Published in final edited form as:

Neurobiol Learn Mem. 2014 February ; 0: 38–51. doi:10.1016/j.nlm.2013.09.016.

Psychological and Neural Mechanisms of Experimental Extinction: A Selective Review

Andrew R. Delamater^{1,2} and R. Frederick Westbrook²

¹Brooklyn College of the City University of New York

²University of New South Wales

Abstract

The present review examines key psychological concepts in the study of experimental extinction and implications these have for an understanding of the underlying neurobiology of extinction learning. We suggest that many of the signature characteristics of extinction learning (spontaneous recovery, renewal, reinstatement, rapid reacquisition) can be accommodated by the standard associative learning theory assumption that extinction results in partial erasure of the original learning together with new inhibitory learning. Moreover, we consider recent behavioral and neural evidence that supports the partial erasure view of extinction, but also note shortcomings in our understanding of extinction circuits as these relate to the negative prediction error concept. Recent work suggests that common prediction error and stimulus-specific prediction error terms both may be required to explain neural plasticity both in acquisition and extinction learning. In addition, we suggest that many issues in the content of extinction learning have not been fully addressed in current research, but that neurobiological approaches should be especially helpful in addressing such issues. These include questions about the nature of extinction learning (excitatory CS-No US, inhibitory CS-US learning, occasion setting processes), especially as this relates to studies of the micro-circuitry of extinction, as well as its representational content (sensory, motivational, response). An additional understudied problem in extinction research is the role played by attention processes and their underlying neural networks, although some research and theory converge on the idea that extinction is accompanied by attention decrements (i.e., habituation-like processes).

Keywords

Rescorla-Wagner Model; Prediction Errors; Contents of Extinction Learning; Attention; Erasure; Amygdala; Infralimbic Prefrontal cortex

Over the last 15 years or so interest in the study of extinction in Pavlovian learning paradigms has exploded. Figure 1 shows the results of a PubMed search from 1927 to 2012 of the phrase “experimental extinction.” This figure shows that prior to the early 1960s there were relatively few publications on this topic (through this search engine), and that aside from a flurry of interest in the early 1960s (likely due to interest in Amsel’s frustration

© 2013 Elsevier Inc. All rights reserved.

Address Correspondence to: Andrew R. Delamater, Psychology Department, Brooklyn College – CUNY, 2900 Bedford Ave, Brooklyn, NY 11210, USA, andrewd@brooklyn.cuny.edu or Fred Westbrook (f.westbrook@unsw.edu.au).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

theory and Capaldi's sequential theory) there has been an exponential growth in more recent years in publications related to experimental extinction phenomena. This explosion in interest is due in large part to an increasing understanding of the neural mechanisms involved in simple fear conditioning, together with the increased recognition that such processes likely play a significant role in certain human psychopathologies (especially anxiety disorders, substance abuse, and eating disorders) as well as their treatment.

Generally speaking, in Pavlovian learning new conditioned responses (CR) are shown to develop to initially neutral conditioned stimuli (CS) as a result of these being repeatedly paired with biologically significant unconditioned stimuli (US). In fear conditioning with rodents, for instance, an auditory CS is usually paired with an aversive electric foot shock US and this results in freezing CRs developing to the CS. Reductions in these freezing CRs occur when the CS is repeatedly presented alone during an extinction phase. However, this reduction in CRs is often temporary because various experimental treatments have been shown to result in recovery of the CR. This creates the main theoretical problem for the learning theorist in explaining extinction, i.e., understanding what occurs both at a theoretical level as well as at the level of underlying neural mechanisms. We take the position that important clues to understanding the nature of the neural mechanisms of extinction learning can be gained by also achieving clarity on the nature of the psychological mechanisms at work in extinction. The two go hand in hand.

While Pavlovian extinction processes have most often been studied in recent years using the fear conditioning paradigm, other aversive learning paradigms as well as appetitive ones have also been used. In addition, extinction has also been extensively studied in various instrumental learning paradigms, and this has played an especially important role in modeling relapse in addiction. The scope of this review will be limited to an analysis of what we take to be the main concepts that have emerged especially in the study of extinction in Pavlovian paradigms (both appetitive and aversive), mostly using non-human subjects, and we will explore some of the implications these psychological concepts have for our understanding of the underlying neural mechanisms of extinction as they have been revealed in recent years.

Psychological Considerations in the Study of Extinction

The psychological study of extinction has a rich history. In the hands of different learning theorists, a number of different processes have been proposed to play an important role in extinction. Here we consider what we take to be several key ideas that either are currently important or will likely soon become important in the study of the neural mechanisms of extinction. We will start by discussing the issue of erasure versus new learning in extinction in connection with the "prediction error" concept. Then we will discuss issues regarding the representational content in extinction, and also touch upon the role of attention in extinction.

Erasure, New Learning, and Prediction Error

Theories that identified learning with a single construct such as the strength of an association or the connection weights between a CS and US explained extinction as the erasure of this strength or the resetting of these weights. The clearest statement in this regard came in the form of the model proposed by Rescorla and Wagner (1972). This model assumed there to be changes in the effectiveness of the US over the course of conditioning, such that it lost its potency in supporting new learning as it became accurately predicted. When, for example, an under-predicted US occurred such "positive prediction errors" produced increments in associative strength, and, conversely, when an over-predicted US occurred (or, as in the case of extinction, when a predicted US failed to occur) such "negative prediction errors" produced decrements in associative strength.

Textbook depictions of the Rescorla-Wagner model show that extinction eventually results in complete erasure of learning (i.e., complete associative decrements) and proceed to cite various response restoration phenomena as evidence that erasure does not occur. However, it is important to note that this model in fact predicts that in many circumstances associative loss in extinction should be far from complete. The model makes this prediction because it holds that associative change in extinction or acquisition is regulated by the discrepancy between the actual outcome of a trial and that predicted by all the cues present, including context. This departs from earlier accounts that assumed the discrepancy to be between the actual outcome of the trial and that predicted by only the cue in question (e.g., Bush & Mosteller, 1955). The former type of discrepancy is often referred to as a “common” error term, since it assumes that all stimuli present will contribute to and be influenced by this prediction error, but the latter type is referred to as an “individual” error term affecting only the cue in question. More will be said about this important distinction later. For now, evidence that a common error term regulates associative change in extinction was provided by Wagner (1969) who reported that a moderately trained CS was less able to elicit responding when extinguished in compound with a well-trained than with a weakly trained CS. The error produced by the compound containing the well-trained CS was greater than that by the one containing the weakly trained CS and, hence, change to the moderately trained CS was greater in the former than in the latter. Conversely, Rescorla (2003) reported that a CS was more able to elicit responding after being extinguished in compound with a conditioned inhibitor than with a stimulus that was equally familiar but lacked inhibitory properties. The error produced by extinction of the compound containing the conditioned inhibitor was smaller than that by the one containing the familiar stimulus, and, hence, change to the excitator was less in the former than in the latter.

Consider now the case of a CS that is accompanied by a novel cue during extinction. Under these circumstances the two CSs will undergo associative decrements with each extinction trial. The target CS will lose some of its associative strength, but the added CS will gain inhibitory associative strength. Importantly, however, when the inhibitory strength to this added CS equals the magnitude of excitatory strength possessed by the target CS, further extinction trials will be without any consequences (because the US will no longer be predicted to occur on the basis of the net associative strength of the two cues present). The target CS in this case will be “protected” from undergoing total associative loss (e.g., Rescorla, 2003). Notice also that when a CS undergoes extinction in a novel context, as in an ABA “renewal” experiment, for instance, the context, B, where extinction occurs can potentially acquire such inhibitory strength that could protect the CS from undergoing total associative loss (e.g., Polack, Laborda, & Miller, 2012). Hence, responding should be renewed when that CS is returned to context A for testing. This type of explanation is challenged by certain instances of a second type of renewal, ABC, where CS1 and CS2 are each trained in A, and then CS1 is extinguished in B and CS2 extinguished in C. Responding is greater when each CS is tested in the other’s extinction context than in its own, e.g., CS1 elicits more responding in C than B (e.g., Campese & Delamater, 2013; Harris, Jones, Bailey, & Westbrook, 2000; Maren & Hobin, 2007). Here, both B and C should have become inhibitors that protected their excitors and should have been interchangeable in their ability to suppress responding to either excitator. The protection from extinction explanation can accommodate these findings, however, if it is assumed that exposure to CS1 in B could have generated a unique cue, X, as could exposure to CS2 in C have generated such a cue, Y. Each of these cues would have acquired inhibition, leading to a suppression of responding when CS1 was tested in B (X was present) but a renewal of responding when tested in C (X was absent).

Consider also the case where a CS is trained, extinguished, and subjected to a reinstatement manipulation such as a US alone presentation, all in the same context (e.g., Rescorla &

Heth, 1975). Assuming the CS enters extinction with a greater associative strength than does context, the CS alone presentations will progressively reduce its excitatory associative strength but will also imbue the context with inhibitory strength, thereby again serving to protect the CS from complete erasure. The US alone presentations in the context remove the inhibition accrued to the context, thus allowing the residual excitatory associative value of the CS to express itself in responding, that is, reinstatement. Similarly, rapid reacquisition reflects the residual associative value of the CS. This associative value allows the CS to elicit responding more rapidly than a novel CS while at the same time undergoing less associative change because its value is closer to asymptote than the novel CS (Rescorla, 2002a).

Thus, many of the so-called signature characteristics of extinction merely show that some of the original association survives extinction; they do not show that the CS has acquired an additional opposing association. Indeed, many of these characteristics are consistent with the predictions from the Rescorla-Wagner model that extinction should result in some unlearning and some new inhibitory learning (see also Delamater, 2012). Figure 2 presents the results of Rescorla-Wagner model simulations of ABA, ABC, AAB renewal, and US reinstatement phenomena. In each case, conditioned responding to a stimulus (X) that had undergone extinction displays recovery when compared to its appropriate control condition. Figure 3 shows a more detailed breakdown of changes in associative strength to the various stimuli in a basic ABA renewal design. Essentially, it illustrates that the A context remains somewhat excitatory at the end of training, the B context becomes inhibitory during extinction, and the CS (X) is protected from losing all of its associative strength during extinction. When the CS is returned to the A context recovery arises because of the removal from the inhibitory B context together with the added strength of A's excitation combining with the CS's own residual excitation. It is worth noting that with some parameters, e.g., very long inter-trial intervals, the model predicts that the context will tend to lose its own associative value during the inter-trial interval, and this would diminish its overall contribution to test responding. However, whether or not extinction can undermine inhibitory strength to a stimulus is a topic that has been debated. Indeed, Zimmer-Hart and Rescorla (1974) suggested a simple modification to permit the model to accommodate a lack of effect of extinction on inhibitory strength without compromising the importance of the prediction error concept itself.

There are some exceptions in the literature where predictions from the Rescorla-Wagner model are not completely met (e.g., Bouton and King, 1986; Delamater, 1996; Rescorla, 1996), but the main point here is that whenever a stimulus has been shown to display any of these various phenomena this does not constitute evidence against an unlearning interpretation. Indeed, an underappreciated fact is that the various recovery phenomena listed above often are themselves incomplete (e.g. Rauhut, Thomas, & Ayres, 2001; Bouton and King, 1986; Robbins, 1990), and this, in itself, is consistent with an unlearning interpretation (though perhaps it is consistent with other views as well - see below).

Theories that rely on error correction mechanisms to explain associative change in extinction make an additional prediction. This is that additional extinction of a CS exhibiting restoration of responding will undergo greater change than one not exhibiting restoration. This will occur because the CS exhibiting restoration again signals the US and the error produced by the non-occurrence of the US across additional extinction will be greater than that for the CS that continues to signal the non-occurrence of the US. Six ways are known to restore responding to an extinguished CS: the lapse of time (spontaneous recovery); a context shift between extinction and test (renewal); US alone presentations (reinstatement); reconditioning with a normally ineffective US intensity (e.g., Zheng, et al, 2013), and compounding an already extinguished CS with either a non-extinguished (Reberg, 1972) or

an extinguished (Hendry, 1982) CS. Additional extinction of a CS exhibiting most of these forms of restoration deepened response loss to that CS (Holmes & Westbrook, 2013; Leung and Westbrook, 2008; 2010; Rescorla, 2006). Moreover, the deepening produced by any one of these ways was evident by its subsequent resistance to any other. For example, a CS subjected to additional extinction while exhibiting spontaneous recovery was more resistant to subsequent reinstatement than an extinguished CS given additional extinction while not exhibiting recovery; a CS subjected to additional extinction while exhibiting reinstatement was more resistant to subsequent renewal than an extinguished CS given additional extinction while not exhibiting reinstatement.

Another key implication stemming from continued application of the Rescorla-Wagner model to extinction is that “extinction learning” should closely resemble many other cases of learning that depends upon negative prediction errors. For example, other cases that involve learning with “negative prediction errors” include Pavlovian conditioned inhibition, backward conditioned inhibition, explicitly unpaired conditioned inhibition, negative contingency learning, overexpectation effects, and others (e.g., LoLordo and Fairless, 1985; Rescorla, 1969). The interesting prediction from the Rescorla-Wagner model is that a common set of underlying neural processes should apply to all of these phenomena. To our knowledge a thorough going analysis of the neural mechanisms of “negative prediction error – based phenomena” has not yet taken place, with some exceptions (e.g., Takahashi, et al., 2009). Our understanding of the neural mechanisms of extinction can only be improved by inviting its comparison to other seemingly highly related phenomena.

A final critical advance made in the behavioral analysis of learning is well worth considering in the context of extinction. This advance concerns the nature of the prediction error rule that gives rise to new learning. In a series of elegant papers, Rescorla observed that depending upon their associative status CSs differ in the degree to which their associative values change on a given conditioning trial in which some prediction error is generated (Rescorla, 2000; 2001a; 2002b). More specifically, he suggested that conditioned excitors will be affected more than conditioned inhibitors on trials with negative prediction errors, but that the opposite will be true on trials with positive prediction errors. Consider an example. Suppose that an organism were to receive Tone – Pellet pairings on some conditioning trials and Tone + Light – No Pellet presentations on other conditioning trials. The Tone should become a conditioned excitor in this arrangement while the Light should become a conditioned inhibitor. Now, if additional Tone + Light – Pellet pairings were given, then on these trials a positive prediction error will result. The total prediction of pellet based upon the Tone and Light associative strengths going into these new trials will be effectively zero because the Tone’s prediction of a pellet would be cancelled by the Light’s prediction of its absence. Thus, when a pellet actually is paired with the Tone + Light compound stimulus, this produces an overall positive prediction error. What Rescorla asked was whether the Tone and Light stimuli would show equal increments in associative strength as a result of this positive prediction error. He observed that the Light’s associative strength incremented more than did the Tone’s as a result of these extra training trials. However, in conceptually similar studies in which negative prediction errors were generated, he observed that the excitor decremented more than the inhibitor (Rescorla, 2000).

The importance of these findings is that there appear to be multiple prediction error terms in the equation that describes learning. A “common” error term describes the overall difference between that which actually occurs and that which is predicted on the basis of all the stimuli present (Rescorla & Wagner, 1972). This is the usual way in which we now think about “prediction error.” However, these findings suggest that the degree to which an individual stimulus predicts the US on the trial – an “individual” error term – also seems to modulate the effectiveness of that overall prediction error (e.g., Bush & Mosteller, 1955). A simple

extension of the R-W model that captures the interaction between common and individual error terms is to assume that the common error term determines the amount of change that is available for distribution across the CSs and the individual error determines how much of that change is allocated to each CS (Le Pelley, 2004; Rescorla, 2000).

Leung and Westbrook (2008) invoked this explanation to account for the effects of extinguishing a CS exhibiting spontaneous recovery. As noted previously, they reported that extinction of such a CS resulted in a greater loss than extinction of a CS not exhibiting spontaneous recovery. They also showed that this deepening was regulated by both common and individual error terms. Specifically, they trained four CSs, A+, B+, C+ and D+, extinguished one (A−) at one point in time and second, (B−) at a later point. Shortly after extinction of B, rats were extinguished to two compounds; one (AC) composed of the CS exhibiting spontaneous recovery, the remotely extinguished A, and the excitator C; the other (BD) composed of the CS not exhibiting recovery, the recently extinguished B, and the excitator D. A subsequent test revealed less responding to C than D showing that the greater error produced by extinction of AC than BD had produced greater associative change to C relative to D. In subsequent experiments, they trained A+ and B+, extinguished A at one point in time, and B at a later point. Shortly after extinction of B, an AB compound was extinguished. If change was regulated solely by a common error term, then the amount of change to A and B would be the same, and subsequent responding would be equivalent. However, it was not: there was less responding to A than B, showing that the CS that entered compound extinction with the greater excitatory value (the remotely extinguished A) underwent a greater associative change. How the brain computes both common and individual error terms has yet to be addressed (but see below in Hong et al, 2011). More generally, how prediction errors, themselves, might come to influence learning within cortico-limbic extinction circuits is far from clear (but see McNally, Johansen, & Blair, 2011; Quirk & Mueller, 2008). We believe that major advances in our understanding of the neural basis of extinction will be made once a solution to these problems has been found.

Representational Content in Extinction

As a CS undergoes extinction training and reductions in responding are observed there are many levels at which one can explore this basic phenomenon. One important issue concerns the content of the initial learning because it is this learning that provides the basic substrate upon which extinction learning must act. There are at least two basic issues regarding the representational content of learning. The first is concerned with identifying the types of associations that the CS may have entered into with the various features of the US (e.g., its sensory, motivational, hedonic, temporal properties). The second issue concerns the nature of extinction learning itself. In other words, it has become popular to assume that extinction involves new learning, but we need to ask: how do we best characterize that new learning? The next two sections will discuss these issues in greater detail.

Sensory and Motivational Components of the US—One of the important characteristics of simple Pavlovian learning is that it is a multifaceted phenomenon (see also Delamater, 2012; Delamater & Oakeshott, 2007). Many different response systems can be engaged, and, undoubtedly, a variety of underlying learning and neural processes can be recruited as well. Konorski's (1967) characterization of Pavlovian learning was among the first to emphasize its multifaceted nature (see also Wagner and Brandon, 1989). As his starting point he accepted the earlier, more ethologically derived, distinction between preparatory and consummatory classes of behavior (Craig, 1918). The former class refers to behaviors that are rather diffuse and non-directed (e.g., general activity), and such behaviors have generally been assumed to be activated by releasing stimuli that are quite temporally distant from the ultimate goal object (e.g., the initial onset of physiological "hunger"

signals). Consummatory behaviors, in contrast, were thought to be highly specific and directed towards the goal object as the animal encounters it (e.g., clawing, clutching, and biting actions as in a predator-prey encounter). Such behaviors have been assumed to be activated by releasing stimuli most proximal to receipt of the goal object (e.g., the hissing sounds of an attacking snake just prior to their strike). Konorski went beyond earlier characterizations of this distinction by supposing that in Pavlovian conditioning different aspects of the US came to control these different classes of behavior, and were potential targets of conditioning: specifically, preparatory responses are controlled by the more general motivational (i.e., appetitive-aversive) features of the US, whereas consummatory responses are controlled by the US's highly specific sensory features (e.g., the specific tastes and textures of a food US or the localized painful qualities of a footshock US). A natural consequence of this way of thinking is that a CS could associate with either or both of these US features, and the underlying psychological and neural systems could be distinct from one another (e.g., Balleine & Killcross, 2006; Corbit & Balleine, 2005; 2011).

This basic distinction between sensory and motivational US components is widely accepted, but there are other properties of a US that enter into the determination of learned behavior. For instance, it is well known that USs are processed in terms of their hedonic qualities (e.g., Berridge & Kringelbach, 2013), their temporal characteristics (e.g., see Delamater and Holland, 2008), their spatial loci (e.g., Williams, Butler, & Overmier, 1990), and they are events that have specific response evoking capabilities. When CS and US are paired, then, separable associations are likely established between the predictive cue and each of these distinct features of the US. In short, a wide number of underlying neural systems may be engaged in any simple Pavlovian learning situation.

When one accepts this multifaceted aspect of Pavlovian learning, then it becomes clear that any answers to questions regarding psychological or neural mechanisms could very well turn out to be response-system dependent. For instance, current research clearly points to a distinction between learning about the motivational and sensory aspects of the US in appetitive tasks (e.g., Corbit and Balleine, 2005; 2011). Balleine and Killcross (2006) have suggested that the same distinctions also apply to aversive Pavlovian learning tasks; specifically, the basolateral complex of the amygdala is assumed to code for the sensory and the central nucleus for the motivational associations. It will become important for future research to establish parallels between learning about sensory and motivational US features in appetitive and aversive learning paradigms. Most current research in aversive paradigms employ a single US, typically foot shock, and designs that employ multiple aversive USs (e.g., Debiec, et al, 2010; Diaz-Mataix, et al, 2011) would be required to explore these issues because analogous multi-US designs are frequently employed in those studies of appetitive Pavlovian learning that give rise to such distinctions.

When examining the effects of extinction upon behavior and underlying psychological and neural mechanisms, these issues become extremely important. For instance, many authors have commented on the possibility that extinction generally involves new learning without being accompanied by much, if any, erasure of original learning (e.g., Bouton, et al., 2006; Delamater, 2004; Myers, Carlezon, & Davis, 2011; Quirk & Mueller, 2008). However, most extinction research has employed a single response system (e.g., freezing, eyeblink, or magazine approach) and any conclusions would necessarily be limited in scope to whatever specific learning processes may be assayed by that paradigm. It is fruitful to consider using paradigms that allow for the separation between learning about motivational and sensory-specific features of the US. This way, one can explore the possibility that extinction may have differential effects on different aspects of learning.

Delamater (1996) explored this possibility when he used a Pavlovian to instrumental transfer (PIT) task to determine if stimuli exerted selective control over instrumental responses with which they shared a reinforcing outcome. Selective PIT (see Kruse, Overmier, Konz, & Rokke, 1983) is widely understood to reflect learning about the highly specific sensory qualities of reward (e.g., see Balleine & Killcross, 2006). Whereas extinction conducted prior to the test session in the study reported by Delamater (1996) had no impact upon the ability of the CS to exert selective PIT, magazine approach responses to the extinguished CS were reduced relative to that seen to the non-extinguished CS. Delamater (1996) suggested that extinction reduced control by motivationally-based learning (the association between the CS and the motivational features of the US) without impacting the sensory-based learning (the association between the CS and the sensory features of the US). In other words, extinction could be viewed as having broken one association while leaving intact another. Very clearly, more experimental confirmation of this is required. The more general point, though, is that extinction may very well have distinct effects upon different aspects of learning. It becomes useful to know in any given situation which of these different aspects of learning is being studied. Experimental designs are often chosen that would not permit for a clear distinction between the many different aspects.

What is Learned in Extinction?—Konorski introduced another psychological distinction critical for furthering our study of the neural mechanisms of extinction. This concerns the nature of what is learned when a previously trained CS is presented without the US. Konorski discussed this in different ways. His earlier idea was that the CS develops an inhibitory association with the US as it undergoes extinction training (Konorski, 1948). In other words, whereas conditioning was thought to result in new excitatory connections forming between the neural representations of the CS and US, extinction would result in the formation of parallel inhibitory connections. However, Konorski (1967) later changed his view and suggested that during extinction the CS formed new excitatory connections with a “no US” memory.

This distinction between extinction involving new inhibitory learning with the US memory or excitatory learning with a “no US” memory has been difficult to address at the purely behavioral level. It seems possible, in principle, for the issue to be decided by a more complete understanding of the neural mechanisms of extinction. For instance, the finding that “extinction neurons” become excited by the CS as it undergoes extinction training (e.g., Herry, et al, 2008) would appear to provide support for Konorski’s later new excitatory learning with a “no US” memory view. However, it should be recognized that when new neural connections are developed between a cellular process that codes the CS and cells that then have inhibitory effects on subsequent processes thought to be important for response generation, such connections could also be interpreted in terms of Konorski’s earlier view that new inhibitory associations are learned during extinction training. In this case, those inhibitory connections happen to go through inhibitory interneurons. This idea is entirely consistent with Pare and colleagues’ finding that extinction results in the formation of enhanced connectivity between basolateral amygdala cells and intercalated cells that have inhibitory effects on cells within the medial central nucleus (Amano, Unal, & Pare, 2010). What is needed is a clear way of isolating CS from US neural representations, as well as a clear identification of where plasticity during extinction might be taking place (e.g., see Pare & Duvarci, 2012). Until this is better understood it will be difficult to fully understand the representational status of such “extinction” neurons, but the approach holds a lot of promise for future work in this regard.

Equally, it will become important for researchers to have a clear understanding of what might be meant by the idea that a CS enters into an excitatory association with a “no US” memory as a result of extinction training. This idea can be conceptualized in different ways,

and one challenge for future research will be distinguishing among them. In line with the distinction raised above between sensory and motivational features of the US, one can think of a “no US” memory in similar terms. When an organism undergoes extinction and is said to learn a CS – no US association, this “no US” representation can be thought of in sensory specific ways, as though the representation has specific referential content – to the fact that this particular US will not occur under these circumstances. Delamater, LoLordo, and Sosa (2003) reported that a CS paired in a backward manner with a particular US acquired the capacity to selectively inhibit instrumental responses that were previously reinforced with that specific US, and so the notion as applied to extinction may have some merit.

An alternative possibility is that the CS forms an excitatory association with another general motivational state that is antagonistic to the state originally conditioned to the CS during the training phase. For instance, if during tone-shock conditioning the tone CS comes to associate with an aversive motivational state, e.g., fear, then during extinction the CS might develop a new excitatory association with an appetitive motivational state that opposes fear (e.g., Denny, 1971; Dickinson & Dearing, 1979; Mowrer, 1960). Similarly, a CS that signals food comes to evoke an appetitive motivational state, but then during extinction the CS might develop a new excitatory association with an opposing aversive state, e.g., frustration (Rescorla, 2001c). In this way, the new learning during extinction diminishes conditioned responding because the net level of motivation supporting the CR in the first place has been reduced by these antagonistic interactions between the two opposing states. Notice that the CS in this case is not referential because it does not convey information regarding the new associative relation. Rather, the fear-conditioned stimulus, for instance, merely comes to activate less defensive responding across extinction because of the balancing act between these two opposing motivational states. How we conceptualize the nature of this new excitatory learning in extinction will have important implications for our understanding of the functional significance of underlying neural mechanisms.

An additional answer to the question of what is learned in extinction comes largely from the work of Bouton (e.g., 1991, 1994, 2004; see also this special issue). In this work, Bouton argues that the extinction context acquires the capacity to function as an “occasion setting” stimulus that is required to retrieve the memory of extinction. This “memory of extinction” can be thought of in different ways, but the most common way has been to assume the context functions as an AND gate enabling the CS to inhibit US activation through its inhibitory association with the US representation (Bouton and Nelson, 1994). When the CS is tested outside of the extinction context, as in renewal designs, then the necessary supports for fully engaging this inhibitory CS-US link are not in place and so the CS will only come to excite the US representation. It is important to note that this occasion setting mechanism is outside of the scope of traditional associative theories, such as the Rescorla-Wagner model, because those theories only admit that learning consists of either excitatory or inhibitory associations between two events but not modulatory relations (see also Holland, 1985; Rescorla, 1985).

A complete discussion of this mechanism of extinction is outside the scope of the present paper. Here we point out merely that the best evidence to support this distinction between simple associative and modulatory functions of cues and contexts is provided by studies that attempt to show (a) that modulatory functions persist even after treatments are applied that undermine any potential associative control, and (b) that such modulatory cues only work on select classes of stimuli. For instance, Holland (1985) demonstrated that a negative occasion setting stimulus did not lose this negative modulatory function even after the stimulus was converted into an excitatory stimulus through separate pairings of the modulatory CS with the US prior to the test for modulation. Moreover, Bouton and King (1986) demonstrated that reinstating US presentations increased responding to a CS tested in the reinstatement

context, but this only occurred when the CS had been trained and then extinguished. Other CSs given either a latent inhibition or partial reinforcement treatment were not affected by this reinstatement treatment (see also Harris et al, 2000). This result is difficult to reconcile with a simple associative function played by the context because if that were true then all stimuli should have been affected, not just extinguished ones.

We may also note that perhaps the best experimental designs used to assess control by this occasion setting function in extinction are those that explicitly attempt to equate the simple associative histories of the contexts in a renewal experiment. For instance, in an ABA renewal design where different CSs are trained in different contexts (Context 1: CS1-US, Context 2: CS2-US) but then extinguished in the other cue's training context (Context 1: CS2-, Context 2: CS1-) both contexts receive equivalent training. Test responding to the two CSs is greater when the stimuli are tested in their training than their extinction context. Such a design is more likely to recruit occasion setting like mechanisms because an appeal to control by simple associative functions by context does not easily work here (but see previous discussion of such designs generating a unique cue that attracts the inhibition). There are several examples of the use of this sort of design in renewal experiments (e.g., Campese & Delamater, 2013; Delamater, Campese, & Westbrook, 2006; Harris, et al, 2000; Maren & Hobin, 2007); however, it is far more common for researchers to use renewal designs that are more amenable to the kinds of analyses advanced above in terms of the mechanisms discussed by the Rescorla-Wagner model. In fact, most studies examining the neural substrates of fear extinction or pharmacological manipulations that impair or facilitate fear extinction have used a design in which a CS is trained in A and then extinguished and tested in B. This leaves open the possibility for manipulations that facilitate extinction, for example, to do so by affecting the development of inhibition to context rather than what is assumed, namely, the development of inhibition to the CS or occasion setting to the context. In a recent study by Baker, McNally, and Richardson (2012), such a design was used and while d-cycloserine was shown to facilitate the extinction of conditioned freezing to a CS, this study went further than others by testing whether the extinction context had developed enhanced inhibitory learning in d-cycloserine-treated animals. The results, unfortunately, are somewhat equivocal on this point because although the extinction context in drug-treated animals was no slower than in controls at acquiring fear of its own, a more traditional summation test for conditioned inhibition was not performed.

The Role of Attention in Extinction

Another important psychological issue for our understanding of extinction concerns the role of attention, and, possibly, habituation (e.g., see McSweeney & Swindell, 2002). Some conditioning theories have assumed that the CS undergoes changes in its processing over the course of acquisition and extinction phases. For instance, Pavlov (1927) originally suggested that new inhibitory learning accrues during extinction, but he assumed that the inhibition was directed to cortical cells controlled by the CS. One interpretation of this suggestion is that extinction is accompanied by decreases in CS processing, or, more simply, decreases in attention to the CS (Robbins, 1990). Mackintosh's theory of learning (Mackintosh, 1975) assumes that attention to a CS increases to the extent that it signals the US more strongly than other cues present, and that it decreases when the CS is no better than other cues at signaling the US. The latter condition will apply to extinction as the CS loses its associative strength during extinction and so this theory would suppose that attention to the CS should decline as well.

Other theories of conditioning also suggest that attention to CSs can fluctuate during acquisition and extinction phases of an experiment (Pearce & Hall, 1980; see also Larrauri & Schmajuk, 2008). However, according to these views attention declines during acquisition training as the CS comes to fully predict the US (or as the overall "novelty" in the situation

declines), increases somewhat during the initial trials of extinction because of a noticeable change in the contingency, but then decrease again during extinction as the CS comes to fully predict the absence of the US. An implication of such theories is that associative change ceases once attention to the CS has declined. Hence, manipulations that maintain attention across extinction or that restore attention should lead to additional associative change, enhancing response loss. Still other, more recent, models of learning exist and these also find a prominent role for attention (e.g., Le Pelley, 2004; Pearce & Mackintosh, 2010). Such models attempt to integrate, in one way or another, the sorts of mechanisms discussed previously by Mackintosh (1975) and Pearce and Hall (1980). The main point is that the study of associative learning has long maintained that attention to the CS is one of the important variables that govern learning. Indeed, it would be surprising if the study of the neural mechanisms of extinction did not reveal that extinction is accompanied by decreases in attention to the CS as well as whatever else might occur during extinction. Further work will be needed to more fully explore this possibility as well.

Psychological Mechanisms and their Neural Implementation

As noted above there has been an explosion of interest in the study of extinction over the last 15 years or so and much of this increased interest is related to progress in identifying the neural mechanisms of conditioning and extinction, largely in fear conditioning paradigms and, more recently, in appetitive learning paradigms as well (where natural and/or drug rewards are employed). While excellent reviews of these literatures already exist (e.g., Herry, et al, 2010; Marchant, Millan, & McNally, 2012; Maren & Quirk, 2004; Pare & Duvarci, 2012; Peters, Kalivas, & Quirk, 2009; Quirk & Mueller, 2008), the current focus will be on examining potential points of convergence between the psychological mechanisms of extinction noted above and what we currently know of the underlying neurobiology of extinction. Thus, our focus will necessarily be limited in scope, and we hope to highlight for future research several areas that may be especially worthwhile in reaching a more integrative understanding of extinction than currently exists. Before addressing these issues, however, it will be useful to briefly consider a model of extinction learning that has emerged over the last decade.

A Neural Systems Model of Extinction Learning in Aversive and Appetitive Paradigms

A large number of studies have identified the ventromedial prefrontal cortex, especially the infralimbic region (IL), the hippocampus, and the amygdala as major sites where plasticity takes place during extinction in fear conditioning. In addition, primary auditory cortex and the thalamus are two other structures shown to participate in extinction learning in fear conditioning preparations, as lesions to these structures retard the development of extinction (e.g., Song, et al, 2010). In instrumental appetitive tasks, in addition to IL and amygdala, other structures thought to be important for extinction learning include the nucleus accumbens shell (NAcSh), the mediodorsal hypothalamus (MDH), and the paraventricular thalamus (PVT) (see Marchant, Millan, & McNally, 2012; Peters, Kalivas, & Quirk, 2009).

Substantial evidence points to a critical role of the IL in long-term extinction of conditioned fear responses (freezing), although some findings suggest that this structure may not be the only critical site (e.g., Chang & Maren, 2010; Garcia, Chang, & Maren, 2006). This evidence includes the poor recall of extinction (i.e., recovery of responding) observed as a result of pre-training IL lesion and local inactivation performed either shortly after extinction training or prior to a retention test (Laurent & Westbrook, 2009; Morgan, Romanski, & LeDoux, 1993; Sierra-Mercado, Padilla-Coreano, & Quirk, 2011). Consistent with these findings, stimulation of IL neurons reduces recovery of responding following extinction (Vidal-Gonzalez, Vidal-Gonzalez, Rauch, & Quirk, 2006), and spontaneous firing during extinction training is correlated with longer lasting extinction retention (Santini,

Quirk, & Porter, 2008). However, the role of the IL during extinction training remains unclear as one study showed that local inactivation impairs fear inhibition late in extinction training (Sierra-Mercado, et al, 2011) while another study found no effect (Laurent & Westbrook, 2009). This discrepancy may be due to differences in training duration or stimulus modality (auditory vs. contextual). Additional evidence that the IL constitutes a site of extinction learning is that inactivation prior to re-extinction disrupts re-learning of extinction. Rescorla (2001b; see also Leung & Westbrook, 2008) showed that retraining an extinguished CS did not erase the learning produced by extinction: such a CS continued to exhibit the signature characteristics of extinction, notably spontaneous recovery. Re-extinguishing a retrained CS presumably recruits the original extinction trace whose site is the IL, as indicated by the impairment in re-extinction when preceded by inactivation of the IL.

These data point to different mechanisms accounting for within session extinction and its retention. Inhibitory neural processes within the amygdala have been shown to play a key role in explaining within session extinction and also point to ways in which longer term extinction retention might be accomplished. Specifically, the discovery of “extinction neurons,” i.e., neurons whose firing is correlated with extinction, within the basal amygdala (Herry, et al, 2008) suggests that new cellular processes within the amygdala are recruited during extinction. Consistent with this are findings suggesting changes in connectivity within the amygdala as a result of extinction. In particular, Amano, Unal, and Pare (2010) observed extinction to produce greater inhibitory post-synaptic potentials (IPSPs) on cells within the central medial nucleus of the amygdala (CNm) induced by stimulation of basolateral amygdala cells (BLA). Furthermore, these authors went on to show that extinction training resulted in enhanced connectivity between BLA cells and inhibitory CNm-projecting intercalated cells (ITC), an effect that was mediated by input from the IL (Amano, et al, 2010). These results together with those mentioned above support the view that significant microcircuitry changes take place within the amygdala during extinction, and that IL plays a prominent role in supporting some of those changes. It is noteworthy, however, that recent work has established that IL projections to ITC are rather light compared to the more dense projections to neighboring regions (Pinard, Mascagni, & McDonald, 2012). How such findings can be reconciled with existing work remains to be resolved.

The hippocampus has strong connections with the IL, as well as the amygdala, and this structure has been implicated in some context-dependent extinction phenomena, e.g., renewal. Maren and his colleagues have documented a role for the dorsal hippocampus in “ABA” and “ABC” fear renewal (e.g., Corcoran and Maren, 2001; 2004; Ji and Maren, 2005; 2007; Maren and Hobin, 2007). The generality of these effects remain to be determined. Using a Pavlovian appetitive conditioning task (conditioned magazine approach) we have failed to find a role for the dorsal hippocampus in either ABA or ABC forms of renewal (Campese and Delamater, 2013), although dorsal hippocampal inactivation at the time of testing did undermine spontaneous recovery (Campese, unpublished dissertation). It will be important to determine the nature of the discrepancy between aversive and appetitive tasks, but some potentially important differences concern the total number of conditioning trials, the total exposure to contexts, and the fixed or variable nature of intertrial intervals in the two situations.

In another appetitive paradigm, drug self-administration, the IL has also been shown to interact with the NAcSh in the mediation of extinction learning. For instance, Peters, LaLumiere, and Kalivas (2008) showed that IL inactivation restored extinguished cocaine-reinforced instrumental responses. Moreover, unilateral IL inactivation together with either ipsilateral or contralateral inactivation of the NAcSh also disrupted the expression of

extinction in this task, suggesting that the two structures communicate in the retrieval of extinction memories. Using an alcohol reinforcement task, Marchant et al. (2009) found with cFOS and retrograde tracing techniques that different subpopulations of lateral hypothalamus-projecting NAcSh neurons are recruited during ABB and ABA renewal tasks. However, using a heroin self-administration paradigm, Bossert et al. (2012) demonstrated that unilaterally inactivating the ventromedial prefrontal cortex (vmPFC) while also administering a dopamine D1 receptor antagonist directly into the NAcSh diminished renewal relative to control conditions, a result that was interpreted to mean that prefrontal-accumbens shell interactions may actually play opposing roles in tasks involving cocaine or alcohol and heroin reward (see also Ghazizadeh, et al., 2012). Retrograde tracing results pointed to the presence of more NAcSh-projecting cells in the vmPFC during ABA than ABB tests in this task. The results from these studies are difficult to compare for many reasons, and before any firm conclusions can be drawn from some of the divergent sets of results it will be important to perform experiments under more similar conditions than has currently been done. To complicate the picture further, Marchant et al (2009) have also provided evidence from tracing studies suggesting the importance of a circuit involving the IL, the mediodorsal hypothalamus (MDH), and the paraventricular thalamus (PVT) in renewal of extinguished alcohol seeking. The involvement of multiple structures in addition to the IL (NAcSh, LH, MDH, PVT) in these appetitive learning paradigms points to a more complex extinction network than is currently thought to modulate extinction in aversive learning (see also Peters, Kalivas, & Quirk, 2009).

Attention and Extinction Circuits

As noted above, one important psychological construct used to explain extinction phenomena at the behavioral level is the notion that attention to the CS declines over the course of extinction training. This psychological mechanism could help explain why excitatory learning may not be completely undermined by extinction – if attention to the CS declines as the subjects learn about its new relation to the US, it should be difficult for the stimulus to lose its conditioned strength. In spite of the apparent simplicity of this idea there has not been a great deal of work at either the behavioral or neural systems levels of analysis that directly bear on this issue. However, recent studies do provide some intriguing clues.

Early work established the auditory cortex as a site of plasticity during fear conditioning. For example, Quirk, Armony, and LeDoux (1997) found cells within auditory association cortex (Te3/Te1v) that displayed conditioning-dependent increases in firing rates. Although such cells showed increased firing with a slower onset latency than those recorded from the lateral amygdala, and only LA cells showed decreases in firing rates during extinction, the results point to auditory cortex as one site of plasticity within fear conditioning. Furthermore, as noted above, it has been reported that pre-training lesions of the auditory cortex retard extinction in an auditory fear conditioning task (Song, et al., 2010). The results from this latter study may especially point to a role for attention because in that study while a 10-s auditory CS was paired with a shock US during the training phase, this CS was presented for a continuous 8 min during extinction testing. Such a testing procedure is likely to involve decrements in attention by recruiting habituation-like processes.

In another line of research that points to the potential involvement of attentional processes during conditioning and extinction, Weinberger and his colleagues (e.g., Edeline & Weinberger, 1992; Weinberger, 2007a; 2007b) have studied receptive field plasticity within the auditory cortex during fear, and, more recently, appetitive tasks in guinea pigs and rats. For instance, during both fear (Edeline & Weinberger, 1992; 1993) and instrumental appetitively-reinforced tasks (Bieszczad & Weinberger, 2010, 2012), these authors have reported that cells within primary auditory cortex, A1, change their firing sensitivities to different auditory frequencies favoring the specific stimulus trained as S+. In other words, as

a result of conditioning there is more space within A1 devoted to representing the training stimulus – a kind of representational expansion. An especially intriguing finding is Bieszczad and Weinberger's recent discovery that just as conditioning results in a representational expansion within A1, extinction results in a representational contraction and a return to normal pre-conditioning levels within A1 (Bieszczad & Weinberger, 2012). While this result is merely suggestive, it is consistent with what might be seen as a minimal requirement for an attentional theory of conditioning and extinction, namely, that relatively direct measures of stimulus processing (representational space in primary sensory areas) show fluctuations during conditioning and extinction treatments. If such an idea is to gain more explanatory power then it will be worth investigating the consequences of such representational contraction for further downstream structures, as well as the dynamics of this process over the course of training (Weinberger, 2007a). Nevertheless, the results are clearly consistent with what would be expected if extinction were to be accompanied by decreases in attention paid to the CS.

One other line of inquiry in this regard is worth noting. Sharpe and Killcross (2012) recently demonstrated that prelimbic lesions of the prefrontal cortex resulted in a failure to down regulate attention paid to a specific stimulus in appetitive Pavlovian blocking experiments. While the prelimbic cortex is generally thought of as promoting conditioned responding and opposing extinction learning (e.g., Peters, et al, 2009), given these findings it may be worth considering its role in attention during extinction.

Erasure, New Learning, & Prediction Error Coding in the Brain

The prevailing view of extinction is that it involves new learning, as various recovery phenomena (spontaneous recovery, renewal, reinstatement, and rapid reacquisition) suggest that extinction does not entirely eliminate learning. However, as we hope is clear from our earlier discussion, each of these phenomena is to a large extent consistent with predictions derived from the Rescorla-Wagner model – a model that assumes extinction to produce a partial reduction in associative strength, i.e., partial erasure, together with new learning. Thus, we think it is quite reasonable to ask not if extinction completely erases old learning, but, rather, whether it partially erases that learning (see also Mauk & Ohya, 2004). While some behavioral studies suggest that extinction may have no effect on prior learning (Delamater, 1996; Rescorla, 1996), recovery phenomena are usually, if not always, incomplete (e.g., see Bouton & King, 1983; Rauhut, Thomas, & Ayres, 2001). This suggests that extinction may at least partially erase prior learning.

More molecular and neural systems level studies provide support for this possibility. First, Richardson and his colleagues working with young rats have provided evidence to suggest that fear conditioning may, indeed, be eliminated by extinction, and that, at the very least, the system mediating extinction in very young rats differs from those employed in the adult. One of the key findings is that 17 (but not 24) day old rats fail to display renewal and reinstatement phenomena (Kim & Richardson, 2007a,b). Further, Kim and Richardson (2008) demonstrated that extinction and re-extinction (i.e., where the extinguished CS is retrained and again extinguished) were both amygdala dependent in 17-day old rats, but 24-day old rats displayed amygdala-dependent extinction and amygdala-independent re-extinction. Thus, it appears that extinction processes in developing rats differ from those in young adult rats. Whether this reflects the possibility that extinction erases the original fear memory or more completely masks it is not obvious from these findings.

Evidence that is consistent with the erasure view, though, comes from Kim & Richardson's additional finding (Kim & Richardson, 2010) that when rats were extinguished while young and then re-extinguished at an older age, re-extinction was shown to be NMDAR-dependent. Control rats that were given both extinction and re-extinction at an older age displayed a

normal transition from NMDAr-dependent extinction to independent re-extinction. Thus, it appears as though when extinction occurs in the young rat, it may reset the memory in such a way to make it more vulnerable to NMDAr manipulations (see also Graham & Richardson, 2011). However, an alternative interpretation is that retraining erased the learning produced by extinction in the young rats, rather than extinction erasing the original learning. As noted above, an extinguished CS that has been re-trained continues to exhibit spontaneous recovery in mature animals (e.g., Rescorla, 2001b), showing that retraining does not erase what had been learned in extinction. It remains to be determined if this is also true in young rats. If it is, then the Kim and Richardson (2010) results could be more definitively ascribed to the erasure of the original learning. While these data are suggestive of an erasure view of extinction, more molecular studies would be required to identify the specific neural mechanisms involved in extinction and re-extinction processes (e.g., see Laurent, Marchand, & Westbrook, 2008; Laurent & Westbrook, 2009).

More direct evidence to support a true erasure account of extinction comes from two-photon microscopy and in vitro slice recording studies. Lai et al (2012) recently examined fear conditioning and extinction in mice using two-photon microscopy to examine dendritic spine remodeling in dorsal medial frontal association cortex (dmFrA). These authors observed significantly enhanced spine elimination in individual neurons in fear-conditioned mice (but no increase in new spines on individual neurons). This effect was behaviorally meaningful as the amount of spine elimination was strongly correlated with conditioned freezing in these mice (i.e., $r = 0.77$ in a test conducted two days following fear conditioning, and $r = 0.90$ in a test conducted nine days following conditioning). However, after extinction training there was a significant *increase* in the formation of new spines on those very same neurons where spines had been eliminated as a result of fear conditioning. Furthermore, a large proportion of the new spinal growth observed on these neurons occurred within two micrometers of those that had been eliminated as a result of conditioning. This result impressively demonstrates that extinction has the effect of restoring the neuron back to its original state, a result that is directly consistent with the erasure notion of extinction discussed more extensively above. It seems possible, however, that if conditioning results in enhanced attention to a CS, extinction might result in reduced attention to that CS and this could be revealed as a restoration of the neuron to its original state. This may especially be true for neurons studied in primary sensory areas or, more generally, at loci other than where CS and US convergence is assumed to take place. However, a somewhat stronger case for true associative erasure (reduced associative strength) in extinction can be made from fear conditioning studies examining conditioning and extinction-induced synaptic changes in the amygdala.

In another series of studies, Kim et al. (2007) studied depotentiation of thalamic-lateral amygdala synapses in tissues taken from animals given fear conditioning or fear conditioning and extinction treatments. These authors found increased synaptic efficacy in tissues taken from fear-conditioned rats, but, importantly, this increased efficacy was not seen in tissues taken from rats given fear conditioning followed by extinction. Furthermore, paired-pulse low frequency stimulation (which induces long-term depression (LTD)) applied to the tissue extracted from fear conditioned rats resulted in a depotentiation effect, but this was not observed in tissues taken from naïve rats or those that had undergone fear conditioning followed by extinction. Finally, the depotentiation effect depended upon AMPA receptor endocytosis as a peptide that blocked this endocytosis reduced the rate at which freezing extinguished and eliminated the depotentiation effect. These results suggest that AMPA receptor reabsorption by the cell is one mechanism of extinction and depotentiation. To support this idea the authors additionally found that AMPA receptor expression in the lateral amygdala (GluR1 and GluR2) was increased by fear conditioning,

but extinction eliminated this effect suggesting a locus at which extinction erases the learning produced by conditioning.

The results from this study fairly directly demonstrate that increased synaptic efficacy brought about by fear conditioning can be reversed by extinction, a result that is consistent with an associative erasure interpretation of extinction. Further, because such findings were observed to take place in the lateral amygdala, the primary site for plasticity in the fear conditioning circuit, the data strongly implicate the kind of associative process envisaged by the Rescorla-Wagner model, namely, that conditioning results in increments in associative strength while extinction results in decrements.

A later study by Hong et al (2011) added to this story. These authors extracted tissue from LA in rats given either fear conditioning, fear conditioning and extinction, conditioning-extinction-reconditioning, unpaired training, or no prior training. The LA tissue was examined for LTP or LTD (i.e., depotentiation induced by administration of the type I mGluR agonist, DHPG), and it was found that LTP occurred only in tissue from naïve, unpaired, or conditioned and then extinguished rats. Depotentiation, on the other hand, only occurred in conditioned or reconditioned rats. Thus, synapses that have already reached their “saturation” point are not subject to either further increases or decreases in plasticity. This interesting conclusion maps fairly well onto the ideas noted above (1) that extinction can result in true associative decrements, and (2) that associative changes brought about by prediction error computations are applied to stimuli in proportion to the degree to which those individual stimuli signal that prediction error. In other words, these studies provide evidence at the synaptic level to support Rescorla’s claims (2000; 2001a; 2002b) that individual stimuli that signal an event most discrepant from what actually occurs on a training trial benefit the most from the prediction error (positive or negative) that may occur on that training trial. It will be important to determine if such a neural realization of this basic idea will be able to account for some of the phenomena noted above regarding the effectiveness of prediction errors on learning to individual stimuli.

While all the studies reviewed in this section converge on the idea that extinction can be accompanied by some loss of prior learning, most behavioral studies in the literature suggest that extinction does not completely undermine that learning. Indeed, we have discussed above some theoretical reasons why that might generally be true. In one additional area of research, however, investigators have recently explored the hypothesis that certain memory reactivation manipulations may more permanently cause a profound associative loss. This is particularly important because according to the sorts of psychological theorizing discussed in this paper such manipulations should not have such an impact.

Monfils et al (2009; see Schiller et al, 2010 for a similar effect in humans) reported that a single presentation of a previously fear-conditioned 20-s auditory CS presented 10 or 60 min, but not 6 hr, prior to a further 18 extinction trials (separated by 3 min inter-trial intervals) eliminated spontaneous recovery of the fear CR, ABA renewal, reinstatement, and rapid reacquisition. Indeed, reactivated rats displayed slower reacquisition relative to a naïve control group. The authors suggested that a single tone presentation prior to extinction retrieves the memory and by doing so places it in a labile state of reconsolidation such that the memory of acquisition can, essentially, be rewritten by the fear extinction memory should further extinction trials occur within this “reconsolidation window” that presumably lasts somewhere between 10 min and 6 hr. Additional data in this report suggested that the dephosphorylation of GluR1 receptors may mediate this retrieval-extinction effect.

One problem with this interpretation is that over the course of the normal extinction series (18 CSs presented with a 3-min inter-trial interval), if the reconsolidation window becomes

activated by 10 min then extinction trials 5 – 18 should have occurred within the reconsolidation window. Thus, it is not clear why such an effect should not have also been seen in the rats given normal extinction training.

An intriguing replication of the basic retrieval-extinction effect was recently reported by Xue, et al (2012) who found that an extinction procedure given 10 or 60 min, but not 6 hr, prior to a longer lasting extinction period (45 min) resulted in reduced reinstatement and spontaneous recovery of a conditioned place preference based on cocaine or heroin reward. These authors also replicated the basic effect using a drug self-administration paradigm as well as a cue-induced craving paradigm in human heroin addicts. In their rat studies, Xue et al (2012) also went on to show that normal extinction resulted in increased expression of PKMzeta in the IL and decreased expression in the BLA. Moreover, the retrieval-extinction task further increased PKMzeta expression in the IL and decreased it in the BLA. This protein has been implicated in several forms of long-term memory stability and synaptic plasticity (Sactor, 2011; Shema, Sacktor, & Dudai, 2007; Serrano et al, 2009; Kwapis, Jarome, Lonergan, & Helmstetter, 2009), although even more recent studies question the functional significance of this protein for learning and memory processes (Volk, et al, 2013).

While these intriguing findings suggest that it may be possible to use a purely behavioral technique to induce long-lasting extinction, a number of other studies have found either no effect of the “retrieval-extinction” manipulation or the opposite effect. For instance, Chan et al (2010) performed six experiments in which a single CS alone presentation occurred 90 min prior to a series of 18 additional extinction trials (with a 3 min ITI) in Group Retrieval. Group No Retrieval rats received equal exposure to the chamber 90 min prior to an extinction series of 19 CS alone presentations. These authors found more profound ABA renewal in rats given the retrieval treatment. Furthermore, they observed that this difference disappeared when the retrieval trial occurred in a different context from where the remaining extinction trials occurred.

Using mice, Stafford, et al (2013) found the retrieval treatment to result in attenuated conditioned freezing in tests conducted 1 or 14 days later, but only when this occurred 24 hr prior to extinction (and not immediately, 1 hr, or 4 hr before). MacPherson, et al (2013) also failed to find an effect of a retrieval treatment in two different strains of mice. Millan, Milligan-Saville, and McNally (2013) using an alcohol self-administration design replicated the effect reported by Xue, et al (2012), however, these authors also found that it did not matter whether the brief 10 min retrieval trial preceded or followed the 50 min extinction trial (see also Baker, McNally, & Richardson, 2013). In both cases reduced ABA renewal was observed relative to a no retrieval control (that received a single 60 min extinction session). Furthermore, these authors found that while ABA renewal was attenuated by the treatment, the retrieval manipulation resulted in faster reacquisition suggesting that the memory erasure (or rewrite) interpretation does not apply here. Similarly, using a human learning paradigm, Kindt and Soeter (2012) failed to replicate the basic retrieval-extinction effect initially reported by Monfils et al (2009) and Schiller et al (2010).

Thus, although the treatment appears to be effective in some circumstances at resulting in durable extinction learning with accompanying changes in underlying neural mechanisms, problems with replication as well as an inadequate theoretical understanding of the phenomenon speak to the need for its further investigation. A further complication is that a closely related procedure, conducting extinction immediately after conditioning, might also be expected to place the acquisition memory in a labile state. Studies that have examined this manipulation have also shown inconsistent results, reporting either enhanced (e.g., Myers, Ressler, & Davis, 2006) or reduced retention of extinction (e.g., Chang & Maren, 2009; Maren & Chang, 2006).

Summary and Conclusions

We have covered much ground in our review of experimental extinction. Clearly, much progress has been made, but, at the same time, many important questions remain. We hope to have convinced the reader that the issue of whether extinction entails “erasure” or “new learning” is a gross oversimplification of the psychological problem in the study of extinction. Psychological theories assume there to be some erasure and some new learning. The nature of that new learning and the nature of the mechanisms that drive that new learning need to be more fully understood, especially at the neural systems level. Theoretical work at the psychological level has pointed to the involvement of two distinct types of prediction errors (based on common or individual error terms) in determining both acquisition and extinction learning. We are beginning to see evidence at the neural level for partial erasure of original learning, as well as the importance of common and individual errors in governing changes in neural plasticity. Additional work will be required, however, before this mapping can be more fully realized. We are only just beginning to understand how prediction error circuits contribute to learning and much progress needs to be made in this area, especially as it involves the use of multiple methods for generating negative prediction error contributions to associative changes.

The role of attention in extinction, surprisingly, has not been extensively explored. When investigators use extinction methods that involve presenting a short duration CS in training for extended periods of time (e.g., 8 min) in test, changes in attention are very likely to be engaged. Current work looking at changes in neural representations in primary sensory areas brought about by conditioning and extinction procedures seem to converge on psychological ideas regarding the role of attention in learning. This would appear to be a good place to see interesting connections between psychological and neural levels of analysis, but studies along these lines are only just beginning.

Perhaps one of the most underdeveloped areas of research reviewed here concerns issues in the content of extinction learning. There are many avenues on which this key issue can be addressed. For one thing, the use of a single response system to study extinction will likely limit the conclusions we may draw because of the multi-faceted nature of simple Pavlovian learning, and, presumably, its extinction. Moreover, how should we best characterize extinction learning? Is it learning a new CS – No US memory or learning an inhibitory CS – US memory or even a hierarchical negative occasion setting associative structure of some sort? If we think of learning in either of these ways, how do we best capture the essence of that US representation? Should we think of it in primarily sensory terms, emotional terms, or response processes, or some combination that may vary with the protocols used? In principle, as we uncover more about the underlying neural mechanisms of extinction we should be able to come to grips with some of these questions, but only if we have behavioral tools that will effectively capture these distinctions.

Overall, there is much to be excited about concerning the continuing study of extinction. The deployment of more sophisticated tools together with an increasing recognition of how to effectively use those tools in interesting behavioral tasks has led and, we suspect, will continue to lead to a more complete understanding of what has become an aggressively studied and historically important topic. Moreover, this more complete understanding of extinction will be critical for developing more effective treatments of anxiety and appetitive disorders.

Acknowledgments

The paper was largely written while ARD was a Visiting Professorial Fellow at the University of New South Wales, and generously supported through a Visiting Research Fellowship awarded by the University of New South Wales and by a National Institute on Drug Abuse (034995) grant awarded to ARD.

References

- Amano T, Unal CT, Paré D. Synaptic correlates of fear extinction in the amygdala. *Nature Neuroscience*. 2010; 13(4):489–494.
- Baker KD, McNally GP, Richardson R. D-cycloserine does not facilitate fear extinction by reducing conditioned stimulus processing or promoting conditioned inhibition to contextual cues. *Learning & Memory*. 2012; 19:461–469. [PubMed: 22984284]
- Baker KD, McNally GP, Richardson R. Memory retrieval before or after extinction reduces recovery of fear in adolescent rats. *Learning & Memory*. 2013; 20:467–473. [PubMed: 23950194]
- Balleine BW, Killcross S. Parallel incentive processing: An integrated view of amygdala function. *Trends in Neurosciences*. 2006; 29(5):272–279. [PubMed: 16545468]
- Beiszcza KM, Weinberger NM. Remodeling the cortex in memory: Increased use of a learning strategy increases the representational area of relevant acoustic cues. *Neurobiology of Learning and Memory*. 2010; 94(2):127–144. [PubMed: 20434577]
- Beiszcza KM, Weinberger NM. Extinction reveals that primary sensory cortex predicts reinforcement outcome. *The European Journal of Neuroscience*. 2012; 35(4):598–613. [PubMed: 22304434]
- Berridge KC, Kringelbach ML. Neuroscience of affect: Brain mechanisms of pleasure and displeasure. *Current Opinion in Neurobiology*. 2013; 23(3):294–303. [PubMed: 23375169]
- Bossert JM, Stern AL, Theberge FR, Marchant NJ, Wang HL, Morales M, Shaham Y. Role of projections from ventral medial prefrontal cortex to nucleus accumbens shell in context-induced reinstatement of heroin seeking. *The Journal of Neuroscience*. 2012; 32(14):4982–4991. [PubMed: 22492053]
- Bouton ME. Context and retrieval in extinction and in other examples of interference in simple associative learning. In: Dachowski, L.; Flaherty, CF., editors. *Current topics in animal learning: Brain, emotion, and cognition*. Hillsdale, NJ: Erlbaum; 1991. p. 25–53.
- Bouton ME. Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes*. 1994; 20(3):219–231.
- Bouton ME. Context and behavioral processes in extinction. *Learning & Memory*. 2004; 11(5):485–494. [PubMed: 15466298]
- Bouton ME, King DA. Effect of context on performance to conditioned stimuli with mixed histories of reinforcement and nonreinforcement. *Journal of Experimental Psychology: Animal Behavior Processes*. 1986; 12(1):4–15.
- Bouton ME, Nelson JB. Context-specificity of target versus feature inhibition in a feature-negative discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*. 1994; 20(1):51–65. [PubMed: 8308493]
- Bouton ME, Westbrook RF, Corcoran KA, Maren S. Contextual and temporal modulation of extinction: Behavioral and biological mechanisms. *Biological Psychiatry*. 2006; 60(4):352–360. [PubMed: 16616731]
- Bush, RR.; Mosteller, F. *Stochastic models for learning*. New York: Wiley; 1955.
- Campese, V. The role of the dorsal hippocampus in the contextual control of appetitive responding (unpublished dissertation). City University of New York; New York: 2011.
- Campese V, Delamater AR. ABA and ABC renewal of conditioned magazine approach are not impaired by dorsal hippocampus inactivation or lesions. *Behavioural Brain Research*. 2013; 248:62–73. [PubMed: 23583520]
- Chang, C-h; Maren, S. Strain difference in the effect of infralimbic cortex lesions on fear extinction in rats. *Behavioral Neuroscience*. 2010; 124:391–397. [PubMed: 20528083]

- Corbit LH, Balleine BW. Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian-instrumental transfer. *The Journal of Neuroscience*. 2005; 25(4):962–970. [PubMed: 15673677]
- Corbit LH, Balleine BW. The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *The Journal of Neuroscience*. 2011; 31(33):11786–11794. [PubMed: 21849539]
- Corcoran KA, Maren S. Hippocampal inactivation disrupts contextual retrieval of fear memory after extinction. *The Journal of Neuroscience*. 2001; 21(5):1720–1726. [PubMed: 11222661]
- Corcoran KA, Maren S. Factors regulating the effects of hippocampal inactivation on renewal of conditional fear after extinction. *Learning & Memory*. 2004; 11(5):598–603. [PubMed: 15466314]
- Craig W. Appetites and aversions as constituents of instincts. *The Biological Bulletin*. 1918; 34:91–107.
- Debiec J, Díaz-Mataix L, Bush DE, Doyère V, LeDoux JE. The amygdala encodes specific sensory features of an aversive reinforcer. *Nature Neuroscience*. 2010; 13(5):536–537.
- Delamater AR. Effects of several extinction treatments upon the integrity of Pavlovian stimulus-outcome associations. *Animal Learning & Behavior*. 1996; 24(4):437–449.
- Delamater AR. Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *The Quarterly Journal of Experimental Psychology: Section B*. 2004; 57(2):97–132.
- Delamater AR. Issues in the extinction of specific stimulus-outcome associations in Pavlovian conditioning. *Behavioral Processes*. 2012; 90(1):9–19.
- Delamater AR, Campese V, Westbrook RF. Renewal and spontaneous recovery, but not latent inhibition, are mediated by GABA in appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*. 2009; 35:224–237. [PubMed: 19364231]
- Delamater AR, Holland PC. The influence of CS-US interval on several different indices of learning in appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*. 2008; 34(2):202–222. [PubMed: 18426304]
- Delamater AR, LoLordo VM, Sosa W. Outcome-specific conditioned inhibition in Pavlovian backward conditioning. *Learning & Behavior*. 2003; 31(4):393–402. [PubMed: 14733487]
- Delamater AR, Oakeshott S. Learning about multiple attributes of reward in Pavlovian conditioning. *Annals of the New York Academy of Sciences*. 2007; 1104:1–20. [PubMed: 17344542]
- Denny, MR. Relaxation theory and experiments. In: Brush, FR., editor. *Aversive conditioning and learning*. New York: Academic Press; 1971. p. 235–295.
- Díaz-Mataix L, Debiec J, LeDoux JE, Doyère V. Sensory-specific associations stored in the lateral amygdala allow for selective alteration of fear memories. *The Journal of Neuroscience*. 2011; 31(26):9538–9543. [PubMed: 21715618]
- Dickinson, A.; Dearing, M. Appetitive-aversive interactions and inhibitory processes. In: Dickinson, A.; Boakes, R., editors. *Mechanisms of learning and motivation*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1979. p. 203–231.
- Edeline JM, Weinberger NM. Associative returning in the thalamic source of input to the amygdala and auditory cortex: Receptive field plasticity in the medial division of the medial geniculate body. *Behavioral Neuroscience*. 1992; 106(1):81–105. [PubMed: 1554440]
- Edeline JM, Weinberger NM. Receptive field plasticity in the auditory cortex during frequency discrimination training: Selective returning independent of task difficulty. *Behavioral Neuroscience*. 1993; 107(1):82–103. [PubMed: 8447960]
- Garcia R, Chang C-h, Maren S. Electrolytic lesions of the medial prefrontal cortex do not interfere with long-term memory of extinction of conditioned fear. *Learning & Memory*. 2006; 13:14–17. [PubMed: 16418435]
- Ghazizadeh A, Ambroggi F, Odean N, Fields HL. Prefrontal cortex mediates extinction of responding by two distinct neural mechanisms in accumbens shell. *The Journal of Neuroscience*. 2012; 32(2):726–737. [PubMed: 22238108]
- Graham BM, Richardson R. Fibroblast growth factor-2 alters the nature of extinction. *Learning & Memory*. 2011; 18(2):80–84. [PubMed: 21233324]

- Harris JA, Jones ML, Bailey GK, Westbrook RF. Contextual control over conditioned responding in an extinction paradigm. *Journal of Experimental Psychology: Animal Behavior Processes*. 2000; 26(2):174–185. [PubMed: 10782432]
- Hendry JS. Summation of undetected excitation following extinction of the CER. *Animal Learning & Behavior*. 1982; 10(4):476–482.
- Herry C, Ciochi S, Senn V, Demmou L, Müller C, Lüthi A. Switching on and off fear by distinct neuronal circuits. *Nature*. 2008; 454(7204):600–606. [PubMed: 18615015]
- Herry C, Ferraguti F, Singewald N, Letzkus JJ, Ehrlich I, Lüthi A. Neuronal circuits of fear extinction. *The European Journal of Neuroscience*. 2010; 31(4):599–612. [PubMed: 20384807]
- Holland, PC. The nature of conditioned inhibition in serial and simultaneous feature negative discriminations. In: Miller, RR.; Spear, NE., editors. *Information processing in animals: Conditioned inhibition*. Hillsdale, NJ: Erlbaum; 1985. p. 267–298.
- Holmes NM, Westbrook RF. Extinction of reinstated or ABC renewed fear responses renders them resistant to subsequent ABA renewal. *Journal of Experimental Psychology: Animal Behavior Processes*. 2013; 39(3):208–220. [PubMed: 23668186]
- Hong I, Kim J, Lee J, et al. Reversible plasticity of fear memory-encoding amygdala synaptic circuits even after fear memory consolidation. *PlosOne*. 2011; 6(9):1–8.
- Ji J, Maren S. Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction. *Learning & Memory*. 2005; 12(3):270–276. [PubMed: 15930505]
- Ji J, Maren S. Hippocampal involvement in contextual modulation of fear extinction. *Hippocampus*. 2007; 17(9):749–758. [PubMed: 17604353]
- Kim J, Lee S, Park K, et al. Amygdala depotentiation and fear extinction. *Proceedings of the National Academy of Sciences*. 2007; 104(52):20955–20960.
- Kim JH, Richardson R. A developmental dissociation in reinstatement of an extinguished fear response in rats. *Neurobiology of Learning and Memory*. 2007a; 88(1):48–57. [PubMed: 17459734]
- Kim JH, Richardson R. A developmental dissociation of context and GABA effects on extinguished fear in rats. *Behavioral Neuroscience*. 2007b; 121(1):131–139. [PubMed: 17324057]
- Kim JH, Richardson R. The effect of temporary amygdala inactivation on extinction and reextinction of fear in the developing rat: Unlearning as a potential mechanism for extinction early in development. *The Journal of Neuroscience*. 2008; 28(6):1282–1290. [PubMed: 18256248]
- Kim JH, Richardson R. Extinction in preweanling rats does not involve NMDA receptors. *Neurobiology of Learning and Memory*. 2010; 94(2):176–182. [PubMed: 20472087]
- Kindt M, Soeter M. Reconsolidation in a human fear conditioning study: A test of extinction as updating mechanism. *Biological Psychology*. 2012; 92(1):43–50. [PubMed: 21986472]
- Konorski, J. *Conditioned reflexes and neuron organization*. New York: Cambridge University Press; 1948. *Conditioned reflexes and neuron organization*.
- Konorski, J. *Integrative activity of the brain: An interdisciplinary approach*. Chicago: University of Chicago Press; 1967.
- Kruse JM, Overmier JB, Konz WA, Rokke E. Pavlovian conditioned stimulus effects upon instrumental choice behavior are reinforcer specific. *Learning and Motivation*. 1983; 14(2):165–181.
- Kwapis JL, Jarome TJ, Lonergan ME, Helmstetter FJ. Protein kinase Mzeta maintains fear memory in the amygdala but not in the hippocampus. *Behavioral Neuroscience*. 2009; 123(4):844–850. [PubMed: 19634944]
- Lai CSW, Franke TF, Gan WB. Opposite effects of fear conditioning and extinction on dendritic spine remodeling. *Nature*. 2012; 483:87–92. [PubMed: 22343895]
- Larrauri JA, Schmajuk NA. Attentional, associative, and configural mechanisms in extinction. *Psychological Review*. 2008; 115(3):640–676. [PubMed: 18729595]
- Laurent V, Marchand AR, Westbrook RF. The basolateral amygdala is necessary for learning but not relearning extinction of context conditioned fear. *Learning & Memory*. 2008; 15(5):304–314. [PubMed: 18463174]

- Laurent V, Westbrook RF. Inactivation of the infralimbic but not the prelimbic cortex impairs consolidation and retrieval of fear extinction. *Learning & Memory*. 2009; 16(9):520–529. [PubMed: 19706835]
- Le Pelley ME. The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology: Section B*. 2004; 57(3):193–243.
- Leung HT, Westbrook RF. Spontaneous recovery of extinguished fear responses deepens their extinction: A role for error-correction mechanisms. *Journal of Experimental Psychology: Animal Behavior Processes*. 2008; 34(4):461–474. [PubMed: 18954230]
- Leung HT, Westbrook RF. Increased spontaneous recovery with increases in conditioned stimulus alone exposures. *Journal of Experimental Psychology: Animal Behavior Processes*. 2010; 36(3):354–367. [PubMed: 20658866]
- LoLordo, VM.; Fairless, J. Pavlovian conditioned inhibition: The literature since 1969. In: Miller, RR.; Spear, NE., editors. *Information processing in animals: Conditioned inhibition*. Hillsdale, NJ: Erlbaum; 1985. p. 1-50.
- Mackintosh NJ. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*. 1975; 82(4):276–298.
- Marchant NJ, Hamlin AS, McNally GP. Lateral hypothalamus is required for context-induced reinstatement of extinguished reward seeking. *The Journal of Neuroscience*. 2009; 29(5):1331–1342. [PubMed: 19193880]
- Marchant NJ, Millan EZ, McNally GP. The hypothalamus and the neurobiology of drug seeking. *Cellular and Molecular Life Sciences*. 2012; 69(4):581–597. [PubMed: 21947443]
- Maren S, Hobin JA. Hippocampal regulation of context-dependent neuronal activity in the lateral amygdala. *Learning & Memory*. 2007; 14:318–324. [PubMed: 17522021]
- Maren S, Chang C-h. Recent fear is resistant to extinction. *Proceedings of the National Academy of Sciences*. 2006; 103:18020–18025.
- Maren S, Quirk GJ. Neuronal signaling of fear memory. *Nature Reviews Neuroscience*. 2004; 5(11):844–852.
- MacPherson K, Whittle N, Camp M, Ozge G-C, Singewald N, Holmes A. Temporal factors in the extinction of fear in inbred mouse strains differing in extinction efficacy. *Biology of Mood & Anxiety Disorders*. 2013; 310.1186/2045-5380-3-13
- Mauk MD, Ohyama T. Extinction as new learning versus unlearning: Considerations from a computer simulation of the cerebellum. *Learning & Memory*. 2004; 11(5):566–571. [PubMed: 15466310]
- McNally GP, Johansen JP, Blair HT. Placing prediction into the fear circuit. *Trends in Neuroscience*. 2011; 34(6):283–292.
- McSweeney FK, Swindell S. Common processes may contribute to extinction and habituation. *The Journal of General Psychology*. 2002; 129(4):364–400. [PubMed: 12494990]
- Millan EZ, Milligan-Saville J, McNally GP. Memory retrieval, extinction, and reinstatement of alcohol seeking. *Neurobiology of Learning and Memory*. 2013; 101:26–32. [PubMed: 23305621]
- Monfils MH, Cowansage KK, Klann E, LeDoux JE. Extinction-reconsolidation boundaries: Key to persistent attenuation of fear memories. *Science*. 2009; 324(5929):951–955. [PubMed: 19342552]
- Morgan MA, Romanski LM, LeDoux JE. Extinction of emotional learning: Contribution of medial prefrontal cortex. *Neuroscience Letters*. 1993; 163(1):109–113. [PubMed: 8295722]
- Mowrer OH. Basic research methods, statistics and decision theory. *The American Journal of Occupational Therapy*. 1960; 14:199–205. [PubMed: 14424815]
- Myers KM, Carlezon WA, Davis M. Glutamate receptors in extinction and extinction-based therapies for psychiatric illness. *Neuropsychopharmacology*. 2011; 36(1):274–293. [PubMed: 20631689]
- Myers KM, Ressler KJ, Davis M. Different mechanisms of fear extinction dependent on length of time since fear acquisition. *Learning & Memory*. 2006; 13:216–223. [PubMed: 16585797]
- Paré D, Duvarci S. Amygdala microcircuits mediating fear expression and extinction. *Current Opinion in Neurobiology*. 2012; 22(4):717–723. [PubMed: 22424846]
- Pavlov, IP. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. 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(6):532–552. [PubMed: 7443916]
- Pearce, JM.; Mackintosh, NJ. Two theories of attention: A review and a possible integration. In: Mitchell, CJ.; Le Pelley, ME., editors. *Attention and associative learning: From brain to behavior*. Oxford: Oxford University Press; 2010. p. 11–40.
- Peters J, Kalivas PW, Quirk GJ. Extinction circuits for fear and addiction overlap in prefrontal cortex. *Learning & Memory*. 2009; 16(5):279–288. [PubMed: 19380710]
- Peters J, LaLumiere RT, Kalivas PW. Infralimbic prefrontal cortex is responsible for inhibiting cocaine seeking in extinguished rats. *The Journal of Neuroscience*. 2008; 28(23):6046–6053. [PubMed: 18524910]
- Pinard CR, Mascagni F, McDonald AJ. Medial prefrontal cortical innervation of the intercalated nuclear region of the amygdala. *Neuroscience*. 2012; 205:112–124. [PubMed: 22249157]
- Polack CW, Laborda MA, Miller RR. Extinction context as a conditioned inhibitor. *Learning & Behavior*. 2012; 40(1):24–33. [PubMed: 21786176]
- Quirk GJ, Armony JL, LeDoux JE. Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. *Neuron*. 1997; 19(3):613–624. [PubMed: 9331352]
- Quirk GJ, Mueller D. Neural mechanisms of extinction learning and retrieval. *Neuropsychopharmacology*. 2008; 33(1):56–72. [PubMed: 17882236]
- Rauhut AS, Thomas BL, Ayres JJ. Treatments that weaken Pavlovian conditioned fear and thwart its renewal in rats: Implications of treating human phobia. *Journal of Experimental Psychology: Animal Behavior Processes*. 2001; 27(2):99–114. [PubMed: 11296492]
- Reberg D. Compound tests for excitation in early acquisition and after prolonged extinction of conditioned suppression. *Learning and Motivation*. 1972; 3(3):246–258.
- Rescorla RA. Pavlovian conditioned inhibition. *Psychological Review*. 1969; 72(2):77–94.
- Rescorla, RA. Conditioned inhibition and facilitation. In: Miller, RR.; Spears, NS., editors. *Information processing in animals: Conditioned Inhibition*. Hillsdale, NJ: Erlbaum; 1985. p. 299–326.
- Rescorla RA. Preservation of Pavlovian associations through extinction. *The Quarterly Journal of Experimental Psychology A*. 1996; 49(3):245–258.
- Rescorla RA. Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes*. 2000; 26(4):428–438. [PubMed: 11056883]
- Rescorla RA. Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *The Quarterly Journal of Experimental Psychology: Section B*. 2001a; 54(1):53–68.
- Rescorla RA. Retraining of extinguished Pavlovian stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*. 2001b; 27:115–124. [PubMed: 11296487]
- Rescorla, RA. Experimental extinction. In: Mowrer, RR.; Klein, SB., editors. *Handbook of contemporary learning theories*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2001c. p. 119–154.
- Rescorla RA. Savings tests: Separating differences in rate of learning from differences in initial levels. *Journal of Experimental Psychology: Animal Behavior Processes*. 2002a; 28:369–377. [PubMed: 12395494]
- Rescorla RA. Effect of following an excitatory-inhibitory compound with an intermediate reinforcer. *Journal of Experimental Psychology: Animal Behavior Processes*. 2002b; 28(2):163–174. [PubMed: 11987873]
- Rescorla RA. Protection from extinction. *Learning & Behavior*. 2003; 31(2):124–32. [PubMed: 12882371]
- Rescorla RA. Deepened extinction from compound stimulus presentation. *Journal of Experimental Psychology: Animal Behavior Processes*. 2006; 32(2):135–144. [PubMed: 16634656]
- Rescorla RA, Heth CD. Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*. 1975; 1(1):88–96. [PubMed: 1151290]

- Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH.; Prokasy, WF., editors. Classical conditioning: Vol. 2. Current research and theory. New York: Appleton-Century-Crofts; 1972. p. 64-99.
- Robbins SJ. Mechanisms underlying spontaneous recovery In autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*. 1990; 16(3):235–249.
- Sacktor TC. How does PKM ζ maintain long-term memory? *Nature Reviews Neuroscience*. 2011; 12(1):9–15.
- Santini E, Quirk GJ, Porter JT. Fear conditioning and extinction differentially modify the intrinsic excitability of infralimbic neurons. *The Journal of Neuroscience*. 2008; 28(15):4028–4036. [PubMed: 18400902]
- Schiller D, Monfils M-H, Raio CM, Johnson DC, LeDoux JE, Phelps EA. Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*. 2010; 463(7):49–54. [PubMed: 20010606]
- Serrano P, Friedman EL, Kenney J, Taubenfeld SM, Zimmerman JM, Hanna J, Alberini C, Kelley AE, Maren S, Rudy JW, Yin JC, Sacktor TC, Fenton AA. PKMzeta maintains spatial, instrumental, and classically conditioned long-term memories. *PLoS Biology*. 2008; 23;6(12):2698–2706.
- Sharpe MJ, Killcross S. The prelimbic cortex contributes to the down-regulation of attention toward redundant cues. *Cerebral Cortex*. 2012;10.1093/cercor/bhs393
- Shema R, Sacktor TC, Dudai Y. Rapid erasure of long-term memory associations in the cortex by an inhibitor of PKM zeta. *Science*. 2007; 317(5840):951–953. [PubMed: 17702943]
- Sierra-Mercado D, Padilla-Coreano N, Quirk GJ. Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear. *Neuropsychopharmacology*. 2011; 36(2):520–529.
- Song EY, Boatman JA, Jung MW, Kim JJ. Auditory cortex is important in the extinction of two different tone-based conditioned fear memories in rats. *Frontiers in Behavioral Neuroscience*. 2010; 4:1–7. [PubMed: 20126432]
- Stafford JM, Maughan DK, Ilioi EC, Lattal KM. Exposure to a fearful context during periods of memory plasticity impairs extinction via hyperactivation of frontal-amygdalar circuits. *Learning & Memory*. 2013; 20:156–163. [PubMed: 23422280]
- Takahashi YK, Roesch MR, Stalnaker TA, Haney RZ, Calu DJ, Taylor AR, Burke KA, Schoenbaum G. The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron*. 2009; 62:269–280. [PubMed: 19409271]
- Vidal-Gonzalez I, Vidal-Gonzalez B, Rauch SL, Quirk GJ. Microstimulation reveals opposing influences of prelimbic and infralimbic cortex on the expression of conditioned fear. *Learning & Memory*. 2006; 13(6):728–733. [PubMed: 17142302]
- Volk LJ, Bachman JL, Johnson R, Yu Y, Haganir RL. PKM- is not required for hippocampal synaptic plasticity, learning and memory. *Nature*. 2013; 493(7432):420–423. [PubMed: 23283174]
- Wagner, AR. Stimulus-selection and a “modified continuity theory”. In: Bower, GH.; Spence, JT., editors. *The psychology of learning and motivation*. Vol. 3. New York: Academic Press; 1969.
- Wagner, AR.; Brandon, SE. Evolution of a structured connectionist model of Pavlovian conditioning (AESOP). In: Klein, SB.; Mowrer, RR., editors. *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory*. Hillsdale, NJ: Erlbaum; 1989. p. 149-189.
- Weinberger NM. Auditory associative memory and representational plasticity in the primary auditory cortex. *Hearing Research*. 2007a; 229(1–2):54–68. [PubMed: 17344002]
- Weinberger NM. Associative representational plasticity in the auditory cortex: A synthesis of two disciplines. *Learning & Memory*. 2007b; 14(1–2):1–16. [PubMed: 17202426]
- Williams DA, Butler MM, Overmier JB. Expectancies of reinforcer location and quality as cues for a conditional discrimination in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*. 1990; 16(1):3–13. [PubMed: 2303792]
- Xue YX, Luo YX, Wu P, Shi HS, Xue LF, Chen C, Zhu WL, Ding ZB, Bao YP, Shi J, Epstein DH, Shaham Y, Lu L. A memory retrieval-extinction procedure to prevent drug craving and relapse. *Science*. 2012; 336(6078):241–245. [PubMed: 22499948]

- Zheng X, Deschaux O, Lavigne J, Nachon O, Claren C, Moreau JL, Garcia R. Prefrontal high-frequency stimulation prevents sub-conditioning procedure-provoked, but not acute stress-provoked, reemergence of extinguished fear. *Neurobiology of Learning & Memory*. 2013; 101:33–38. [PubMed: 23313613]
- Zimmer-Hart CL, Rescorla RA. Extinction of Pavlovian conditioned inhibition. *Journal of Comparative & Physiological Psychology*. 1974; 86:837–845. [PubMed: 4833588]

Figure 1. Number of publications by year through PubMed with the search terms "experimental extinction."

Figure 2.

Rescorla-Wagner (1972) model simulation results showing the associative strengths (US activation values) to a hypothetical stimulus X that received CS-US pairings during an acquisition phase, nonreinforced presentations during an extinction phase, and further nonreinforced presentations during a test phase. The upper left graph depicts associative strengths in an ABA renewal design where stimulus X is trained in Context A, extinguished in Context B, and then tested in either Context A or B. The upper right and lower left graphs show, respectively, corresponding simulation results from AAB and ABC renewal experiments. In all three cases greater associative strengths are produced under conditions that generate renewal, and the magnitude of ABA renewal is greater than either ABC or AAB renewal. The lower right graph simulates the US reinstatement effect, where training, extinction, and test all occur in the same context but prior to testing either the US is separately presented or not. The US reinstatement procedure is simulated to produce higher levels of responding to stimulus X.

Figure 3.

Detailed Rescorla-Wagner model (1972) simulation of an ABA renewal experiment. Associative strengths (US activation values) are shown separately for stimulus X, conditioning context A, and extinction context B, as well as the combined strengths of stimulus X with context A (Ctx A + X) and stimulus X with context B (Ctx B + X). Notice, in the left panel, that the combined associative strength to stimulus X and context A shows a negatively accelerated learning curve during the acquisition phase, and that this reflects the additive sum of the separate associative strengths conditioned to stimulus X and context A. In the middle panel, there is a modest decrease in associative strength to stimulus X together with the development of negative associative strength to context B. The additive sum of these leads to overall decreased US activation across Context B + X trials. When testing occurs back in Context A (right panel) there is a spontaneous increase in associative strength to the Context A + stimulus X compound. This is caused by (a) removal of the inhibitory Context B and (b) the reintroduction of stimulus X into the still excitatory Context A.