory, a population of neurons in the LA have been identified in which the CS response is maintained despite a decrease in the expression of conditioned fear with extinction, along with a second more transiently responsive population (Repa et al., 2001). This finding provides further evidence that the amygdala supports the maintenance of the original fear memory while simultaneously facilitating extinction learning (see Hartley and Phelps, 2010 for review).

Although the amygdala may be critical for the acquisition of extinction learning, the vmPFC is also necessary for the acquisition and recall of extinction. This was first demonstrated by Morgan et al. (1993) who found that rodents with vmPFC lesions required many more presentations of the CS to extinguish conditioned fear. It was later found that the infralimbic (IL) region of the vmPFC is the site of extinction consolidation (Quirk et al., 2000). Disruption of protein synthesis (Santini et al., 2004), MAPk blockade (Hugues et al., 2006), and administration of an NMDA antagonist (Burgos-Robles et al., 2007) within the vmPFC impairs retrieval of extinction, indicating that the plasticity in this region supports extinction consolidation. Electrophysiological studies suggest that the IL inhibits the expression of conditioned fear during extinction through reciprocal connections with the amygdala. IL neurons show increased activity to the CS during extinction retrieval (Milad and Quirk, 2002) and stimulation of IL neurons both decreases the responsiveness of CE neurons (Quirk et al., 2003) and diminishes conditioned responding to a non-extinguished CS (Milad et al., 2004). Inhibition of fear expression during extinction may therefore occur through IL activation of the inhibitory ITC projections to the CE, or through IL activation of inhibitory interneurons in the LA (see Milad and Quirk, 2012 for review).

Following extinction, contextual information plays a critical role in determining whether the original fear memory or the new extinction memory controls fear expression (see Bouton, 2004). Rats with hippocampal lesions show impaired contextual renewal of the CR (Wilson et al., 1995), and inactivation of the hippocampus after extinction learning prevents the renewal of conditioned fear (Hobin et al., 2006). In addition, inactivation of the hippocampus before extinction learning impairs extinction recall on the subsequent day (e.g., Corcoran et al., 2005). This suggests that the hippocampus may mediate fear expression both outside and within the extinction context. The hippocampus is proposed to control the context-specific retrieval of extinction both indirectly through projections to the vmPFC, and directly through projections to the LA (see Maren et al., 2013 for a review).

Consistent with studies in animal models, functional neuroimaging, lesion and morphology studies in humans indicate that extinction learning depends on the integrated functioning of a neural circuit that includes the amygdala, the vmPFC, and the hippocampus (Milad and Quirk, 2012). This convergent evidence suggests that the neural mechanisms supporting fear extinction are phylogenetically conserved across species.

Psychological and Cognitive factors

It is widely recognized that whatever is learned in extinction is more fragile than the original associations trained through CS-US conditioning, as evidenced by findings that the acquisition CR returns in a variety of situations. This apparent inability to abolish the memory of a conditioning experience may be adaptive: In nature, signals for danger may rarely coincide with actual threat. On the occasion when threat does exist, however, a rapid defensive response could promote survival. From this perspective, the fragility and transience of extinction seems appropriately balanced against the strength and persistence of conditioning. In fear learning, the term 'adaptive conservatism' or 'anxiety conservation' have been used to describe this better-safe-than-sorry approach (Solomon and Wynne, 1954); the survival cost of inappropriately disregarding a danger signal is higher than the cost of inappropriately responding to those signals when threat is not imminent. Thus, despite repeated presentations of a CS in the absence of the US, maintaining some trace of the original memory could provide defense against even the remote possibility of future threat.

A number of psychological factors may help support the maintenance of the conditioning memory after extinction (Lovibond, 2004). One factor is beliefs or contingency knowledge regarding the CS-US relationship. For example, if during extinction another stimulus is presented at the same time as the CS, or a novel action is enabled that prevents the occurrence of the US, this other stimulus or action can prevent the original CS from acquiring inhibitory properties, an effect referred to as *protection from extinction* (Rescorla, 2003). Indeed, once the other stimulus or action are removed the CR returns, suggesting that the absence of the US had been attributed to the (now absent) additional factor. To clinicians, protection from extinction may be reflected in safety behaviors that interfere with the success of exposure-based therapy.

Cognitive mechanisms are also involved in complex forms of inhibitory learning that involve retrospective revaluation (Dickinson and Burke, 1996). In backward blocking, for example, subjects learn that a compound of two stimuli (e.g., a light and a tone) predicts a US. Presented alone, each element will elicit some amount of conditioned responding. However, if one element of the compound (e.g., the light) is then paired alone with the US, then the second element (the tone) ceases to elicit a conditioned response. It seems that since the light can fully predict the US, the tone is retrospectively regarded as unrelated to the US. Such effects that arise from retrospective revaluation provide strong evidence that the memory representation of a CS and its predictive value can be updated even when the CS is absent. In part because such updating is challenging (though not insurmountable) for classic associative learning mechanisms, and in part because, in humans, many of these experiments were framed in causal learning terms, these effects have been interpreted in terms of cognitive beliefs and expectancies about the causal nature of the CS (Lovibond, 2004). However, retrospective revaluation and protection from extinction effects occur in other species (e.g., Miller and Matute, 1996; Rescorla, 2003), and although these effects may implicate explicit causal reasoning in humans, some modern theoretical accounts reconceptualize standard associative learning in similar terms, as effectively a mechanism for inferring the causal relationships underlying observed events (Courville et al., 2005;

Courville et al., 2003; Gershman and Niv, 2012) - as described in detail below. Viewed from this perspective, extinction can generate a number of beliefs about the CS-US relationship: for instance, the CS no longer predicts the US, the CS predicts the US less reliably than before, or the CS predicts the US just as reliably as it did before, but something else is temporarily preventing the US from occurring. Belief in each proposition could result in the same reduction of fear at the time of extinction, but which belief predominates can determine whether expression favors the extinction memory or the fear memory in the future.

Associative learning theories of extinction

A number of influential learning theories explain acquisition and extinction of Pavlovian conditioning (see Figure 1). The following section is not an exhaustive review of these theories, but instead describes how extinction is generally conceptualized within an associative learning framework. As discussed above, theoretical views of extinction fall broadly within two general classes: associative loss or 'unlearning' (e.g., Rescorla and Wagner, 1972), and new inhibitory learning or interference (e.g., Pearce and Hall, 1980). But these two mechanisms are not mutually exclusive, a point that has come into clearer focus in a newer series of statistical learning models, which have reconceptualized the key mechanisms of classic associative learning models as each arising from different aspects of sound statistical inference. Applied to extinction, this class of models points toward a single account balancing contributions from both unlearning and interference, and which may help to clarify the experimental circumstances that may favor either mechanism (Gershman et al., 2010; Redish et al., 2007). Furthermore, whereas associative learning theories have historically been more successful at explaining initial extinction rather than post-extinction recovery effects (e.g., Miller et al., 1995), newer theories aims more explicitly at a unified account of both.

Rescorla-Wagner and the Kalman Filter—The most influential associative learning account of Pavlovian conditioning is the Rescorla and Wagner model (1972). The model suggests that discrepancies between the predicted and actual outcome drives learning ('error correcting learning'). Associative strength (Figure 1A, top) increases when a surprising US (positive prediction error) occurs, and decreases due to the absence of a predicted US (negative prediction error). This model has been used with great success to describe a number of conditioning-related phenomena, including simple acquisition curves and more complex forms of learning involving cue competition such as blocking (Kamin, 1969) and over-expectation (Rescorla, 1970). However, one of the more notable failures of the model is in describing post-extinction recovery effects (Miller et al., 1995). This is because, in this model, extinction engenders a simple decrease of the associative value of the CS. Thus, extinction is viewed as a form of unlearning and, consequently, recovery is not predicted.

This failure in explaining extinction notwithstanding, the core error-driven learning (and unlearning) mechanism of Rescorla-Wagner has received support from two directions. First, a neural substrate for prediction error signals has been identified in the phasic firing of dopamine neurons in the midbrain (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). These neurons' activities correlate with prediction errors postulated by Rescorla-Wagner,

and evidence from various manipulations of their activity suggests their causal involvement in conditioning. However, this particular system has been examined predominantly in appetitive rather than aversive conditioning; evidence about dopamine's involvement in the latter remains mixed (Matsumoto and Hikosaka, 2009; Ungless et al., 2004) and there may well be additional neural systems playing a similar role in the aversive domain (Daw et al., 2002).

Rescorla-Wagner's error-driven learning principle also arises independently from statistically principled accounts of conditioning. In particular, alongside the rise of Bayesian accounts in psychology more generally, recent theoretical work has aimed to reconceptualize classic associative learning accounts of conditioning (which are more mechanistic) in terms of normative accounts of statistical reasoning about events given noisy evidence (Dayan and Long, 1998). One of the early successes of this program of research was the observation that, given a particular set of assumptions about the structure of noise in the world, standard statistical reasoning about the relationship between CSs and USs gives rise to a rule (known independently in engineering as the Kalman filter) that corresponds closely to the Rescorla-Wagner model (Kakade and Dayan, 2002). In particular, this rule includes the key error-driven learning mechanism, but generalizes the model to include CS-processing mechanisms similar to the Pearce-Hall model (described below) (Courville et al., 2006; Dayan et al., 2000) and to account for retrospective revaluation (Daw et al., 2008; Kakade, 2001). These accounts, however, do not alone shed light on the Rescorla-Wagner model's original failure to account for recovery and renewal following extinction.

Pearce-Hall—A second key account of conditioning is that of Pearce and Hall (1980). Though Rescorla-Wagner and Pearce-Hall models are most famous for differing with respect to their accounts of cue competition phenomena (discussed below), another major departure between these two models are their accounts of extinction. Pearce-Hall (1980) recognized that: "The problem we face in supplying an adequate account of inhibitory learning is rather more fundamental than that met when we first considered excitatory learning. In that case there was, at least, fairly general agreement about the way in which the relationship between CS and US is represented internally. There is no such agreement in the case of inhibitory learning" (pp.543). According to the Pearce-Hall theory, extinction involves new inhibitory learning. Thus, a CS-no US association develops due to omission of the expected US, and can be expressed behaviorally and psychologically through stimulus omission responses, such as frustration due to withdrawal of reward (Amsel, 1958), relief due to omission of an expected threat (Gerber et al., 2014), or orienting in response to a missing stimulus (Dunsmoor and LaBar, 2012). Pearce and Hall (1980) viewed these unconditional no-US responses as evidence that absence of the US is in itself an outcome processed with the currently activated CS representation, thereby generating the CS-no US association (Figure 1A, middle). In this way, Pearce and Hall make no distinction between excitatory and inhibitory learning: "we regard extinction as a new form of conditioning" (pp. 546).

The Pearce-Hall model thus assumes that expression of the excitatory CR diminishes due to an inhibitory relationship between the CS-US association and the CS-no US association. This idea that conditioning can invoke parallel, positive and negative associations

simultaneously goes back at least to Konorski (1967), and may have a neural substrate in the existence of two distinct pathways out of the striatum (known as the direct and indirect pathways), which have opposing effects on behavior. Although this idea has been examined mostly in instrumental conditioning, these two pathways appear to serve as parallel targets for positive and negative plasticity (see Frank et al., 2004).

The other core component of the Pearce-Hall model that distinguishes it from the Rescorla-Wagner model is an emphasis on dynamic changes in a CS's susceptibility to Pavlovian conditioning, referred to as the CS's 'associability'. According to their model, the associability of a CS increases when surprise (absolute prediction error) is high, and diminishes when surprise is low. Since a CS's associability governs the extent of learning about its associations, these dynamics give rise to a number of effects involving stronger or weaker learning following different experiences. In particular, during acquisition, the surprising US increases CS associability, promoting CS-US learning. This same mechanism is thought to be invoked, symmetrically, during extinction, as the surprising omission of the US increases CS associability once again, promoting CS-no US learning. Notably, the more reliably a CS predicts the US at the end of acquisition (and therefore the lower its associability), then the slower extinction will be on the first few trials, as associability is restored. Increasing surprise just prior to extinction sessions should increase the rate of extinction, a prediction confirmed in an experiment reported in Pearce and Hall (1980).

Although the associability gating mechanism of Pearce-Hall and the prediction-error learning of Rescorla-Wagner were initially seen as two competing explanations for conditioning phenomena, they are, in fact, complementary and may both coexist in the brain (Le Pelley, 2004). Evidence for neural associability signals have been reported in rodent and human amygdala, alongside prediction errors observed in the midbrain dopamine system and striatum (Li et al., 2011; Roesch et al., 2010).

Associability-like effects also arise naturally, gating the strength of error-driven learning, in the Kalman filter and related statistical models (Behrens et al., 2007; Courville et al., 2006; Dayan et al., 2000). In particular, a hallmark of statistical learning, which follows directly from Bayes' theorem, is that the extent to which a learner should be willing to update their beliefs about a CS's associations in the face of each new prediction error depends upon the extent to which they were uncertain (or, conversely, confident) about those beliefs beforehand. The centerpiece of Bayesian learning models is the dynamic accounting of this uncertainty, which serves as their formal counterpart to the older construct of associability and helps to clarify its interpretation. The correspondence is good; uncertainty in these models behaves both qualitatively and quantitatively similarly to associability. Finally, as discussed below, uncertainty's role in gating learning extends beyond simply controlling how fast extinction occurs; it should also affect the balance between different types of learning that might arise during extinction, notably between unlearning and interference.

Extinction as a form of memory interference

The predominant theoretical basis of post-extinction recovery effects is that proposed by Bouton (1993, 2004). Similar to the Pearce-Hall explanation, Bouton views extinction as a context-dependent form of new inhibitory learning, and retrieval of the inhibitory memory

interferes with expression of the excitatory memory. However, because in this view extinction is a context-dependent memory, retrieval rarely survives a shift in context: extinction is tied to where it was learned. A key element in Bouton's theory of extinction is that new inhibitory learning renders the CS ambiguous because its presence now signals either the presence or the absence of the US. Resolving this ambiguity after extinction relies on the current context, much as the context of a sentence determines the meaning of an ambiguous word. If the context at test is similar to the context in which extinction occurred, retrieval tends to favor the inhibitory CS-no US memory. Otherwise, retrieval tends to favor the CS-US memory, since this association was learned first or is simply more prominent. Bouton proposes that time is also a context, and therefore spontaneous recovery can be seen as renewal. Reinstatement is similarly context-dependent, as it occurs only if unpaired presentations of the US occur in the same context as subsequent presentations of the CS (see Bouton, 2004).

Latent cause models

A further refinement of the statistical learning models of conditioning suggests a more formal underpinning for Bouton's ideas about context and extinction (Courville et al., 2005; Gershman et al., 2010; Gershman and Niv, 2012). The key idea here is that conditioning is conceived as inference about the causal structure that gives rise to observed stimuli such as CSs and USs. However, unlike the Kalman filter (and the associative learning theories that are its cousins) it is not assumed that the CS is directly (e.g., causally) linked to the US. Instead, some third, not observable event causes them both. Such an event is known as a latent cause. This class of models use statistical inference to figure out how likely it is that different underlying structures of latent causes produced the experienced patterns of observable stimuli, including CSs and USs but also other stimuli that comprise the context. Then, on any particular trial, the CR is determined by using this structure to predict which cause is likely to be active at this point in time, and thus whether a US is expected (Courville et al., 2005; Courville et al., 2003; Gershman et al., 2010; Gershman and Niv, 2012).

Informally, this process of inferring the latent cause responsible for each trial is similar to clustering trials into different categories based on patterns of CSs and USs. CS-US associations following acquisition training are clustered together (Figure 1B) and are represented via a single latent cause that is likely to produce the CS, the US, and any other available internal and external contextual stimuli. When a US is omitted, the mismatch between this event and the pattern predicted by the previous learning causes the model to infer that this trial is likely to have been produced by a new and previously unobserved latent cause, which predicts the CS but not the US, that is, to assign the trial to a new cluster (Figure 1B). Subsequent renewal or recovery, and their context sensitivity, depend on the organism judging which of these two latent causes, old or new, is likely active at the current time, based on how well either one can account for the full panoply of cues currently available. These cues include spatial and temporal context, which gives rise to the sensitivity of recovery phenomena to contextual manipulations. In this way the clusters or latent causes formalize Bouton's notion of context, and the model captures much of the context sensitivity of extinction learning (Gershman et al., 2010).

Importantly, latent cause models subsume both mechanisms of extinction discussed above: the notion of new interfering learning (the creation of a new latent cause), together with the possibility that extinction experience will attenuate or update the original fear association (which happens to the extent that the extinction trial is assigned to the old latent cause), as suggested by Rescorla & Wagner. In particular, during extinction the model continues to learn about the associations of each possible latent cause, with learning distributed between them depending on inference of how likely each cause is to be active (Figure 1A, bottom). To the extent to which an extinction trial is judged to be attributed to the original latent cause rather than a new one, the original CS-US association will be updated, thus reducing the prediction of the US associated with the original latent cause. Such updating follows similar statistical principles to the Kalman filter and other Bayesian parameter-learning models discussed so far; in particular, larger prediction errors and more uncertainty about the weights of the latent cause (standing in for associability) increase the rate of updating, combining the key features of the classic associative learning models of Rescorla-Wagner and Pearce-Hall. Conversely, to the extent to which an extinction trial is judged to be attributed to a new latent cause, that new cause will come to be associated with the CS but with the absence of the US.

The latent cause model therefore predicts that different training patterns may give rise to different balances of updating the original fear memory or the formation of a new extinction memory. For instance, new latent causes should be most often and most confidently created for the most surprising events (those producing larger prediction errors) – since these are the ones that the existing latent cause can least adequately explain. Conversely, this suggests that accomplishing extinction by a series of smaller prediction errors rather than a large one will promote erasure of the original memory over interference (see below).

Moreover, the judgment of whether a new experience matches a previously trained latent cause versus requiring a new one will also depend on how uncertain, or certain, are the old cause's associations, since this, in part, determines how surprising an anomalous experience should be. This model thus predicts that uncertainty (formalizing associability), and the various factors that affects it, such as surprise, overtraining, and environmental volatility, will not only modulate the modification of an existing cause's associations, but will also affect the likelihood of creating a new cause.

A possible biological foundation for some of these mechanisms was also suggested in an earlier model of extinction by Redish et al. (2007), which envisions that something similar to latent causes is implemented by attractor states in the cortex. The stability of the attractor landscape (and hence the tendency, effectively, to interpret events as arising from an existing cause vs. splitting out a new one) is hypothesized to depend on tonic levels of dopamine, giving rise to an influence of dopaminergically signaled prediction errors on the likelihood of creating causes.

Altogether, then, the latent cause framework comprises both of the major mechanisms of extinction from associative learning models, together with further machinery (generalizing earlier ideas about prediction errors and associability) for balancing their effects. This framework provides a promising basis, which has not yet been fully explored, for

developing new, more effective extinction procedures, and for understanding why some approaches have previously shown to be more successful than others in preventing fear recovery. In the next section, we review many of these approaches, highlighting their possible connections (especially those yet to be fully understood) with the hypothesized computational elements.

Augmenting extinction

As discussed above, one of the most reliable findings from fear extinction research is that defensive behaviors are recovered over time, reinstated via presentation of the US, renewed following a change in context, and quickly reacquired: Collectively referred to as the return of fear. Consequently, if the goal of extinction is to permanently reduce unwanted behavior, as in the case of exposure therapy, traditional extinction protocols seem a rather unsatisfactory approach. Establishing safe and effective techniques to strengthen extinction is a fundamental goal of translational research that adopts conditioning-based approaches towards psychotherapy.

What is the matter with extinction?

Return of fear phenomena are elegantly explained by the view that extinction is a context-dependent form of inhibitory learning, and that relapse is due to a failure to retrieve inhibitory CS-no US associations (Bouton, 1993). As mentioned earlier, failure to retrieve the extinction (or ‘safety’) memory can be adaptive from the perspective that mistakenly treating safe stimuli as dangerous is often far less costly than the alternative. Viewed from this perspective, clinical treatments based on extinction principles face an uphill struggle, because mechanisms are in place at the basic level of learning and memory to ‘forget’ safety much more readily than threat (Bouton, 1993; Solomon and Wynne, 1954).

Traditional CS-alone extinction procedures have a number of shortcomings that challenge their usefulness as the basis for exposure therapy. First, extinction relies on negative prediction errors, reliably generated only if the CS predicted the US consistently enough to render its absence a violation of expectancy in the first place. In most real world situations, however, highly feared outcomes may occur infrequently (or not at all). For example, an individual fearful of heights may maintain the fear despite never falling. Indeed, extinction proceeds more slowly following partial reinforcement, a result known as the partial reinforcement extinction effect (Jenkins and Stanley, 1950). Statistical learning models also capture this effect by accounting for the uncertainty and variability of events; if a US occurs rarely, its omission is unsurprising and should not engender much learning (Courville et al., 2006; Kakade and Dayan, 2002).

Another shortcoming of extinction procedures is that there is often little correlation between behavioral measures and memory strength. For instance, fear conditioning in rodents shows that within-session decreases in the CR are not predictive of between-session recovery of the CR (Plendl and Wotjak, 2010), and in fact *higher* fear responses during extinction are in some cases associated with stronger and more persistent extinction learning (e.g., Rescorla, 2006). Relatedly, Craske et al. (2008) have convincingly shown that within-session fear reduction during exposure treatment does not predict therapeutic outcomes. Latent cause

models of extinction, in fact, predict this inverse relationship between the rate of extinction and magnitude of recovery. That is, if during extinction a new latent cause is inferred, the CR will decrease rapidly, but the original memory (old latent cause) will not be updated. If, however, a new latent cause is *not* inferred, fear responses may persist while predictions contingent on the original latent cause are gradually updated. The latter effect should in principle lead to gradual erasure of the fear memory (Gershman and Hartley, 2015).

Finally, extinction procedures render the CS ambiguous (Bouton, 2004), which, in the clinic, may create an unfavorable situation for individuals averse to ambiguity and uncertainty. For instance, Dunsmoor et al. (2014b) found that individuals with high self-reported intolerance of uncertainty expressed greater spontaneous recovery after fear extinction.

Given these demonstrated difficulties with the persistence of extinction learning, techniques to augment extinction are needed. Below, we discuss the idea of modifying traditional extinction protocols to reduce the return of unwanted behavior, and review emerging approaches that have shown success in animal, human, preclinical, and clinical applications. As an organizing principle, each approach is described in terms of whether it is thought to target the CS-US association (the 'fear' memory), strengthen inhibitory learning (the extinction or 'safety' memory), or promote retrieval of the inhibitory memory (see Figure 2) (see also Craske et al., 2014; Fitzgerald et al., 2014; Laborda et al., 2011).

Targeting the fear memory

Consolidation—The most effective procedure to abolish the conditioned response permanently would be to eliminate the memory of the CS-US association altogether. One approach is to block consolidation of the fear memory by blocking protein synthesis in the amygdala around the time of fear conditioning. If protein synthesis inhibitors are applied to the LA soon after conditioning, then the immediate expression of fear (short-term memory) is left intact, but expression at a later time (long-term memory) is impaired (for review of consolidation processes related to fear conditioning, see Johansen et al., 2011).

Blocking protein synthesis directly in the amygdala is possible in animal models, but is not practical or safe for human therapeutics. One feasible alternative is to administer pharmacological agents that reduce noradrenergic system activity soon after an aversive experience, which is thought to influence protein synthesis in the amygdala (Gelinas and Nguyen, 2005), in order to impede consolidation of emotional memory. A limited number of clinical studies have found that administration of the β-adrenergic antagonist propranolol shortly after trauma may reduce PTSD symptoms (Pitman et al., 2002; Vaiva et al., 2003), but results of more recent studies have shown less promise (Sharp et al., 2010). Importantly, null findings may be due to propranolol administration several days following trauma, which is likely beyond the time window of consolidation.

Another potential method to target consolidation of the CS-US memory is to administer extinction training immediately after fear conditioning in order to interfere with ongoing consolidation processes. In clinical practice, early intervention by exposure therapy following trauma may reduce PTSD symptoms (Rothbaum et al., 2014). Notably, some immediate interventions for trauma that are not based on extinction principles, like

psychological debriefing, may in fact exacerbate anxiety symptoms (Bisson et al., 1997) and have received considerable criticism as a treatment option for PTSD (e.g., Litz et al., 2002). Laboratory studies of immediate versus delayed extinction are mixed —whereas some labs have found that immediate extinction eliminated spontaneous recovery in rats (Myers et al., 2006), others have found that immediate extinction is less effective than delayed extinction in preventing the return of fear in rats and humans (Huff et al., 2009; Maren and Chang, 2006; Schiller et al., 2008) (see Maren, 2014 for review).

Reconsolidation—Of course, most individuals who seek treatment for fear and anxiety disorders do so long after negative memories have consolidated. Moreover, in many psychiatric illnesses, with the notable exception of PTSD, the etiology is unclear, and so identifying what constitutes early intervention is challenging. However, later treatment may also target the original memory by taking advantage of the phenomenon of reconsolidation, in which re-exposure or reactivation of a previously created long-term memory brings it to a labile state. Studies of reconsolidation suggest that previously consolidated long-term memories can be modified, weakened or even erased, via interventions timed after reactivation of the memory trace and before it is reconsolidated. In a landmark study by Nader and colleagues (2000), conditioned fear responses in rats were effectively abolished by administration of protein synthesis inhibitors into the LA following reactivation of a previously consolidated fear memory. This finding has generated excitement for the idea of targeting and disrupting specific fear memories.

As with blocking consolidation of the original fear memory, administering protein synthesis inhibitors directly into the amygdala in humans is an unrealistic solution to the problem of persistent and intrusive trauma memories. One possibility is to administer safer pharmacological agents, like propranolol, to disrupt memory reconsolidation. Thus far, this protocol has yielded mixed results in PTSD patients (Brunet et al., 2008; Wood et al., 2015). However, primary outcome measures for this research tend to rely on physiological responses during script-driven traumatic imagery, which may be susceptible to individual variability not assessed prior to trauma (Wood et al., 2015). Soeter and Kindt (2015) administered propranolol to participants with arachnophobia following a 2-minute exposure to spiders, and found reduced fear behaviors towards spiders one year after treatment. However, results of laboratory studies of propranolol administration following emotional memory reactivation in humans have also been mixed, with some fear-conditioning studies showing reductions in fear-potentiated startle but not conditioned skin conductance responses (Kindt et al., 2009), and other studies finding no effect on either physiological measure (Bos et al., 2014).

A non-pharmacological approach developed by Monfils and colleagues (2009) takes advantage of the reconsolidation period by incorporating traditional extinction trials 10 minutes after reactivation of the consolidated fear memory. This technique was effective at preventing spontaneous recovery, renewal, reinstatement, and rapid reacquisition in rats (Monfils et al., 2009). The technique has also been effective at reducing the return of fear in humans, even one year after the original training and extinction (Schiller et al., 2010), and has also been applied to the domain of drug seeking behaviors in rats and humans (Xue et al., 2012), suggesting that it generalizes across both appetitive and aversive persistent

associations. Neurobiologically, extinction following reactivation has been shown to both induce plasticity-related changes in the LA in rodents (Clem and Huganir, 2010; Monfils et al., 2009), and reduce involvement of vmPFC inhibitory networks during extinction in humans (Schiller et al., 2013), consistent with the notion that this behavioral intervention is targeting the original fear memory.

This extinction following reactivation approach has received considerable attention, largely due to its straightforward potential as a therapeutic strategy. A major advantage is that the technique does not depend on pharmacological agents. However, some laboratories have failed to show an effect of extinction following reactivation in rodents (Chan et al., 2010) and humans (Golkar et al., 2012). From a theoretical standpoint, an important question that remains unclear is why extinction trials following an earlier (CS-alone) memory-reactivation trial overwrite prior learning, while extinction trials without an earlier reactivation trial initiate new learning. That is, it is not clear why the first trial of standard extinction does not act as a reactivation trial, leading to subsequent updating of the fear memory by extinction trials that fall within the reconsolidation time window (Delamater and Westbrook, 2014).

More broadly, across both pharmacological and behavioral techniques aimed at targeting reconsolidation, the precise conditions that initiate reconsolidation of the original memory trace following reactivation remain unclear (Suzuki et al., 2004). This brings these procedures back into contact with theoretical models of learning. That is, although consolidation and reconsolidation have mostly been discussed in terms of the biological processes of plasticity, this focus is not inconsistent with simultaneously conceptualizing them in the computational terms of statistical or associative learning models, whose principles are implemented by the biological processes. Specifically, reconsolidation (and the concomitant susceptibility to memory alteration) may arise under circumstances when experiences would lead to a memory being retrieved and modified, as opposed to a new memory being created. If so, this suggests that the same sorts of statistical factors that modulate the dominance of old versus new latent causes in Gershman et al.’s (2010) model will also affect the susceptibility to reconsolidation.

Consistent with this idea that susceptibility to reconsolidation varies based on the learning context, recent research suggests several boundary conditions that limit the effectiveness of targeting reconsolidation, many which bear similarity to situations highlighted in latent cause models as leading to the formation of a new memory (i.e. inferring a new latent cause), as opposed to memory modification (i.e. inferring the old latent cause). These include strength and generalization of initial learning and age of the memory (Clem and Huganir, 2010; Suzuki et al., 2004; Taubenfeld et al., 2009). Even more directly paralleling memory modification in latent cause models, one factor that may initiate memory destabilization and reconsolidation is the detection of prediction errors due to mismatch in the expectations of CS-US association between initial acquisition and memory reactivation (Díaz-Mataix et al., 2013).

Gradual extinction—Another behavioral method for modification of the original fear memory builds more directly on the latent cause framework. According to this framework, the large difference in observed stimuli at extinction (CS only) compared to acquisition (CS

and US) provides evidence for the existence of a new latent cause at the beginning of extinction. Once a new latent cause is inferred, further extinction trials are attributed in large part to that new latent cause, and thus extinction learning is no longer applied to the original latent cause. In essence, the new latent cause serves to protect the original memory from new learning. The model thus predicts that making extinction more similar to acquisition should help prevent inference of a new latent cause, and direct all learning in the extinction phase to the old cause, resulting in updating of the original fear memory. Gershman and colleagues tested this prediction using a 'gradual extinction' technique in which some extinction trials were reinforced with a US. The frequency of reinforced trials diminished throughout the extinction session, essentially 'weaning' the rats off the shock US. This technique was effective in preventing spontaneous recovery and reinstatement in rats (Gershman et al., 2013). Notably, rats who received five shocks during gradual extinction exhibited less recovery and reinstatement than rats who received no shocks during traditional extinction. These results are consistent with demonstrations that rapid extinction is actually accompanied by more spontaneous recovery (Gershman and Hartley, 2015), and suggest that clinical practices that aim to speed up extinction might actually be counterproductive (Craske et al., 2008).

General issues in targeting the fear memory—A practical concern for therapeutic use of approaches that target the CS-US memory directly is that fearful experiences are prone to generalization beyond the details of the CS (Dunsmoor and Paz, 2015). However, in rats, higher-order (indirectly associated) fear memories do not become labile via reactivation of the first-order (direct CS-US association) fear memory (Debiec et al., 2006), thus it is not clear how to target the whole network of related memories. Of course, it makes sense that reconsolidation would be specific to the actual reactivated memory—presumably, the role of reconsolidation is to update this memory with new relevant knowledge, not to erroneously alter or update all other memories within that network. For treatment purposes, however, the ability to update fear memories at a generalized level is desired, since fear memories consist of multiple elements that become interweaved within a broad associative network (Dunsmoor and Murphy, 2015). For example, in PTSD, panic, phobias, and other anxiety disorders, a multitude of objects, places, sensations, or abstract concepts can act as triggers to induce anxiety symptoms (e.g., Bouton et al., 2001). Recent research suggests that one approach to target a broader associative fear network is to use the US, as opposed to CS, as a reactivation cue (Liu et al., 2014).

Importantly, the goal of most anxiety treatments is not to eliminate memories altogether, but to make negative memories less persistent and intrusive, and to decouple episodic content from emotional responses. According to the influential 'emotional processing theory' (Foa and Kozak, 1986), fear representations are cognitive in nature and are maintained within informational structures (fear structures, or schemas). Activating fear structures during therapy allows corrective information to weaken the association between informational elements and fear responses. Although treatment approaches differ among clinicians (e.g., Craske et al., 2008), this model of fear memory remains a dominant view that continues to guide anxiety research and psychiatric treatment. Importantly, the intent of effective exposure therapy is to target only the pathological elements of a memory structure.

Finally, it is important to consider whether positive findings of techniques that putatively target the CS-US association are due to memory modification or erasure, or, alternatively, are due to strengthened inhibitory learning (Lattal and Wood, 2013). At this stage, there is not a definitive neurobiological marker of persistent memory alteration, and the absence of the CR at tests of return of fear is necessary but not sufficient evidence of erasure.

Strengthening extinction

Compound or 'deepened' extinction—As nearly all associative models describe extinction as new inhibitory learning rather than unlearning, many efforts to prevent the return of fear focus on ways to promote better extinction learning so that the association learned in extinction later outcompetes the original fear memory for expression. One recently developed strategy is to conduct extinction in the presence of another fear conditioned stimulus (i.e., a second exciter). In this technique, two or more CSs (e.g., CSA and CSB, a light and a tone) are paired separately with the US. Next, one CS is presented during extinction (or both CSs are extinguished separately in an alternate version of this task; Leung et al., 2012). The two CSs are then combined (e.g., a light/tone compound), and extinction continues with the compound. Rescorla (2000) was the first to demonstrate that extinction is enhanced by the presence of an additional excitatory CS, a technique referred to as *deepened extinction* (Rescorla, 2006). In a number of subsequent animal studies, deepened extinction was shown to reduce spontaneous recovery, reinstatement, reacquisition, and renewal (e.g., Leung et al., 2012), and reduced spontaneous recovery of conditioned SCRs in a recent human fear conditioning study (Culver et al., 2014).

In theoretical terms, the key principle of deepened extinction seems to be summation, i.e., the idea that when two CSs are presented together, the net US expectancy, and therefore the potential prediction error, reflects the sum over both CS's separate associations (Rescorla and Wagner, 1972). Here, when CS A and CS B are trained and CS A is extinguished, presenting the two CSs together increases the (joint) prediction of the US. The prediction error generated by the continued omission of the US in the face of this heightened US expectancy can therefore decrease the associative value of CS A below the level it could attain had it continued to be extinguished alone. Rescorla-Wagner's summation principle even allows stimuli to acquire negative associative strength, i.e. to cancel US predictions that would otherwise be expected. This effect – known as conditioned inhibition – traditionally arises when an otherwise US-predicting stimulus is paired with a neutral stimulus and no US is presented. The neutral stimulus then acquires inhibitory (negative) associative strength, which can serve to 'cancel' the positive predictions of other concurrently presented stimuli. Negative prediction errors during AB pairings, after A is already extinguished, may thus make CS A a conditioned inhibitor, contributing to the deepened extinction effect (Leung et al., 2012).

However, summation effects are not ubiquitous in conditioning, occurring in some circumstances but not others. Statistical models of conditioning can explain boundary conditions and apparently arbitrary effects of experimental protocols on prediction learning. For instance, Soto and colleagues (2014) generalized Gershman et al.'s (2010) model to allowing for multiple latent causes to be active simultaneously (e.g., one each for CS A and

CS B), capturing summation and related effects such as conditioned inhibition This model could effectively explain the results of a wide variety of summation and generalization experiments, within one statistical learning framework (Courville et al., 2005; Courville et al., 2003). Latent cause theories may therefore be helpful in understanding what circumstances promote or oppose the deepened-extinction effect. Another avenue for future work is to clarify how the events in a deepened extinction protocol affect the tradeoff between modification and interference, which would also affect the efficacy of deepened extinction.

Massive extinction training—Pavlov observed that continuing extinction beyond the point that the animal has stopped responding reduces spontaneous recovery (i.e., *silent extinction beyond the zero;* Pavlov 1927). More recently, Denniston et al. (2003) showed that a substantial number of extinction trials (800) diminished contextual renewal in rats, providing evidence that extinction learning can be expressed outside the extinction context if training is immense. According to interference models of extinction, however, the depth of extinction training should have minimal effect on expression of the CS-US association outside the extinction context; that is, even if inhibitory learning is extraordinarily strong, the context is still expected to gate expression of the extinction memory (Bouton et al., 2006; Maren et al., 2013). In contrast, in latent cause models, it is possible that massive extinction increases the generality of the interfering memory (e.g. via increasing its prior probability and its generality over different temporal contexts). The concept of massive extinction relates clinically to prolonged exposure therapy, an effective treatment for PTSD (Powers et al., 2010).

Exposure to novelty—A cornerstone of most associative learning models is that learning is induced by the presence of novel or surprising events (Pearce and Hall, 1980; Rescorla and Wagner, 1972). A number of neural systems crucial for attention, learning, and memory respond to novel events, and memory systems seem to favor consolidation of novel information (Lisman et al., 2011). Novel neutral stimuli also promote dopaminergic responses, and surprising events activate a number of other neuromodulators including norepinephrine and acetylcholine, which may also be involved in gating learning (e.g., Yu and Dayan, 2005). As novelty appears to promote learning processes, one approach to strengthening extinction is by increasing novelty during or around the time of learning.

Recently, Dunsmoor et al. (2014) used novel events to augment extinction by replacing, rather than simply omitting, an aversive electrical shock US with a surprising non-aversive outcome (a tone). Compared to groups that received traditional extinction training through shock omission alone, this modified extinction paradigm, referred to as *novelty-facilitated extinction*, reduced spontaneous recovery of conditioned skin conductance responses in humans and freezing in rats at a 24-hour test of extinction retention. One possibility for the effectiveness of this procedure is that shock omission reduced the CR while, simultaneously, the novel tone maintained attention to the CS, increasing its associability and therefore the rate at which the associative strength of the CS was updated in extinction. This effect is captured in statistical models, including latent cause models, by the idea that new learning should be gated by uncertainty about previously inferred causal structures (Courville et al.,

2006; Dayan et al., 2000). Indeed, the core of statistical inference models is formally tracking and manipulating uncertainty, thus these models may help to design protocols that maximally leverage uncertainty in the service of extinction. This is especially important in light of the fact that novelty during extinction may enhance learning of a competing association (a new latent cause), at the expense of modification of the original fear memory – a tradeoff that should be carefully titrated.

Another, related line of research uses novelty exposure either before or after fear extinction to enhance memory consolidation (reviewed in Moncada et al., 2015). In one design (de Carvalho Myskiw et al., 2013), animals are initially fear conditioned to a context (cage), a hippocampal-dependent form of learning (Maren et al., 2013). Rats are then *weakly* extinguished 24 hours later by leaving them in the cage for 10 minutes without any shocks. Weak training leads to short-term reductions in freezing, but does not lead to a long-term extinction memory as demonstrated by near-complete recovery of freezing the next day. However, rats who explore a novel open field 1 or 2 hours before or 1 hour after weak extinction training showed significantly less freezing at a long-term memory test than rats without novelty-exploration.

Why does exposure to a novel open field enhance weak memory of a separate experience like contextual fear extinction? The answer may lie in an evolving neurobiological view of how salient experiences strengthen memory for weakly learned experiences occurring around the same time. Frey and Morris (1997) proposed a process by which action potentials at a synapse induce an early phase of long-term potentiation (LTP) that initiates a local synaptic tag. This tag represents the potential for lasting change, but only if it is “captured” by plasticity-related proteins required for late LTP and thus long-term memory. These proteins can be induced by activity in a shared neural ensemble prior or following initiation of the tag, in a time delimited manner (see Redondo and Morris, 2011 for review).

One way such “synaptic tagging” may work at the behavioral level is that ‘strong’ experiences boost consolidation for weakly learned behavioral experiences occurring around the same time and that involve similar neural substrates. For instance, novelty-exploration benefits long-term memory for weakly learned hippocampal-dependent tasks like context conditioning (Ballarini et al., 2009), context extinction (de Carvalho Myskiw et al., 2013), object recognition (Ballarini et al., 2009), and inhibitory avoidance (Moncada et al., 2011). Critically, exploration of novel, but not familiar, environments upregulates immediate early gene expression (Li et al., 2003) and dopamine release in the CA1 region of the dorsal hippocampus (Lisman et al., 2011). Indeed, exposure to a familiar open field, blockade of protein synthesis in CA1, or blockade of hippocampal D1/D5 dopamine receptors before or following novelty exposure prevents behavioral tagging effects (de Carvalho Myskiw et al., 2013; Moncada et al., 2011). As behavioral tagging effects may be a general process of long-term memory consolidation across species (Ballarini et al., 2009; Dunsmoor et al., 2015), one intriguing possibility to enhance extinction is to combine extinction with other novel or rewarding tasks that recruit regions involved in extinction consolidation, like the vmPFC. In line with this idea, post-extinction administration of L-DOPA has been shown to strengthen extinction in rats and humans (Haaker et al., 2013).

Stressor Controllability and Active Avoidance—It has long been known that exposure to uncontrollable stress later results in a host of maladaptive behavioral responses and health consequences (i.e., learned helplessness; Maier and Seligman, 1976). Less known is that exposure to controllable stress can enhance behavioral performance and neurochemical responses to subsequent stress (Williams and Maier, 1977). More recently it has been demonstrated that the benefits of stressor controllability extends to enhanced extinction learning and reduced spontaneous recovery, relative to no-stress controls (Baratta et al., 2007). In these paradigms, stress is operationalized as exposure to shocks that the animal can either avoid (escapable shock; ES) or not (inescapable shock; IS). Using a triadic design that included ES, IS and no shock groups, Baratta et al. (2007) found that a session of ES in a different context 24 hours after fear conditioning facilitated subsequent extinction learning relative to IS and control groups, and eliminated spontaneous recovery. A similar study in humans found that an ES session a week prior to fear conditioning, extinction and a spontaneous recovery test also enhanced extinction relative to IS and control groups, and eliminated spontaneous recovery (Hartley et al., 2014). Stressor controllability effects have been shown to depend on plasticity within the vmPFC, which facilitates inhibitory control over brainstem nuclei and the amygdala (Maier and Watkins, 2010). Injecting muscimol into the vmPFC during ES eliminates any benefit on later conditioned fear expression (Baratta et al., 2007). These results suggest that stressor controllability may augment extinction via a general, lasting facilitation of the mechanisms of fear inhibition.

Similar effects of reducing fear recovery occur in studies of active avoidance. In signaled active avoidance, after fear conditioning a rodent learns a behavioral response in the presence of the CS to avoid the US (Moscarello and LeDoux, 2013). In escape from fear, a rodent learns a behavioral response to avoid the CS (Cain and LeDoux, 2007). Interestingly compared to rodents who undergo standard extinction training, both of these paradigms result in the elimination of later spontaneous recovery, even though during the spontaneous recovery test there was no opportunity to avoid the US or CS. In other words, the active avoidance experience during extinction, like stressor controllability, enhances future fear control. Furthermore, Moscarello and LeDoux (2013) showed that injection of a protein synthesis inhibitor into either IL or CE impaired or facilitated active avoidance, respectively. These data support a model in which active avoidance learning recruits IL to inhibit CE-mediated conditioned fear behaviors, leading to a robust suppression of conditioned responding that generalizes across contexts.

Although on the surface it may seem that active avoidance and escape from fear paradigms could result in protection from extinction in which a new behavior that eliminates the CS or US prevents extinction learning (see *Psychological and Cognitive Factors* for a description), recent research on stressor controllability suggests a key difference between the augmentation of extinction with active avoidance and the impairment of extinction observed in protection from extinction paradigms is the subjective perception of internal control in eliminating the presentation of the CS or US (Hartley et al., 2014). In active avoidance and stressor controllability the source of eliminating the aversive event is attributed to the learned actions of the animal, whereas in protection from extinction the source is attributed to external circumstances.

Pharmacological Enhancement—Over the last decade, a broad range of neuropharmacological tools have been suggested to enhance learning and memory processes during extinction to help prevent the return of extinguished behavior. These include agents acting on a variety of neurotransmitter systems, including modulation of glutamatergic and GABAergic receptors, and modulators of the monoamine, cholinergic, cannabinoid, and steroid hormone systems (see Fitzgerald et al., 2014 for a full review). As our understanding of the cellular and systems neuroscience of fear extinction improves even further through the use of tools with temporal and spatial precision like optogenetics (Do-Monte et al., 2015), pharmacological agents will become increasingly directed to specific neural targets to modulate extinction learning. As it currently stands, pharmaceutical adjuncts to extinction learning in humans tend to incorporate systematic administration of putative cognitive enhancers, most prominently the partial NMDA agonist D-cycloserine (DCS).

In rodents, both systematic administration of DCS or infusion into the BLA directly either before or after extinction training enhances learning (Fitzgerald et al., 2014). These findings have been translated to the clinic-based exposure therapies in humans. For example, in an initial demonstration, patients suffering from acrophobia (fear or heights) who were given DCS showed similar improvement in symptoms after two exposure therapy sessions, as participants given a placebo demonstrated after seven sessions (Ressler et al., 2004). Since this study, the benefits of DCS in augmenting exposure therapy has been documented for a number of different anxiety disorders, although its efficacy may be limited to clinically significant disorders and initial exposure training (see Myers et al., 2011 for a review).

Improving retrieval of the extinction memory

Finally, a complementary approach to strengthening within-session extinction learning is to promote the retrieval of extinction memories at test. In the memory literature, retrieval is enhanced if the encoding and retrieval context are similar, an effect known as encoding specificity (Tulving and Thomson, 1973). Hence, the goal of these approaches, broadly speaking, is to enhance similarity to extinction training so that retrieval favors the inhibitory CS-no US association and not the original CS-US association.

Retrieval cues—One approach to promote retrieval of the extinction memory is to place a cue at extinction that is also present at test. Using appetitive conditioning in rats, Brooks and Bouton showed that extinction cues reduce spontaneous recovery (Brooks and Bouton, 1993) and renewal (Brooks and Bouton, 1994). However, the mechanisms by which extinction retrieval cues function to reduce the return of fear are not entirely clear. From an associative learning framework, extinction cues may act as occasion setters helping to retrieve the CS-noUS association. But extinction cues may also become conditioned inhibitors (‘safety signals’) that could interfere with effective extinction (Lovibond et al., 2000; Rescorla, 2003). Additional steps can help ensure that extinction cues do not turn into conditioned inhibitors. For instance, Brooks and Bouton (1993, 1994) paired extinction cues with the CS on some, but not all extinction, trials. Additionally, the extinction cue was presented several seconds prior to the CS and the two stimuli did not overlap. This also avoided the potential for the CS to be processed as an entirely unique cue. In sum, retrieval

cues may be most effective as reminders of the extinction session, helping “bridge” extinction and test (Laborda et al., 2011).

Multiple contexts—Another approach to reduce the contextual-specificity of fear extinction is to conduct extinction across multiple contexts. From the point of view of latent cause models, the rationale for this is to increase the scope and generality of the extinction cause, so that it matches many contexts and is not tied to a single one. Accordingly, relative to single context extinction, multi-context extinction reduces fear renewal (e.g., Gunther et al., 1998; Shiban et al., 2013) and reinstatement (Dunsmoor et al., 2014a). In essence, extinction under multiple contexts increases the chance that cues present at extinction will be present at test, therefore promoting generalization, similar to the use of extinction cues. Extinction over multiple contexts may also reduce the chance that the context would acquire inhibitory properties and block the CS from complete extinction. As compared to extinction in a single context, switching between contexts may also help maintain the high associability of the CS, as well as increase novelty - processes also in line with strengthening learning.

Silencing the hippocampus—The dorsal hippocampus gates expression of the extinction memory so that extinction is usually confined to the context where it occurred (Maren et al., 2013). One technique to reduce the context-specificity of extinction is to therefore temporarily inactive or down-regulate the hippocampus. In rats, reversible inactivation of the dorsal hippocampus with muscimol after extinction training, prior to test, prevented fear renewal to an extinguished CS when it was tested in a novel environment, but did not prevent renewal in the acquisition context (reviewed in Bouton et al., 2006). The latter finding was consistent with earlier studies showing that permanent hippocampal lesions made prior to fear conditioning did not prevent renewal when tested in the acquisition context (Frohardt et al., 2000; Wilson et al., 1995). Yet, in another study, permanent electrolytic lesions to the dorsal hippocampus in rats, either prior to fear conditioning or following extinction, reduced fear renewal to the CS irrespective of the test context (Ji and Maren, 2005).

In humans, methods that are more practical include pharmacological manipulations with minimal risk. One potential agent is scopolamine, a cholinergic antagonist used to treat motion sickness that also disrupts context-dependent learning in rats (Anagnostaras et al., 1995). Pharmacological disruption of the hippocampus with scopolamine prevents renewal in rats when administered prior, but not following extinction (Zelikowsky et al., 2013). Disrupting the hippocampus during extinction may prevent the context from being fully processed, making learning context-independent and therefore resilient to context shifts. Whether procedures that target hippocampal activity, like pre-extinction administration of scopolamine, are effective in humans awaits study.

Conclusions

Extinction of conditioned responses is one of the oldest and most widely known findings from psychological science. And yet researchers continue to make new discoveries that illuminate behavioral and neurobiological mechanisms underlying the disruption of prior learning. Important questions remain, but a surge of interest in extinction across a number of

psychological and neuroscience domains have started to tackle issues relevant to the disruption of unwanted behaviors and persistent alteration of fear memories. In this review, we highlighted advances in our understanding of extinction and discussed areas that warrant a reexamination. This includes the question of whether extinction always yields a new inhibitory memory trace that competes against the original CS-US association, and what conditions lead to persistent alteration of a memory trace. Beyond targeting the original memory trace, a host of recently developed techniques can compensate for the shortfalls of traditional extinction protocols as a tool to prevent the return of fear. These techniques have clear implications for improving clinical treatment for fear and anxiety disorders. Finally, these techniques, together with new statistical conceptualizations of learning and unlearning, can illuminate central mechanisms implicated in learning and memory above and beyond the phenomenon of extinction.

Acknowledgments

This work was funded by NIMH grants K99MH106719 to JED and RO1MH097085 to EAP and by Army Research Office grant W911NF-14-1-0101 to YN. The information in this manuscript does not necessarily reflect the opinion or policy of the federal government and no official endorsement should be inferred.

References

Amsel A. The role of frustrative nonreward in noncontinuous reward situations. Psychological bulletin. 1958; 55:102–119. [PubMed: 13527595]

Anagnostaras, SG.; Maren, S.; Fanselow, MS. Scopolamine selectively disrupts the acquisition of contextual fear conditioning in rats. 1995.

Ballarini F, Moncada D, Martinez MC, Alen N, Viola H. Behavioral tagging is a general mechanism of long-term memory formation. Proc Natl Acad Sci U S A. 2009; 106:14599–14604. [PubMed: 19706547]

Baratta M, Christianson J, Gomez D, Zarza C, Amat J, Masini C, Watkins L, Maier S. Controllable versus uncontrollable stressors bi-directionally modulate conditioned but not innate fear. Neuroscience. 2007; 146:1495–1503. [PubMed: 17478046]

Barto, AG. Adaptive Critics and the Basal Ganglia. In: Houk, JC.; Davis, JL.; Beiser, DG., editors. Models of information processing in the basal ganglia. Cambridge: MIT Press; 1995. p. 215-232.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nature neuroscience. 2007; 10:1214–1221. [PubMed: 17676057]

Bisson JI, Jenkins PL, Alexander J, Bannister C. Randomised controlled trial of psychological debriefing for victims of acute burn trauma. The British journal of psychiatry : the journal of mental science. 1997; 171:78–81. [PubMed: 9328501]

Bos MGN, Beckers T, Kindt M. Noradrenergic blockade of memory reconsolidation: a failure to reduce conditioned fear responding. Frontiers in Behavioral Neuroscience. 2014:8. [PubMed: 24478655]

Bouton ME. Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. Psychological bulletin. 1993; 114:80–99. [PubMed: 8346330]

Bouton ME. Context and behavioral processes in extinction. Learn Mem. 2004; 11:485–494. [PubMed: 15466298]

Bouton ME, Mineka S, Barlow DH. A modern learning theory perspective on the etiology of panic disorder. Psychological Review. 2001; 108:4–32. [PubMed: 11212632]

Bouton ME, Westbrook RF, Corcoran KA, Maren S. Contextual and temporal modulation of extinction: Behavioral and biological mechanisms. Biological psychiatry. 2006; 60:352–360. [PubMed: 16616731]

Brandon SE, Vogel EH, Wagner AR. A componential view of configural cues in generalization and discrimination in Pavlovian conditioning. Behavioural brain research. 2000; 110:67–72. [PubMed: 10802304]

Brooks DC, Bouton ME. A retrieval cue for extinction attenuates spontaneous recovery. Journal of Experimental Psychology: Animal Behavior Processes. 1993; 19:77. [PubMed: 8418218]

Brooks DC, Bouton ME. A retrieval cue for extinction attenuates response recovery (renewal) caused by a return to the conditioning context. Journal of Experimental Psychology: Animal Behavior Processes. 1994; 20:366.

Brunet A, Orr SP, Tremblay J, Robertson K, Nader K, Pitman RK. Effect of post-retrieval propranolol on psychophysiologic responding during subsequent script-driven traumatic imagery in post-traumatic stress disorder. J Psychiatr Res. 2008; 42:503–506. [PubMed: 17588604]

Burgos-Robles A, Vidal-Gonzalez I, Santini E, Quirk GJ. Consolidation of fear extinction requires NMDA receptor-dependent bursting in the ventromedial prefrontal cortex. Neuron. 2007; 53:871–880. [PubMed: 17359921]

Cain CK, LeDoux JE. Escape from fear: a detailed behavioral analysis of two atypical responses reinforced by CS termination. Journal of Experimental Psychology: Animal Behavior Processes. 2007; 33:451. [PubMed: 17924792]

Chan WYM, Leung HT, Westbrook RF, McNally GP. Effects of recent exposure to a conditioned stimulus on extinction of Pavlovian fear conditioning. Learning & memory (Cold Spring Harbor, NY). 2010; 17:512–521.

Chhatwal JP, Myers KM, Ressler KJ, Davis M. Regulation of gephyrin and GABAA receptor binding within the amygdala after fear acquisition and extinction. The Journal of neuroscience. 2005; 25:502–506. [PubMed: 15647495]

Clem RL, Huganir RL. Calcium-permeable AMPA receptor dynamics mediate fear memory erasure. Science. 2010; 330:1108–1112. [PubMed: 21030604]

Corcoran KA, Desmond TJ, Frey KA, Maren S. Hippocampal inactivation disrupts the acquisition and contextual encoding of fear extinction. The Journal of neuroscience. 2005; 25:8978–8987. [PubMed: 16192388]

Courville AC, Daw ND, Touretzky DS. Similarity and discrimination in classical conditioning: A latent variable account. Advances in neural information processing systems. 2005; 17:313–320.

Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. Trends in cognitive sciences. 2006; 10:294–300. [PubMed: 16793323]

Courville AC, Gordon GJ, Touretzky DS, Daw ND. Model uncertainty in classical conditioning. Advances in neural information processing systems. 2003 None.

Craske MG, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A. Optimizing inhibitory learning during exposure therapy. Behaviour research and therapy. 2008; 46:5–27. [PubMed: 18005936]

Craske MG, Treanor M, Conway CC, Zbozinek T, Vervliet B. Maximizing exposure therapy: An inhibitory learning approach. Behaviour research and therapy. 2014; 58C:10–23. [PubMed: 24864005]

Culver NC, Vervliet B, Craske MG. Compound Extinction: Using the Rescorla–Wagner Model to Maximize Exposure Therapy Effects for Anxiety Disorders. Clinical Psychological Science. 2014

Daw ND, Courville AC, Dayan P. Semi-rational models of conditioning: The case of trial order. The probabilistic mind. 2008:431–452.

Daw ND, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Networks. 2002; 15:603–616. [PubMed: 12371515]

Dayan P, Kakade S, Montague PR. Learning and selective attention. Nature Neuroscience. 2000; 3:1218–1223. [PubMed: 11127841]

Dayan P, Long T. Statistical models of conditioning. Advances in neural information processing systems. 1998:117–123.

de Carvalho Myskiw J, Benetti F, Izquierdo I. Behavioral tagging of extinction learning. Proceedings of the National Academy of Sciences USA. 2013; 110:1071–1076.

Debiec J, Doyere V, Nader K, LeDoux JE. Directly reactivated, but not indirectly reactivated, memories undergo reconsolidation in the amygdala. Proc Natl Acad Sci U S A. 2006; 103:3428–3433. [PubMed: 16492789]

Delamater AR. Issues in the extinction of specific stimulus-outcome associations in Pavlovian conditioning. Behavioural processes. 2012a; 90:9–19. [PubMed: 22465262]

Delamater AR. On the nature of CS and US representations in Pavlovian learning. Learn Behav. 2012b; 40:1–23. [PubMed: 21786019]

Delamater AR, Westbrook RF. Psychological and neural mechanisms of experimental extinction: a selective review. Neurobiology of learning and memory. 2014; 108:38–51. [PubMed: 24104049]

Denniston JC, Chang RC, Miller RR. Massive extinction treatment attenuates the renewal effect. Learn Motiv. 2003; 34:68–86.

Díaz-Mataix L, Martinez RCR, Schafe GE, LeDoux JE, Doyère V. Detection of a temporal error triggers reconsolidation of amygdala-dependent memories. Curr Biol. 2013; 23:467–472. [PubMed: 23453952]

Dickinson A, Burke J. Within-compound associations mediate the retrospective revaluation of causality judgements. Q J Exp Psychol Sect B-Comp Physiol Psychol. 1996; 49:60–80.

Do-Monte FH, Manzano-Nieves G, Quiñones-Laracuente K, Ramos-Medina L, Quirk GJ. Revisiting the Role of Infralimbic Cortex in Fear Extinction with Optogenetics. The Journal of Neuroscience. 2015; 35:3607–3615. [PubMed: 25716859]

Dunsmoor JE, Ahs F, Zielinski DJ, LaBar KS. Extinction in multiple virtual reality contexts diminishes fear reinstatement in humans. Neurobiology of learning and memory. 2014a

Dunsmoor JE, Campese VD, Ceceli AO, LeDoux JE, Phelps EA. Novelty-Facilitated Extinction: Providing a Novel Outcome in Place of an Expected Threat Diminishes Recovery of Defensive Responses. Biological psychiatry. 2014b

Dunsmoor JE, LaBar KS. Brain activity associated with omission of an aversive event reveals the effects of fear learning and generalization. Neurobiology of learning and memory. 2012; 97:301–312. [PubMed: 22387662]

Dunsmoor JE, Murphy GL. Categories, concepts, and conditioning: how humans generalize fear. Trends in cognitive sciences. 2015

Dunsmoor JE, Murty VP, Davachi L, Phelps EA. Emotional learning selectively and retroactively strengthens memories for related events. Nature. 2015

Dunsmoor JE, Paz R. Fear generalization and anxiety: behavioral and neural mechanisms. Biological psychiatry. 2015

Fitzgerald PJ, Seemann JR, Maren S. Can fear extinction be enhanced? A review of pharmacological and behavioral findings. Brain research bulletin. 2014; 105C:46–60. [PubMed: 24374101]

Foa EB, Kozak MJ. Emotional processing of fear - exposure to corrective information. Psychological bulletin. 1986; 99:20–35. [PubMed: 2871574]

Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004; 306:1940–1943. [PubMed: 15528409]

Frey U, Morris RG. Synaptic tagging and long-term potentiation. Nature. 1997; 385:533–536. [PubMed: 9020359]

Frohardt RJ, Guarraci FA, Bouton ME. The effects of neurotoxic hippocampal lesions on two effects of context after fear extinction. Behav Neurosci. 2000; 114:227. [PubMed: 10832785]

Gelinas JN, Nguyen PV. β-Adrenergic receptor activation facilitates induction of a protein synthesis-dependent late phase of long-term potentiation. The Journal of neuroscience. 2005; 25:3294–3303. [PubMed: 15800184]

Gerber B, Yarali A, Diegelmann S, Wotjak CT, Pauli P, Fendt M. Pain-relief learning in flies, rats, and man: basic research and applied perspectives. Learning & memory (Cold Spring Harbor, NY). 2014; 21:232–252.

Gershman SJ, Blei DM, Niv Y. Context, learning, and extinction. Psychol Rev. 2010; 117:197–209. [PubMed: 20063968]

Gershman SJ, Hartley CA. Individual differences in learning predict the return of fear. Learn Behav. 2015:1–8. [PubMed: 25488021]

Gershman SJ, Jones CE, Norman KA, Monfils M-H, Niv Y. Gradual extinction prevents the return of fear: implications for the discovery of state. Frontiers in behavioral neuroscience. 2013:7. [PubMed: 23443667]

Gershman SJ, Niv Y. Exploring a latent cause theory of classical conditioning. Learn Behav. 2012; 40:255–268. [PubMed: 22927000]

Golkar A, Bellander M, Olsson A, Ohman A. Are fear memories erasable?-reconsolidation of learned fear with fear-relevant and fear-irrelevant stimuli. Frontiers in Behavioral Neuroscience. 2012:6. [PubMed: 22375108]

Gunther LM, Denniston JC, Miller RR. Conducting exposure treatment in multiple contexts can prevent relapse. Behaviour research and therapy. 1998; 36:75–91. [PubMed: 9613018]

Haaker J, Gaburro S, Sah A, Gartmann N, Lonsdorf TB, Meier K, Singewald N, Pape HC, Morellini F, Kalisch R. Single dose of L-dopa makes extinction memories context-independent and prevents the return of fear. Proc Natl Acad Sci U S A. 2013; 110:E2428–2436. [PubMed: 23754384]

Hartley CA, Gorun A, Reddan MC, Ramirez F, Phelps EA. Stressor controllability modulates fear extinction in humans. Neurobiology of learning and memory. 2014; 113:149–156. [PubMed: 24333646]

Hartley CA, Phelps EA. Changing Fear: The Neurocircuitry of Emotion Regulation. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2010; 35:136–146. [PubMed: 19710632]

Herry C, Trifilieff P, Micheau J, Lüthi A, Mons N. Extinction of auditory fear conditioning requires MAPK/ERK activation in the basolateral amygdala. Eur J Neurosci. 2006; 24:261–269. [PubMed: 16882022]

Hobin JA, Ji J, Maren S. Ventral hippocampal muscimol disrupts context-specific fear memory retrieval after extinction in rats. Hippocampus. 2006; 16:174–182. [PubMed: 16358312]

Huff NC, Hernandez JA, Blanding NQ, LaBar KS. Delayed Extinction Attenuates Conditioned Fear Renewal and Spontaneous Recovery in Humans. Behav Neurosci. 2009; 123:834–843. [PubMed: 19634943]

Hugues S, Chessel A, Lena I, Marsault R, Garcia R. Prefrontal infusion of PD098059 immediately after fear extinction training blocks extinction-associated prefrontal synaptic plasticity and decreases prefrontal ERK2 phosphorylation. Synapse. 2006; 60:280–287. [PubMed: 16786530]

Jenkins WO, Stanley JC. Partial reinforcement: a review and critique. Psychological bulletin. 1950; 47:193. [PubMed: 15417676]

Ji J, Maren S. Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction. Learn Mem. 2005; 12:270–276. [PubMed: 15930505]

Johansen, Joshua P.; Cain, Christopher K.; Ostroff, Linnaea E.; LeDoux, Joseph E. Molecular Mechanisms of Fear Learning and Memory. Cell. 2011; 147:509–524. [PubMed: 22036561]

Kakade, PDS. Explaining away in weight space. In Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference; MIT Press; 2001. p. 451

Kakade S, Dayan P. Acquisition and extinction in autoshaping. Psychological review. 2002; 109:533. [PubMed: 12088244]

Kamin LJ. Predictability, surprise, attention, and conditioning. Punishment and aversive behavior. 1969:279–296.

Kim J, Lee S, Park H, Song B, Hong I, Geum D, Shin K, Choi S. Blockade of amygdala metabotropic glutamate receptor subtype 1 impairs fear extinction. Biochemical and Biophysical Research Communications. 2007; 355:188–193. [PubMed: 17292864]

Kindt M, Soeter M, Vervliet B. Beyond extinction: erasing human fear responses and preventing the return of fear. Nature Neuroscience. 2009; 12:256–258. [PubMed: 19219038]

Laborda, MA.; McConnell, BL.; Miller, RR. Behavioral techniques to reduce relapse after exposure therapy: Applications of studies of experimental extinction. In: Schachtman, TR.; Reilly, S., editors. In Associative learning and conditioning theory: Human and non-human applications. New York, NY: Oxford University Press; 2011. p. 79-103.

Larrauri JA, Schmajuk NA. Attentional, associative, and configural mechanisms in extinction. Psychological review. 2008; 115:640–676. [PubMed: 18729595]

Lattal KM, Wood MA. Epigenetics and persistent memory: implications for reconsolidation and silent extinction beyond the zero. Nat Neurosci. 2013; 16:124–129. [PubMed: 23354385]

Le Pelley ME. The role of associative history in models of associative learning: A selective review and a hybrid model. Quarterly Journal of Experimental Psychology Section B. 2004; 57:193–243.

LeDoux JE. Emotion circuits in the brain. Annu Rev Neurosci. 2000; 23:155–184. [PubMed: 10845062]

Leung HT, Reeks LM, Westbrook RF. Two ways to deepen extinction and the difference between them. Journal of experimental psychology Animal behavior processes. 2012; 38:394–406. [PubMed: 23066980]

Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND. Differential roles of human striatum and amygdala in associative learning. Nature Neuroscience. 2011

Li S, Cullen WK, Anwyl R, Rowan MJ. Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. Nature neuroscience. 2003; 6:526–531. [PubMed: 12704392]

Lisman J, Grace AA, Duzel E. A neoHebbian framework for episodic memory; role of dopamine-dependent late LTP. Trends Neurosci. 2011; 34:536–547. [PubMed: 21851992]

Litz BT, Gray MJ, Bryant RA, Adler AB. Early Intervention for Trauma: Current Status and Future Directions. Clinical Psychology: Science and Practice. 2002; 9:112–134.

Liu J, Zhao L, Xue Y, Shi J, Suo L, Luo Y, Chai B, Yang C, Fang Q, Zhang Y. An unconditioned stimulus retrieval extinction procedure to prevent the return of fear memory. Biological psychiatry. 2014; 76:895–901. [PubMed: 24813334]

Lovibond PF. Cognitive processes in extinction. Learning & memory (Cold Spring Harbor, NY). 2004; 11:495–500.

Lovibond PF, Davis NR, O'Flaherty AS. Protection from extinction in human fear conditioning. Behaviour research and therapy. 2000; 38:967–983. [PubMed: 11004736]

Maier SF, Seligman MEP. Learned helplessness - theory and evidence. J Exp Psychol-Gen. 1976; 105:3–46.

Maier SF, Watkins LR. Role of the medial prefrontal cortex in coping and resilience. Brain research. 2010; 1355:52–60. [PubMed: 20727864]

Maren S. Nature and causes of the immediate extinction deficit: A brief review. Neurobiology of learning and memory. 2014; 113:19–24. [PubMed: 24176924]

Maren S, Chang CH. Recent fear is resistant to extinction. Proc Natl Acad Sci U S A. 2006; 103:18020–18025. [PubMed: 17090669]

Maren S, Phan KL, Liberzon I. The contextual brain: implications for fear conditioning, extinction and psychopathology. Nature reviews Neuroscience. 2013; 14:417–428. [PubMed: 23635870]

Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 2009; 459:837–841. [PubMed: 19448610]

Milad M, Vidal-Gonzalez I, Quirk G. Electrical stimulation of medial prefrontal cortex reduces conditioned fear in a temporally specific manner. Behav Neurosci. 2004; 118:389. [PubMed: 15113265]

Milad MR, Quirk GJ. Neurons in medial prefrontal cortex signal memory for fear extinction. Nature. 2002; 420:70–74. [PubMed: 12422216]

Milad MR, Quirk GJ. Fear extinction as a model for translational neuroscience: ten years of progress. Annual Review of Psychology. 2012; 63:129–151.

Miller RR, Barnet RC, Grahame NJ. Assessment of the Rescorla-Wagner model. Psychological bulletin. 1995; 117:363–386. [PubMed: 7777644]

Miller RR, Matute H. Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. Journal of Experimental Psychology: General. 1996; 125:370. [PubMed: 8945788]

Moncada D, Ballarini F, Martinez MC, Frey JU, Viola H. Identification of transmitter systems and learning tag molecules involved in behavioral tagging during memory formation. Proceedings of the National Academy of Sciences. 2011; 108:12931–12936.

Moncada D, Ballarini F, Viola H. Behavioral Tagging: A Translation of the Synaptic Tagging and Capture Hypothesis. Neural Plasticity. 2015; 501:650780. [PubMed: 26380117]

Monfils MH, Cowansage KK, Klann E, LeDoux JE. Extinction-Reconsolidation Boundaries: Key to Persistent Attenuation of Fear Memories. Science. 2009; 324:951–955. [PubMed: 19342552]

Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of neuroscience. 1996; 16:1936–1947. [PubMed: 8774460]

Morgan MA, Romanski LM, LeDoux JE. Extinction of emotional learning - contributions of medial prefrontal cortex. Neuroscience Letters. 1993; 163:109–113. [PubMed: 8295722]

Moscarello JM, LeDoux JE. Active avoidance learning requires prefrontal suppression of amygdala-mediated defensive reactions. The Journal of Neuroscience. 2013; 33:3815–3823. [PubMed: 23447593]

Myers KM, Carlezon WA, Davis M. Glutamate receptors in extinction and extinction-based therapies for psychiatric illness. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2011; 36:274–293. [PubMed: 20631689]

Myers KM, Ressler KJ, Davis M. Different mechanisms of fear extinction dependent on length of time since fear acquisition. Learn Mem. 2006; 13:216–223. [PubMed: 16585797]

Nader K, Schafe GE, Le Doux JE. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. Nature. 2000; 406:722–726. [PubMed: 10963596]

Pavlov, IP. Conditioned Reflexes. London: Oxford University Press; 1927.

Pearce JM, Hall G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review. 1980; 87:532–552. [PubMed: 7443916]

Pitman RK, Sanders KM, Zusman RM, Healy AR, Cheema F, Lasko NB, Cahill L, Orr SP. Pilot study of secondary prevention of posttraumatic stress disorder with propranolol. Biological psychiatry. 2002; 51:189–192. [PubMed: 11822998]

Plendl W, Wotjak CT. Dissociation of within-and between-session extinction of conditioned fear. The Journal of Neuroscience. 2010; 30:4990–4998. [PubMed: 20371819]

Powers MB, Halpern JM, Ferenschak MP, Gillihan SJ, Foa EB. A meta-analytic review of prolonged exposure for posttraumatic stress disorder. Clinical psychology review. 2010; 30:635–641. [PubMed: 20546985]

Quirk GJ, Likhtik E, Pelletier JG, Pare D. Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. J Neurosci. 2003; 23:8800–8807. [PubMed: 14507980]

Quirk GJ, Russo GK, Barron JL, Lebron K. The role of ventromedial prefrontal cortex in the recovery of extinguished fear. J Neurosci. 2000; 20:6225–6231. [PubMed: 10934272]

Razran G. Extinction re-examined and re-analyzed: a new theory. Psychol Rev. 1956; 63:39–52. [PubMed: 13289975]

Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychological review. 2007; 114:784. [PubMed: 17638506]

Redondo RL, Morris RG. Making memories last: the synaptic tagging and capture hypothesis. Nat Rev Neurosci. 2011; 12:17–30. [PubMed: 21170072]

Repa JC, Muller J, Apergis J, Desrochers TM, Zhou Y, LeDoux JE. Two different lateral amygdala cell populations contribute to the initiation and storage of memory. Nature neuroscience. 2001; 4:724–731. [PubMed: 11426229]

Rescorla RA. Reduction in the effectiveness of reinforcement after prior excitatory conditioning. Learn Motiv. 1970; 1:372–381.

Rescorla RA. Extinction can be enhanced by a concurrent excitor. Journal of experimental psychology Animal behavior processes. 2000; 26:251–260. [PubMed: 10913990]

Rescorla RA. Protection from extinction. Learn Behav. 2003; 31:124–132. [PubMed: 12882371]

Rescorla RA. Deepened extinction from compound stimulus presentation. Journal of experimental psychology Animal behavior processes. 2006; 32:135–144. [PubMed: 16634656]

Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement (Appleton-Century-Crofts). 1972

Ressler KJ, Rothbaum BO, Tannenbaum L, Anderson P, Graap K, Zimand E, Hodges L, Davis M. Cognitive enhancers as adjuncts to psychotherapy: use of D-cycloserine in phobic individuals to facilitate extinction of fear. Arch Gen Psychiatry. 2004; 61:1136–1144. [PubMed: 15520361]

Roesch MR, Calu DJ, Esber GR, Schoenbaum G. Neural correlates of variations in event processing during learning in basolateral amygdala. The Journal of Neuroscience. 2010; 30:2464–2471. [PubMed: 20164330]

Rothbaum BO, Kearns MC, Reiser E, Davis JS, Kerley KA, Rothbaum AO, Mercer KB, Price M, Houry D, Ressler KJ. Early intervention following trauma may mitigate genetic risk for PTSD in civilians: a pilot prospective emergency department study. The Journal of clinical psychiatry. 2014; 75:1380–1387. [PubMed: 25188543]

Santini E, Ge H, Ren K, de Ortiz SP, Quirk GJ. Consolidation of fear extinction requires protein synthesis in the medial prefrontal cortex. The Journal of neuroscience. 2004; 24:5704–5710. [PubMed: 15215292]

Schiller D, Cain CK, Curley NG, Schwartz JS, Stern SA, LeDoux JE, Phelps EA. Evidence for recovery of fear following immediate extinction in rats and humans. Learn Mem. 2008; 15:394–402. [PubMed: 18509113]

Schiller D, Kanen JW, Ledoux JE, Monfils MH, Phelps EA. Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. Proc Natl Acad Sci U S A. 2013

Schiller D, Monfils MH, Raio CM, Johnson DC, LeDoux JE, Phelps EA. Preventing the return of fear in humans using reconsolidation update mechanisms. Nature. 2010; 463:49–53. [PubMed: 20010606]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]

Sharp S, Thomas C, Rosenberg L, Rosenberg M, Meyer W III. Propranolol does not reduce risk for acute stress disorder in pediatric burn trauma. Journal of Trauma and Acute Care Surgery. 2010; 68:193–197.

Shiban Y, Pauli P, Muhlberger A. Effect of multiple context exposure on renewal in spider phobia. Behaviour research and therapy. 2013; 51:68–74. [PubMed: 23261707]

Soeter M, Kindt M. An abrupt transformation of phobic behavior following a post-retrieval amnesic agent. Biological psychiatry. 2015

Solomon RL, Wynne LC. Traumatic avoidance learning: the principles of anxiety conservation and partial irreversibility. Psychological review. 1954; 61:353. [PubMed: 13215688]

Soto FA, Gershman SJ, Niv Y. Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. Psychological review. 2014; 121:526. [PubMed: 25090430]

Sotres-Bayon F, Bush DE, LeDoux JE. Acquisition of fear extinction requires activation of NR2B-containing NMDA receptors in the lateral amygdala. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2007; 32:1929–1940. [PubMed: 17213844]

Suzuki A, Josselyn SA, Frankland PW, Masushige S, Silva AJ, Kida S. Memory reconsolidation and extinction have distinct temporal and biochemical signatures. The Journal of neuroscience. 2004; 24:4787–4795. [PubMed: 15152039]

Taubenfeld SM, Riceberg JS, New AS, Alberini CM. Preclinical assessment for selectively disrupting a traumatic memory via postretrieval inhibition of glucocorticoid receptors. Biological psychiatry. 2009; 65:249–257. [PubMed: 18708183]

Tulving E, Thomson DM. Encoding specificity and retrieval processes in episodic memory. Psychological review. 1973; 80:352.

Ungless MA, Magill PJ, Bolam JP. Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. Science. 2004; 303:2040–2042. [PubMed: 15044807]

Vaiva G, Ducrocq F, Jezequel K, Averland B, Lestavel P, Brunet A, Marmar CR. Immediate treatment with propranolol decreases posttraumatic stress disorder two months after trauma. Biological psychiatry. 2003; 54:947–949. [PubMed: 14573324]

Vervliet B, Craske MG, Hermans D. Fear extinction and relapse: state of the art. Annual review of clinical psychology. 2013; 9:215–248.

Williams JL, Maier SF. Transituational immunization and therapy of learned helplessness in the rat. Journal of Experimental Psychology: Animal Behavior Processes. 1977; 3:240.

Wilson A, Brooks DC, Bouton ME. The role of the rat hippocampal system in several effects of context in extinction. Behav Neurosci. 1995; 109:828. [PubMed: 8554708]

Wood NE, Rosasco ML, Suris AM, Spring JD, Marin MF, Lasko NB, Goetz JM, Fischer AM, Orr SP, Pitman RK. Pharmacological blockade of memory reconsolidation in posttraumatic stress disorder: Three negative psychophysiological studies. Psychiatry research. 2015; 225:31–39. [PubMed: 25441015]

Xue YX, Luo YX, Wu P, Shi HS, Xue LF, Chen C, Zhu WL, Ding ZB, Bao YP, Shi J, et al. A Memory Retrieval-Extinction Procedure to Prevent Drug Craving and Relapse. Science. 2012; 336:241–245. [PubMed: 22499948]

Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005; 46:681–692. [PubMed: 15944135]

Zelikowsky M, Hast TA, Bennett RZ, Merjanian M, Nocera NA, Ponnusamy R, Fanselow MS. Cholinergic blockade frees fear extinction from its contextual dependency. Biological psychiatry. 2013; 73:345–352. [PubMed: 22981655]

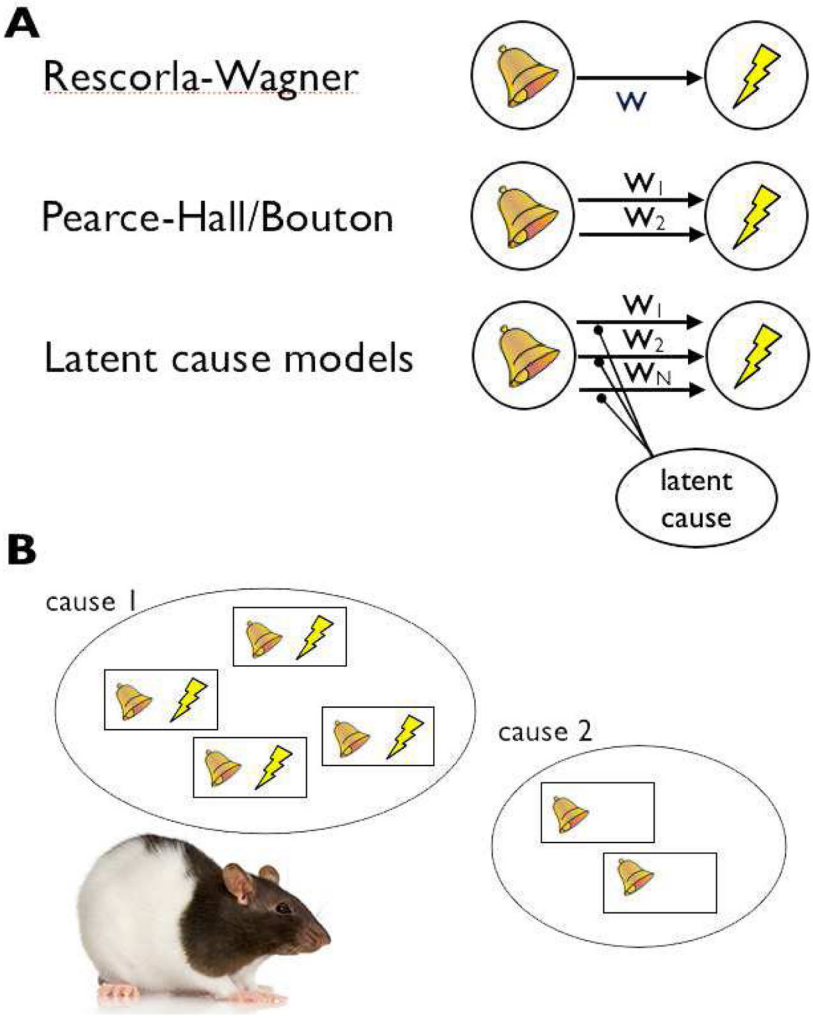

Figure 1. Simplified illustration of theoretical models of extinction

Different theoretical models of associative learning imply different processes in extinction. **A.** In the Rescorla-Wagner model (**top**), associative weights (w) between CSs and USs can increase and decrease based on prediction errors. Here acquisition involves a neutral weight (w=0) acquiring value (e.g., w = 1) over time. Extinction in this model causes 'unlearning' as the negative prediction errors due to the omission of the expected US decrease w back to zero. In contrast, in the Pearce-Hall or Bouton models (**middle**), extinction training causes learning of a new association, here denoted by a new weight w_2 that predicts the absence of the US. Thus extinction does not erase the value that w_1 acquired during the original training. The latent cause model (**bottom**) formalizes and extends this latter idea—here multiple associations (denoted by the arbitrary number N) can exist between a CS and a US, and inference about which latent cause is currently active affects how learning from the prediction error is distributed among these associations. In particular, the theory specifies the statistical conditions under which a new association (weight) is formed, and how learning on each trial is distributed among all existing weights. **B.** Another way to view the latent cause framework is as imposing a clustering of trials, before applying learning.

Similar trials are clustered together (i.e., attributed to the same latent cause), and learning of weights occurs within a latent cause (that is, each latent cause has its own weight). Note that while the illustration suggests that each trial (tone and shock, or tone alone) resides in one cluster only, this is an oversimplification. In practice, the model assigns trials to latent causes probabilistically (e.g., 90% to cause 1 and 10% to cause 2). Since on every trial there is some probability that a new latent cause has become active, the total number of clusters is equal to the number of trials so far; however, many clusters are effectively empty.

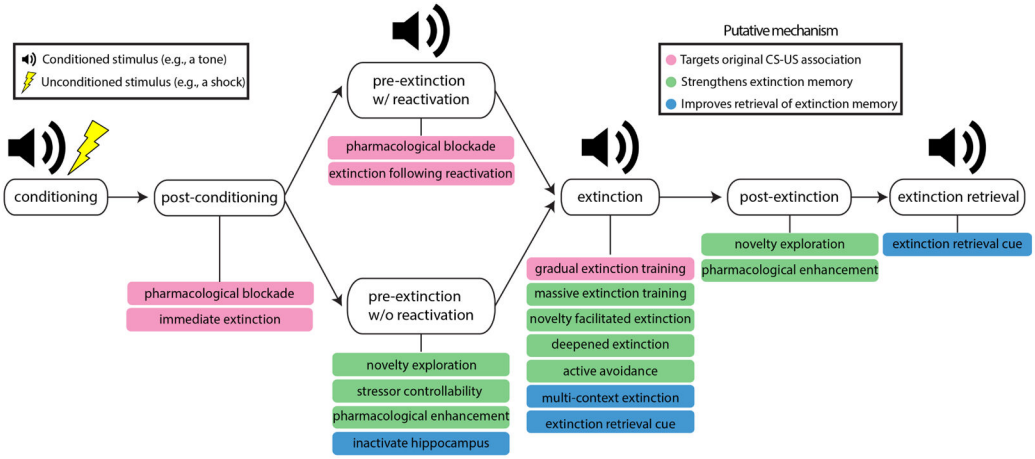

Figure 2.

Behavioral and pharmacological techniques to augment standard extinction, the time point at which each technique could be applied, and the putative mechanism by which each technique operates to prevent the return of unwanted behaviors.