Cell Press

Rethinking Extinction

Joseph E. Dunsmoor,[1,*] Yael Niv,[2] Nathaniel Daw,[2] and Elizabeth A. Phelps[1,3,*]
[1]Department of Psychology and Center for Neural Sciences, New York University, New York, NY 10003, USA
[2]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[3]Nathan Kline Institute, Orangeburg, NY 10962, USA
*Correspondence: joseph.dunsmoor@nyu.edu (J.E.D.), liz.phelps@nyu.edu (E.A.P.)

http://dx.doi.org/10.1016/j.neuron.2015.09.028

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

Introduction

Along with the discovery of the conditioned response (CR), one of Pavlov's most significant contributions to physiology and to psychological science was the observation that absence of reinforcement resulted in a weakening or disappearance of acquired behavior. Termed by Pavlov as the "internal inhibition of conditioned reflexes" (Pavlov, 1927), experimental extinction generated theoretical and empirical research interest throughout the 20th century, but research on extinction paled in comparison to studies of conditions that generate acquisition of CRs. In the past decade, however, there has been a surge of interest in experimental extinction for its own sake. The topic spans neurobehavioral studies in laboratory animals and humans, cellular, molecular and genetic research, and computational learning models. Beyond interest in the basic mechanisms of learning and memory, renewed attention to extinction is due in large part to the clinical significance of extinction for the treatment of a variety of psychiatric disorders (Milad and Quirk, 2012; Vervliet et al., 2013). Specifically, extinction serves as the basis for exposure-based therapy, a primary treatment for anxiety disorders, addiction, and trauma- and stress-related disorders (Powers et al., 2010). Experimental extinction is also considered within the National Institute of Mental Health's Research Domain Criteria as a scientific paradigm to provide objective neurobehavioral measures of mental illness in the domain of Negative Affect. It is hoped that advances in our understanding of extinction across multiple fronts will translate to new, effective treatments for psychiatric conditions characterized by the inability to regulate pathological fear or anxiety.

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

The first section is composed of a brief background on the theoretical foundation upon which contemporary views of extinction rest, a description of the neurobiology of extinction, psychological factors, and major associative learning models. A primary question is whether the mechanisms supporting extinction involve new learning that inhibits or interferes with original learning, as is the current mainstay, or also cause erasure of the original learning, as suggested by recent theoretical and experimental work. In particular, we survey a recent framework that reinterprets extinction in terms of sound statistical reasoning about the causes of events in the world, and suggest that this framework can conceptualize the trade-off between new learning and memory modification. In the second section, we detail the shortfalls of traditional extinction techniques in preventing the return of unwanted behaviors and discuss novel approaches to augment extinction that compensate for these shortfalls. We attempt to understand the success of these approaches in terms of several distinct theoretical mechanisms, including interference and erasure, which might contribute to extinction. Of note, we focus almost exclusively on extinction in the domain of fear or threat conditioning, as it is in this arena that many of the advances in neuroscience, behavior, learning theory, and clinical translational research have been made.

Foundational Research and Theories of Extinction

The canonical expression of experimental extinction rests on Pavlovian conditioning, in which a conditional stimulus (CS; e.g., a tone or light) is paired with a naturally salient unconditional stimulus (US; e.g., food or an electric shock). Once a relationship between the CS and US is established, presentation of the CS

Box 1. Current Status of the Field

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

Box 2. Future Directions

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

initiates a conditioned response (e.g., increases in salivation). In the domain of fear conditioning, in which the US is naturally unpleasant or painful, the CR often takes the form of defensive behaviors or emotional reactions such as increases in sweating, heart rate, pupil size, freezing, and blood pressure. With continuing presentation of the CS in the absence of the US, the CR gradually diminishes or is eliminated altogether.

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

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

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