The Role and Mechanism of Prediction Error in Fear Memory Updating

Authors: Li Junjiao, Chen Wei, Shi Pei, Dong Yuanyuan, Zheng Xifu, Zheng Xifu

Date: 2023-11-01T00:00:00+00:00

Abstract

According to error-driven learning theory, the mismatch between expected and actual behavioral outcomes—prediction error (PE)—serves as the driving force for learning. As a form of salience information, PE differs from physical salience, surprise, and novelty in information processing stages, and its relationship with memory updating also varies. In recent years, the memory reconsolidation interference paradigm has been demonstrated to be effective for updating human conditioned fear memories, wherein the prediction error embedded in the memory retrieval/activation phase plays a critical role in triggering memory "destabilization" and initiating memory reconsolidation. Regarding the behavioral mechanisms that facilitate fear memory updating, PE is considered a necessary but not sufficient condition for memory destabilization. Memory retrieval must contain an appropriate level of PE, yet whether it triggers memory destabilization, extinction, or an intermediate state requires determination based on the intrinsic properties of the memory itself. In terms of the neural mechanisms underlying fear memory updating, the amygdala, periaqueductal gray (PAG), and hippocampus all play important roles in PE detection and computation; the prefrontal cortex (PFC) and its subregions assume a significant role in PE initiating memory reconsolidation. These processes are further modulated by specific neurotransmitters in the nervous system, particularly dopaminergic and glutamatergic systems. Future research should further explore quantitative studies based on PE computational models, integrate the interactions between PE and other boundary conditions, and investigate the role of different types of salience in memory reconsolidation; moreover, there is an urgent need to employ multidisciplinary approaches to explore the neural and molecular mechanisms underlying PE' s role in fear memory updating. Simultaneously, further studies on individual differences in PE effects are required to facilitate the translation of research findings into clinical applications.

Full Text

The Role and Mechanisms of Prediction Error in Fear Memory Updating

LI Junjiao[1], CHEN Wei[2,3,4], SHI Pei[2,3,4], DONG Yuanyuan[2,3,4], ZHENG Xifu[2,3,4]

[1]College of Teacher Education, Guangdong University of Education, Guangzhou 510303, China
[2]School of Psychology, South China Normal University, Guangzhou 510631, China
[3]Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China
[4]Guangdong Key Laboratory of Mental Health and Cognitive Science, Guangzhou 510631, China

Abstract

According to error-driven learning theory, the mismatch between expected and actual behavioral outcomes—known as prediction error (PE)—drives new learning. As a type of salient information, PE differs from physical salience, surprise, and novelty in its information processing stages and its relationship with memory updating. In recent years, the reconsolidation interference paradigm has proven effective for updating conditioned fear memories in humans, wherein PE during memory retrieval plays a critical role in triggering memory "destabilization" and initiating reconsolidation. At the behavioral level, PE is considered a necessary but insufficient condition for memory destabilization; memory retrieval must contain an appropriate degree of PE, though whether this leads to destabilization, extinction, or an intermediate state depends on the nature of the memory itself. At the neural level, the amygdala, periaqueductal gray (PAG), and hippocampus are important for PE detection and computation, while the prefrontal cortex (PFC) and its subregions play crucial roles in PE-initiated memory reconsolidation. These processes are further modulated by specific neurotransmitters, particularly dopaminergic and glutamatergic systems. Future research should further explore quantitative studies based on PE computational models, integrate the interactions between PE and other boundary conditions, and investigate the role of different types of salience in memory reconsolidation. Multidisciplinary approaches are urgently needed to explore the neural and molecular mechanisms of PE in fear memory updating, and further studies on individual differences in PE effects are essential to facilitate translation from laboratory findings to clinical applications.

Keywords: prediction error, conditioned fear, memory updating, reconsolidation, retrieval-extinction paradigm

Phobias, anxiety disorders, and post-traumatic stress disorder (PTSD) represent major categories of mental illness in China. Recent natural disasters and

public health emergencies have caused varying degrees of psychological distress among victims and witnesses. Exposure therapy based on extinction training is a primary clinical treatment for phobias and anxiety disorders, but it suffers from high relapse rates. Research indicates that traditional extinction training does not eliminate or update original memories but rather establishes a new safety memory that competes with the original fear memory, making relapse likely under various conditions. In recent years, the retrieval-extinction paradigm based on memory reconsolidation theory has proven effective in eliminating fear memories and preventing relapse. This approach works by inducing and interfering with the destabilized state of memory, preventing its restabilization and thereby disrupting the original memory association. However, this paradigm is constrained by boundary conditions of reconsolidation—factors that limit effective memory activation into an unstable state. Among these, prediction error (PE) during memory retrieval represents a crucial boundary condition. Therefore, investigating the role and neural mechanisms of PE in fear memory updating (including elimination and modification) is essential for resolving theoretical challenges and facilitating translation from laboratory findings to clinical practice.

1.1 Error-Driven Learning Theory

Error-driven learning theory posits that learning occurs only when a reinforcer is surprising or unpredictable. When a behavior produces an unexpected outcome, new learning emerges; if the actual result perfectly matches expectations and aligns with stored memory, no new learning occurs; and if a previously learned behavior no longer produces the expected outcome, that behavior extinguishes (Schultz, 2000). The mismatch between expected and actual outcomes is called prediction error (PE). Thus, learning is fundamentally an error-driven process.

This principle is exemplified by the blocking phenomenon in classical conditioning. For instance, when a tone repeatedly predicts food, the tone alone eventually elicits salivation, indicating a tone-food association has formed. If a light is then added to the tone (tone+light) to predict food, subsequent presentation of the light alone does not elicit salivation, showing no light-food association has formed. Because prior learning allowed the tone alone to predict food, this experience fully matched expectations, preventing new learning when additional stimuli were paired with the expected reinforcer (Schultz et al., 1997).

In fear extinction learning, when a conditioned stimulus (CS) that previously predicted a negative outcome no longer does so, individuals gradually form a CS-safety association, reflecting adaptation to a changing environment. A key factor in extinction or safety learning is therefore the change in expectation. When expectations and outcomes are highly consistent, prior experience suffices and new learning does not occur. Only when expectations and outcomes mismatch—when individuals realize prior experience no longer applies—does the motivation for learning emerge. Thus, under this theoretical framework, PE plays an important role in both fear acquisition and extinction, serving as the

basic driver of learning.

1.2 Theoretical Models of Prediction Error

Classic computational models of prediction error 主要包括三类，分别是 Rescorla & Wagner 模型、Pearce-Hall 模型，以及时间预期错误 (temporal difference, TD) 的模型。这三类模型既是预期错误模型，又是关于学习的模型 (model of learning).

The Rescorla-Wagner (RW) model proposes that in conditioning, learning is controlled by an error signal reflecting the difference between actual and expected unconditioned stimulus (US) intensity (Rescorla & Wagner, 1972). If actual US is denoted as and expected US as ΣV (the sum of all CS-US association strengths V), the error signal is -ΣV. The synaptic modification rule for changes in association strength can be expressed as:

$$\Delta V = S(\lambda - \Sigma V)$$

where S is the learning rate.

The Pearce-Hall (PH) model posits that learning occurs only when the reinforcer is surprising (Pearce & Hall, 1980). This model uses the absolute value of RPE, equivalent to unsigned RPE (which does not distinguish positive from negative PE). The PH model suggests that the error signal regulates the amount of attention allocated to each conditioning trial. If attention magnitude is denoted as α, its value on trial n is proportional to the prediction error on the previous trial:

$$\alpha_n = |\lambda - \Sigma V|_{n-1}$$

The signal for synaptic plasticity modification can then be expressed as:

$$\Delta V = \alpha_n S\lambda$$

If the CS had low predictive validity for the US on the previous trial, α will be larger, driving stronger synaptic plasticity; conversely, if α is small, the drive for association strength change on the current trial will be low (McNally et al., 2011).

The temporal difference (TD) model proposes that learning occurs when expectations about reward change between two time points (t, t+1). Unlike the previous models, TD does not compute error between actual and expected US, but rather compares the sum of actual and expected US intensity at time t (+ ΣV) with the expected US intensity at the previous moment. Thus, the error signal in TD can be expressed as:

$$\Delta V = S(\lambda_t + \Sigma V_t - \Sigma V_{t-1})$$

where t and t-1 are consecutive moments. TD is therefore defined as actual or expected US intensity (or their sum) exceeding what was expected at the previous time point (Ergo et al., 2020; McNally et al., 2011).

These three learning models each have distinct emphases. The PH model does not distinguish PE direction, considering only absolute differences. The RW model incorporates directionality, considering cases where outcomes are greater or less than expected, thus distinguishing PE types while emphasizing external reinforcement. The TD model emphasizes internal reinforcement and the timing of reinforcement, proposing that temporal information can transmit prediction error, expanding the scope and form of PE. These theories, particularly the RW and TD models, have profoundly influenced research on PE-driven learning and memory.

1.3 Classification of Prediction Error

Reviewing previous research reveals that PE can be classified differently across learning models. The main categories include:

First, **reward prediction error (RPE)** and **punishment prediction error (PPE)**. In operant conditioning, PE is divided based on outcome valence: RPE and PPE. Schultz defined reward as an operational concept describing positive attributes assigned to objects, actions, or internal physical states (Schultz et al., 1997). Reward value associated with a stimulus is not an inherent property of the stimulus itself. RPE signals are closely linked to the midbrain dopamine (DA) system, while PPE pathways are thought to be associated with mental disorders. The relationship between RPE and midbrain dopaminergic neurons has been extensively validated (Colombo, 2014; Kim et al., 2014; Papalini et al., 2020; Starkweather et al., 2017).

Second, **positive PE** and **negative PE**. In Pavlovian fear conditioning, based on the relative magnitude of actual versus expected US, PE can be positive (actual US greater than expected) or negative (actual US less than expected, including US absence). Negative PE can be achieved by increasing expectations or decreasing US intensity, with the simplest model being extinction, where CS presentation without US constitutes negative PE. Some researchers propose that positive PE leads to fear learning while negative PE leads to fear extinction (Rescorla & Wagner, 1972). To enhance extinction, one strategy is to maximize PE: larger negative PE produces stronger learning drive. As expectations adjust, PE gradually approaches zero, learning completes and stops, achieving a new equilibrium.

Third, **signed prediction error (SPE)** and **unsigned prediction error (UPE)**. Based on whether direction is specified, PE can be classified as signed or unsigned. The "sign" refers to the valence indicator (+ or -) that denotes whether the actual outcome is greater or smaller than expected (Ergo et al., 2020). SPE explicitly indicates direction, while UPE only signals mismatch without specifying direction. Research shows these two types have different

neural substrates and processing mechanisms. According to the three computational models, both RW and TD models depend on SPE. Positive SPE direction increases dopamine release, while negative SPE direction decreases dopamine consumption.

Comparing RPE/PPE with positive/negative PE reveals differences in their underlying memory models and meanings. RPE and PPE are common in instrumental conditioning, distinguished by outcome properties resulting from behavior. Positive and negative PE are more common in Pavlovian conditioning, which lacks operant behavior and is defined by CS-US relationships. When a CS is expected to be followed by US but isn' t, or when the actual US is smaller than expected, this constitutes negative PE. When a CS is not expected to be followed by US but is, or when the actual US is larger than expected, this constitutes positive PE. Theoretically, reward-negative PE or punishment-positive PE are possible. Finally, signed and unsigned PE represent a classification independent of specific learning models, based solely on directionality.

1.4 Relationship Between Prediction Error and Other Types of Salience

For evolutionary reasons, attention is captured by salient or conspicuous stimuli, collectively termed salience. Salience is a macro-concept, and fundamentally, all types of PE possess salience. Researchers suggest different types of salience reflect different depths of stimulus processing (Diederen & Fletcher, 2021). Because the salience concept encompasses multiple types—including physical salience, surprise, novelty, and incomplete cues—recent literature has become confusing, necessitating clarification of these concepts and their functions. Hierarchically, salience includes novelty, valence evaluation, rarity, and other salience types. Novelty further includes physical salience, surprise (unexpected novelty), and non-surprising novelty. Surprise corresponds to unsigned PE (UPE), while valence processing corresponds to signed PE (SPE) [Figure 1: see original paper].

These salience types follow two logical relationships: first, they belong to different information processing stages. Stimulus processing involves sensory input, preliminary perception of stimulus outcomes, and valence evaluation of associated reinforcers—each stage may contain salient information. Second, they correspond to different patterns of midbrain dopamine activity: brief enhancement versus substantial release [Figure 1: see original paper].

1.4.1 Physical Salience Physically salient sensory inputs can trigger extremely rapid phasic dopamine neuron responses within 50–110 ms, such as VTA dopaminergic neuron responses to light stimuli. This brief timeframe is insufficient for detailed processing, recognition, and evaluation, suggesting physical salience is not directly associated with reward or reinforcement outcomes, though physically salient stimuli may have reinforcement potential. Some argue

physical stimulus salience itself constitutes reinforcement due to evolutionary advantages in rapidly identifying potential threats (Diederen & Fletcher, 2021).

1.4.2 Surprise Surprise generally includes outcome properties, not merely sensory input like physical salience. However, surprise does not contain valence —it simply indicates a mismatch between actual and expected outcomes without positive or negative direction, making it an unsigned PE (UPE). Studies show surprise and signed PE (SPE) are processed in different brain regions: surprise primarily activates superior frontal regions, while SPE is processed in striatal or midbrain areas. Some studies therefore equate surprise with PE when examining its ability to induce memory destabilization (Sinclair & Barense, 2018).

1.4.3 Novelty Both physical salience and surprise belong to novelty, which is itself a type of salience. Compared to novelty, salience is a broader concept. Dopamine neurons typically increase firing in response to novel stimuli, but this release habituates when the novel stimulus becomes familiar and unreinforced. Human fMRI studies show the substantia nigra (SN)/VTA responds to novel stimuli, while other salience types like rarity and negative emotion do not activate this region. Novelty processing includes early recognition and later processing, influenced differently by dopamine. Importantly, novelty does not always trigger dopamine release—only unexpected novelty induces dopamine release, similar to prediction error (Diederen & Fletcher, 2021).

Based on these findings and integrating previous models (Schultz, 2016), the relationships among these concepts and their processing stages can be clearly represented [Figure 1: see original paper].

Figure 1. Schematic diagram of relationships among salience-related concepts

Note: Salience includes stimulus novelty, valence evaluation, stimulus rarity, and other salience. Novelty includes physical salience, surprise (unexpected novelty), and non-surprising novelty, with only unexpected novelty triggering dopamine release. Physical salience alone, without direct relationship to outcomes, only causes brief dopamine enhancement insufficient for release. Prediction error (including UPE and SPE) primarily involves recognition and outcome perception or valence evaluation processes.

Finally, the concept of "incomplete reminders" deserves mention. Recent researchers have attempted to integrate associative memory (e.g., conditioned fear memory) and declarative memory reconsolidation studies using "incomplete retrieval cues" (Sinclair & Barense, 2019). In declarative memory studies, using initial learning stimuli as retrieval cues with missing or altered subsequent content makes prior memories more susceptible to post-retrieval interference, demonstrating PE' s critical role in declarative memory reconsolidation. Researchers argue that both memory types require incomplete cues during retrieval for successful reconsolidation intervention, which essentially constitutes prediction error (Sinclair & Barense, 2019).

2.1 The Role of Prediction Error in Fear Acquisition and Extinction

Prediction error is considered essential for fear acquisition, mediating the transition from no reinforcer following CS to actual reinforcer presentation (Furlong et al., 2010). As noted, the midbrain periaqueductal gray (PAG) is part of the neural circuit mediating PE-driven fear acquisition. Recent research has further examined the causal relationship between the ventrolateral PAG (vlPAG) and fear conditioning formation. Using differential protocols in rats—threatening CS paired with certain shock versus uncertain cues paired with probabilistic shock (the latter learned through PE)—research found that shock-induced responses aligned with SPE changes, and vlPAG inhibition reduced subsequent fear responses. This indicates PE is necessary for maintaining fear responses under uncertainty, establishing a causal relationship (Walker et al., 2020), consistent with previous conclusions about PE and learning (Fernandez, Boccia, et al., 2016).

During extinction, the RW model suggests that greater mismatch between expected CS-US and actual CS-US increases the likelihood of new learning (Rescorla & Wagner, 1972). In extinction learning, CS+ presentation without US produces two possible outcomes: fear activation and fear extinction. Initial CS+ presentations merely retrieve the original fear memory, eliciting fear responses. When CS+ is repeatedly presented without US, an inhibitory memory association forms and original fear responses decrease. This process critically depends on expectation adjustment. The mismatch between expectation and outcome motivates the need to establish new memory associations, suggesting PE initiates fear extinction, with negative PE being the origin of successful extinction.

Gershman and Monfils (2017) proposed a memory modification model based on structure learning mechanisms, suggesting standard fear acquisition and extinction create two memory associations: CS-US and US-no US. PE serves dual functions: as a signal for associative learning that guides adjustment of CS-US association weights, and as a segmentation signal indicating when a new latent cause becomes active. During extinction, the expectation of CS-US relationship formed during acquisition is violated, generating PE. This PE can be reduced through two pathways: unlearning or forgetting the original CS-US association, or assigning extinction trials to a new latent cause. Researchers suggest that at PE onset, a simple bias toward few latent causes favors forgetting; as extinction progresses, cumulative PE eventually elevates a new latent cause, creating a CS-no US association (Gershman et al., 2017).

However, research on the neural mechanisms of PE in extinction learning remains limited in both animal and human models. The hippocampus (HIP) and ventral tegmental area (VTA) are generally thought to be involved, with the HIP-VTA circuit playing an important role in reinforcement-based memory encoding, though its function in human fear extinction requires verification (Sevenster et al., 2018). Additionally, PE' s role in fear extinction is modulated

by various factors including sleep (particularly REM sleep) and stress hormones. Regarding neurotransmitters, substantial evidence demonstrates that midbrain dopamine neuron activity represents the degree to which actual outcomes are better or worse than expected (Schultz, 2016; Schultz et al., 1997), and the absence of negative US (e.g., shock) during fear extinction can be viewed as an outcome "better than expected" (Raczka et al., 2011; Thiele et al., 2021).

2.2.1 Prediction Error as a Critical Boundary Condition for Memory Destabilization

Memory reconsolidation theory posits that after stable long-term memories are retrieved via cues, they return to an unstable state, become susceptible to interference, and must undergo a process to restabilize—this phase is called "reconsolidation." This theory identifies two key windows for memory modification: consolidation and reconsolidation. Numerous studies have confirmed the independence of the reconsolidation phase and the feasibility of disrupting it to eliminate fear memories (Alberini et al., 2006; Duvarci & Nader, 2004; Lee et al., 2006; Nader et al., 2000). Behaviorally, Monfils and Schiller demonstrated in animals and humans that extinction training after memory retrieval effectively eliminates fear memories and prevents their return (Monfils et al., 2009; Schiller et al., 2010), termed the "retrieval-extinction" (RE) paradigm. A typical RE procedure spans three consecutive days: Day 1 establishes CS-US memory; 24 hours later, a CS retrieves the memory, followed by extinction after 10 minutes; Day 3 tests fear relapse (Chen et al., 2021; Schiller et al., 2010). Recent research suggests reconsolidation can be further divided into destabilization and restabilization phases (Elsey & Kindt, 2017; Faliagkas et al., 2018). Only when retrieval makes prior memories unstable again can updating (elimination or modification) occur. This process is called destabilization, and the conditions required to enter this state are termed boundary conditions of memory reconsolidation (Zuccolo & Hunziker, 2019). Therefore, memory reactivation and opening the fear memory reconsolidation window are two essential prerequisites for ensuring extinction efficacy and preventing fear return.

Prediction error, as a violation of expectation, has been extensively studied in reward learning, but its special significance for initiating reconsolidation was only recently discovered. In 2009, Lee suggested that surprise or PE might play a potential role in memory reconsolidation (Lee, 2009). Subsequently, Kindt' s team demonstrated in human subjects that PE during conditioned fear memory activation is necessary for entering an unstable state (Sevenster et al., 2012, 2013, 2014), sparking international interest and follow-up research. Various forms of PE have proven effective in initiating memory reconsolidation, including TD, PE from learning rules, and PE from US frequency (Chen et al., 2020; Diaz-Mataix et al., 2013; Junjiao et al., 2019; Li et al., 2017; Sevenster et al., 2013, 2014). PE' s role has shown similar conclusions in human and animal subjects and has been validated across different memory types including addiction memory and declarative memory, suggesting it is a common component of memory updating

processes (Fernandez, Bavassi, et al., 2016; Forcato et al., 2007; Das et al., 2018; Sinclair & Barense, 2019). New learning initiation requires novel information as a driving force. When no meaningful new information exists in the environment, synaptic plasticity of memory neurons remains closed, preventing new dendrites and neuronal connections. Only when novel, adaptively relevant, and survival-significant information appears can synaptic plasticity be activated, generating new dendrites and altering connection strength—a process requiring new protein synthesis for restabilization (Diaz-Mataix et al., 2013).

Sevenster et al. (2013) first reported using PE to trigger memory reconsolidation, defining PE as mismatch between acquisition and retrieval phases. They created three retrieval conditions: no PE, positive PE, and negative PE. No PE meant the CS-US relationship during retrieval matched acquisition. Using propranolol to verify reconsolidation, they found both PE groups underwent reconsolidation while the no-PE group did not (Sevenster et al., 2013). To further examine PE' s role, researchers tested different PE generation methods in the RE paradigm. Díaz-Mataix et al. (2013) used TD for retrieval and compared fear relapse on Day 3, finding the TD group successfully prevented fear relapse through retrieval-extinction while the non-TD group showed significant fear return, proving that temporal PE is necessary for initiating memory reconsolidation.

Building on this work, we used compound stimuli in the RE paradigm to establish conditioned fear, verifying the effects of positive and negative PE in activating compound fear memories (Chen et al., 2018) and comparing CS novelty versus CS-US novelty (PE) in initiating reconsolidation (Junjiao et al., 2019). Results showed positive and negative PE had equivalent effects in initiating reconsolidation—whether the actual US was larger or smaller than the original US, both triggered the need to update the original CS-US association. However, CS novelty alone was insufficient to initiate reconsolidation; CS-US novelty (i.e., PE) was the critical factor and necessary condition for activating memory into reconsolidation.

2.2.2 The Degree of Prediction Error Determines Whether Memory Enters an Unstable State

PE is both a key factor in fear extinction and a necessary condition for entering reconsolidation. However, these two processes (memory updating and new learning) are diametrically opposed. Memory retrieval under specific experimental settings may induce either process or an intermediate "limbo state" (Faliagkas et al., 2018). Whether PE-containing retrieval can trigger reconsolidation thus becomes critical. Research indicates PE is a necessary but insufficient condition for reconsolidation trigger. Even when retrieval contains PE, whether it successfully induces memory destabilization depends on the degree of PE. Sevenster and Kindt created three retrieval conditions based on PE degree: no PE, single PE, and multiple PE, examining their effects on initiating reconsolidation. Results showed significant fear relapse on Day 3 in both no-PE and multiple-PE conditions, while only the single-PE condition prevented relapse (Sevenster

et al., 2014). This study was the first to investigate PE degree as a crucial factor determining memory entry into reconsolidation, providing important inspiration for subsequent research. We further translated this paradigm from pharmacological to behavioral intervention, obtaining consistent results (Chen et al., 2018). Recent studies in declarative memory have used confidence ratings combined with feedback to quantify PE, achieving some progress (Pine et al., 2018). However, in conditioned fear research, PE quantification remains at a relatively crude categorical level. Since these studies, the shift from qualitative to quantitative examination of PE' s precise role in memory updating has become an important marker of progress in this field.

2.2.3 The Degree of Prediction Error Required for Memory Destabilization Relates to Original Memory Strength

Recently, researchers have recognized that the two types of boundary conditions in memory reconsolidation are not independent but likely interact to influence retrieval effects. Consequently, interactions between memory properties (e.g., strength) and retrieval boundaries have gained attention. Research on these interactions not only clarifies mechanisms underlying boundary effects on retrieval-extinction but also provides foundations for improving experimental paradigms and developing new clinical treatments. As a retrieval boundary condition, does PE' s role in initiating reconsolidation vary with memory strength? Does stronger memory require more PE for destabilization? These are important questions. We recently explored whether different strengths of conditioned fear memory require different degrees of PE for destabilization in human subjects. Using the RE paradigm, we established weaker fear memories using predictable shock timing (CS-predictable US) and stronger fear memories using unpredictable shock timing (CS-unpredictable US) during acquisition (Amadi et al., 2017). During retrieval, we examined three conditions: single PE, multiple PE, and single CS with two US omissions. Results showed that retrieval-extinction with single PE prevented relapse of predictable CS-US fear memories but not unpredictable ones, indicating single PE was insufficient for destabilizing stronger fear memories. Both multiple PE and single CS with two US omissions suppressed spontaneous recovery, but only multiple PE prevented fear reinstatement, suggesting strong fear memories may require more PE to disrupt stability. The degree of PE needed for destabilization depends on fear memory strength (Chen et al., 2020).

In another study, we re-examined PE' s role across different fear memory strengths and explored the potential role of post-retrieval acute stress in initiating reconsolidation for stronger fear memories, based on glutamatergic neurons' role in activating synaptic plasticity. For moderate-strength fear memories, single PE retrieval-extinction significantly suppressed spontaneous recovery, while single PE was ineffective for stronger fear memories, which showed significant relapse on Day 3. Exogenous acute stress after retrieval further increased fear recovery. These findings demonstrate the variability

of PE as a boundary condition: PE requirements may change according to memory strength (Li et al., 2021). However, since neither study directly verified whether stronger fear memories require larger PE, this question awaits future investigation.

In summary, prediction error or novel information during retrieval plays a crucial role in transitioning memory from a stable state to an unstable state via synaptic plasticity. Whether the US following the retrieval cue CS is larger or smaller than the original US, both can trigger the need to update the original CS-US association. However, PE's specific effect—whether it triggers destabilization, extinction, or an intermediate state—must be determined in combination with the memory' s boundary conditions. Based on these findings, we propose an integrated model of fear memory reconsolidation that combines retrieval boundary conditions with memory conditions [Figure 2: see original paper].

Figure 2. Integrated model of memory reconsolidation combining retrieval boundary conditions and memory conditions
Note: Fear memory properties (e.g., strength) interact with retrieval boundary conditions (e.g., prediction error). For moderate-strength fear memories, single PE can activate memory into reconsolidation for subsequent modification. Excessive PE induces extinction. However, for strong fear memories, single PE cannot activate the original memory for destabilization; the degree of PE required for reconsolidation remains largely unknown.

3. Neural Mechanisms of Prediction Error in Fear Memory Updating

Given PE' s importance, both basic research and clinical applications urgently need operational indicators of PE occurrence (Chen et al., 2020). Researchers note that human experiments have an advantage over animal studies in allowing verbal reports of US expectancy, which can serve as an explicit indicator of PE. However, as an internal process, PE' s explicit indicators remain an open question whose resolution requires understanding mechanisms at multiple levels. Currently, research on PE' s neural mechanisms during memory reconsolidation is particularly lacking (Cao et al., 2019). It is important to distinguish between PE generation mechanisms and PE action mechanisms. Although substantial research has accumulated on PE neural signals and molecular mechanisms, particularly in reward learning, few studies have examined the neural and molecular mechanisms through which PE initiates memory reconsolidation.

3.1.1 Amygdala Previous research shows that synaptic plasticity changes induced by temporal relationships between CS and US are manifested in the lateral nucleus of the amygdala (LA). The amygdala' s role in PE detection has been demonstrated in both human and non-human mammalian studies of conditioned fear and addiction. The amygdala responds to unexpected temporal changes or unanticipated events. Belova et al. (2007) found that amygdala neurons in mammals respond to unexpectedly presented rewards or negative stimuli during reinforcement learning, but not to expected stimuli. Different amygdala

neuron populations may exist: one responding to surprising stimuli that trigger arousal and attention, and another specifically responding to behavioral valence (reward or punishment) (Belova et al., 2007).

Another study using Fos protein expression (a product of the immediate early gene c-fos closely related to learning) as a dependent variable examined its expression in surprise versus consistent groups after establishing light-tone associations. When surprise occurred (light no longer predicting tone), Fos protein expression increased in the central nucleus of the amygdala (CN), while consistent groups showed increased Fos expression in the basolateral amygdala (BLA). This suggests different amygdala regions may have distinct roles in PE mechanisms (Bucci & Macleod, 2007).

3.1.2 Ventrolateral Periaqueductal Gray (vlPAG) Previous studies indicate the periaqueductal gray (PAG) participates in PE computation, with neural activity corresponding to PE magnitude (Roy et al., 2014). Recent research using signed PE (SPE) examined vlPAG' s role in uncertainty-driven fear memory updating. Studies in rats showed that single neuron activity in vlPAG during fear discrimination aligned with SPE and updated subsequent fear memory. Inhibiting vlPAG attenuated fear to uncertain but not threatening cues, suggesting vlPAG may be a center for SPE computation and has a causal relationship with fear memory updating (Walker et al., 2020).

Regarding vlPAG' s mechanism, researchers propose it likely involves the midbrain VTA. Studies show vlPAG sends excitatory or inhibitory signals to VTA, a key region for motivation and reinforcement processing that receives GABAergic or glutamatergic inputs from vlPAG, generating positive and negative PE signals that trigger behavioral changes (Waung et al., 2019).

3.1.3 Hippocampus Given the hippocampus' s important role in memory retrieval, its potential function in memory reconsolidation has attracted attention. Its involvement has been demonstrated in other memory types such as implicit or explicit associative memory in cognitive tasks (Duncan et al., 2009; Kumaran & Maguire, 2006; Long et al., 2016). The neural signal of negative PE from US omission after CS presentation in conditioned fear is also thought to involve the hippocampus (Spoormaker et al., 2011).

New technologies have advanced neural mechanism research. Optogenetics in rodents can precisely activate or silence specific hippocampal neurons to study their function in fear conditioning. Researchers successfully established false contextual fear memories in mice by optogenetically activating dentate gyrus (DG) and CA1 regions (Ramirez et al., 2013), proving the hippocampus' s role in memory activation. Similarly, silencing hippocampal neurons caused memory retrieval failure that could be rescued by activating neocortical neurons (Cowansage et al., 2014), demonstrating hippocampal-neocortical interactions during retrieval. Researchers hypothesize the hippocampus plays a primary role in processing abstract, higher-order PE, generating expected outcomes in CA3

that are conveyed to CA1 for comparison with sensory input from neocortex, producing match/mismatch signals—the latter being PE (Vinogradova, 2001).

Thus, the hippocampus participates in both memory retrieval and online expectation generation, with hippocampal-neocortical interactions being fundamentally different during retrieval versus expectation updating. Through PE' s action, various neurotransmitters play key roles in determining whether memory retrieval or expectation updating dominates (Barron et al., 2020).

3.1.4 Prefrontal Cortex The prefrontal cortex (PFC) always plays an important role in fear memory processing, with different subregions participating in acquisition, consolidation, and extinction. However, PFC' s role differs significantly between traditional extinction and retrieval-extinction. A human fMRI study using temporal difference (TD) PE found that negative PE signals from expected outcome omission were associated with activation in ventromedial PFC (vmPFC), dorsolateral PFC (dlPFC), and left orbitofrontal cortex (LOFC) (Spoormaker et al., 2011). In the RE paradigm, our recent fMRI study on fear extinction in humans showed that using PE during retrieval created significant differences in brain activation patterns during subsequent extinction compared to extinction without PE retrieval, specifically reducing activation in inferior temporal gyrus (IT) and dlPFC, and decreasing functional connectivity between dlPFC-ACC (anterior cingulate cortex) and IT-dlPFC (Junjiao et al., 2019). Another human fear memory retrieval-extinction fMRI study without explicit PE manipulation found that the mechanism distinguishing retrieval-extinction from standard extinction was significantly reduced vmPFC involvement (Schiller et al., 2013). That study used single CS without US as retrieval trials, which could be considered implicit PE due to US omission. Thus, PFC and its subregions play important roles in PE-initiated reconsolidation, though human studies remain limited and specific functions await further investigation.

Based on our human behavioral and neuroimaging research in this field and previous findings, we propose a model of PFC and related brain regions' role in PE-driven memory updating [Figure 3: see original paper].

Figure 3. Role of prefrontal cortex and related brain regions in PE-driven memory updating

Note: At the retrieval boundary condition, novelty in CS-US relationship during retrieval is necessary for entering reconsolidation, while CS novelty alone triggers new extinction learning. Retrieval-extinction and traditional extinction show significant differences in activated brain regions, particularly in PFC, especially dlPFC. Retrieval-extinction significantly reduces dlPFC and IT activation, and decreases functional connectivity between dlPFC and ACC. In contrast, traditional extinction shows significant dlPFC activation.

3.2 Neuromodulation in Circuits

Various neurotransmitters are closely associated with PE function. Error-based learning is importantly modulated by specific neurotransmitters, most notably dopamine. Dopamine neurons respond directly to unexpected outcomes and are thought to be involved in memory updating and destabilization. Midbrain dopamine increases during positive PE, remains unchanged without PE, and decreases during negative PE. Unlike dopamine depletion patterns in midbrain VTA, PE responses in BLA are valence-independent (Schultz, 2016).

Recent research also reveals that midbrain dopamine responds not only to unexpected reward presentation but also to unexpected punishment omission, providing important insights for fear memory elimination and exposure therapy improvement (Hernandez et al., 2018). Papalini and Beckers et al. (2020) proposed a three-step model of PE and midbrain dopamine effects on fear extinction, from molecular mechanisms to clinical application: (1) Omission of negative US triggers PE that initiates dopamine firing in the nucleus accumbens (NAcc)/VTA; (2) This DA signal drives new safety memory formation, possibly involving D2 receptor (D2R)-mediated dopaminergic signaling. DA transmission to PFC updates threat expectations in vmPFC and may implement extinction memory retrieval in lateral PFC (lPFC), potentially involving D1 receptor (D1R)-mediated dopaminergic signaling in PFC and hippocampus; (3) At the behavioral level, further investigation of dopaminergic interventions is needed to promote extinction and maintain exposure therapy effects (Papalini et al., 2020).

Additionally, glutamate (Glu) neurotransmission' s role in PE-initiated reconsolidation warrants attention. As early as 2006, researchers proposed that PE-induced DA and Glu might be co-released in addiction memory, termed Glu-DA co-transmission (Lapish et al., 2006), suggesting these neurotransmitters may co-modulate PE function, though research remains limited. On the other hand, glutamatergic mechanisms in memory destabilization have been gradually revealed. Molecular studies show that memory destabilization depends on the content of NMDA receptor subunit GluN2B, particularly the GluN2B/GluN2A ratio, which determines whether fear memory can enter an unstable state (Milton et al., 2013; Shipton & Paulsen, 2014). A recent study found that increasing GluN2B levels in BLA under strong fear memory conditions can modify memories resistant to reconsolidation intervention due to their strength (Solis et al., 2019). Since the molecular mechanisms through which PE initiates memory destabilization remain unknown, such studies suggest its neural circuits are also closely related to glutamatergic systems.

4.1 Quantitative Research Based on Prediction Error Computational Models

As discussed, a major advance in recent PE and memory reconsolidation research has been the shift from qualitative to quantitative investigation—from examining

PE presence/absence to different PE magnitudes—making research increasingly refined. PE quantification has thus become a primary issue in the field. We propose that quantitative research based on PE computational models can draw from three areas:

First, **studies quantifying PE and memory updating in other memory types**. Pine et al. (2018) first used a "recall-based choice-confidence rating-feedback" paradigm in declarative memory to explore how PE magnitude in striatal regions drives memory updating at behavioral and neuroimaging levels. We suggest that PE manipulation and quantification in fear memory can borrow from successful paradigms in declarative memory to examine relationships between PE magnitude and fear memory updating across different levels.

Second, **psychiatric mechanism research in animal models**. A recent study operationalized hallucinations—a primary symptom of schizophrenia—as high-confidence false alarms using signal detection theory from psychophysics. This study used a belief update formula containing PE to computationally model PE magnitude, learning ability, and belief updating, validating these in animals and humans (Schmack et al., 2021). Such PE calculation methods in psychiatric models also offer important insights for quantifying PE in human fear memory.

Third, **new theoretical models of prediction and learning**. Recent learning models have emerged that more flexibly explain new phenomena than traditional models. Osan et al. (2011) proposed a neural network model of memory reconsolidation and extinction based on information mismatch, suggesting Hebbian learning strengthens synapses while PE causes synaptic weakening. The balance between them forms synaptic weights. When sensory input slightly differs from original memory traces, the original memory updates synaptic weight configurations. When input significantly differs from memory engrams, new memory traces form in opposition. The reconsolidation process is triggered by differences between network input signals (similar but not identical to original patterns) and recalled fear memories, characterized by prediction error (Osan et al., 2011). This model has received empirical support and can predict behavioral outcomes, enabling experimental measurement of boundary conditions (Radiske et al., 2017).

Future research should draw from these three advances to establish more precise quantitative relationships between PE and fear memory elimination within the reconsolidation framework, deepening fundamental understanding of its role.

4.2 Exploring Interactions Between Prediction Error and Other Boundary Conditions

As a retrieval boundary condition, PE interacts with memory properties [Figure 2: see original paper], yet current research rarely combines these factors directly, especially in human studies. Investigating interactions between retrieval boundaries and memory properties (e.g., strength) represents an important future direction, crucial for deepening understanding of PE mechanisms and promoting

clinical application of this paradigm.

This also highlights the need to explore viable memory reconsolidation indicators. A persistent problem in reconsolidation research is the lack of explicit indicators—no study has truly identified a marker for memory entry into reconsolidation. Since PE is a necessary but insufficient condition for destabilization, it cannot serve as such an indicator. Clinically, it is difficult to determine whether patients experience PE and its magnitude. This lack of reconsolidation indicators also leads some researchers to question the existence of this phase (Miller & Matzel, 2006), arguing that reconsolidation is not the only explanation for experimental results and that traditional memory encoding theories could explain the findings. Based on substantial neurobiological evidence for reconsolidation across memory types, we believe the reconsolidation phase objectively exists, but explicit indicators—including PE-related markers—must continue to be explored.

4.3 Examining the Role of Different Salience Types in Memory Reconsolidation

As previously summarized, different salience types are distinguished by memory processing depth [Figure 1: see original paper]. PE as salient information focuses on the relationship between behavior and outcomes, while another type of novelty—stimulus salience—has received increasing attention. The differential roles of stimulus novelty versus stimulus-outcome PE in retrieval interventions, their influencing factors, and their corresponding neural pathways should be investigated.

Previous research shows that while physical stimulus changes enhance dopamine, they are insufficient to trigger dopamine release. Stimulus changes must be paired with reinforcers to be motivating (Schultz, 2016). However, stimulus changes engage attention systems, increasing norepinephrine (NE) and triggering orienting responses, thereby influencing retrieval interventions (Li et al., 2017). The combined effect of stimulus change superimposed on PE versus PE alone warrants further investigation. At the neural mechanism level, research shows different striatal subregions process different salience types: ventral striatum (VS) primarily processes reward expectation, while the tail of striatum (TS) primarily processes perceptual expectation (Schmack et al., 2021). These findings demonstrate that different salience types likely have distinct mechanisms affecting different behavioral aspects.

Future research should therefore focus on how different salience types influence human fear memory retrieval interventions and their mechanisms.

4.4 Individual Differences in Prediction Error Effects

The pattern of PE-driven fear updating may show individual differences, yet such research is particularly lacking. Differences between clinical and non-clinical populations are important and include:

First, **PE signal dysfunction in abnormal populations**. Research shows that while clinically depressed subjects do not differ from non-depressed subjects in reinforcement learning, they exhibit larger errors in valence evaluation during contextual classification tasks. No memory differences exist between groups, but depressed subjects show larger negative PE, which enhances episodic memory more than positive PE. Non-depressed subjects show larger positive PE (Rouhani & Niv, 2019). Yaple and Yu et al. (2021) also found abnormal PE processing in schizophrenia and depression, with clinical subjects showing consistent activity in posterior cingulate regions during PE processing—an effect absent in healthy controls.

Second, **reward processing dysfunction in reward learning**. Anhedonia is a core feature of depression, and reward processing dysfunction in reward learning is part of depression' s pathogenesis, including dysfunctional expectation regulation. Depressed patients show significantly smaller reward PE magnitude, with positive valence having less reinforcing effect. Researchers suggest PE and its regulatory systems may participate in psychiatric disorder development (Beckers & Kindt, 2017). Improving understanding of reward processing dysfunction in depression could enhance diagnosis and treatment (Admon & Pizzagalli, 2015), necessitating future PE function studies in clinical populations including depression.

Third, **stress condition differences**. Clinical populations differ from healthy populations in stress conditions accompanying psychiatric disorders. In PTSD, memory retrieval is accompanied by elevated stress, placing subjects in high-stress states. Stress constrains memory destabilization boundaries, limiting PE' s effectiveness in activating original memories. In both animal models and human studies of depression, stress is a common trigger for initial depressive episodes, with chronic stress suppressing hippocampal neurogenesis, inhibiting mesolimbic dopamine neurons, and sensitizing amygdala responses to negative information. These mechanisms may explain memory disruption for positive material and enhancement for negative material in depressed adults (Dillon & Pizzagalli, 2018).

Future research should therefore emphasize comparative studies between clinical and healthy populations, using psychiatric disorder-related memory models and considering acute and chronic factors like stress to facilitate better clinical translation of basic research findings.

4.5 Using Multidisciplinary Approaches to Explore Neural and Molecular Mechanisms of PE in Memory Updating

The fundamental question remains: How does prediction error open the memory reconsolidation window? What are the underlying neurophysiological mechanisms? This can be decomposed into two issues: PE' s neural signals per se, and PE' s action mechanisms. For the latter, no studies have yet identified differences in neural processes between single CS retrieval that does versus does

not trigger reconsolidation, or which time interval PE primarily affects. Since identifying reconsolidation indicators is crucial for paradigm application, clarifying PE' s mechanisms is critical. Future research should continue exploring PE' s mechanisms for opening reconsolidation at multiple levels.

At the molecular level, multiple classical neurotransmitters play important roles in memory destabilization, including dopamine, norepinephrine (NE), acetylcholine (Ach), serotonin (5-HT), GABA, and brain-derived neurotrophic factor (BDNF) (see Wideman et al., 2018 for review). How PE relates to these neurotransmitters and how their interactions with other boundary conditions influence memory updating are important questions for further exploration.

Memory research at various levels is not fragmented but complementary and mutually explanatory. Biotechnology enables investigation of factors inaccessible at the behavioral level. For example, optogenetics can specifically express light-sensitive ion channels in particular neuron types, altering their firing patterns via light stimulation to study function. Currently, many aspects of PE' s internal mechanisms remain unknown, with most PE research still based on the model of mismatch between external and internal representations driving memory updating (Fernandez, Boccia, et al., 2016). Researchers assert that future non-invasive brain stimulation (NIBS), including transcranial magnetic stimulation (TMS), transcranial electric stimulation (tES), and optogenetics, will enable more specific manipulation of fear memory neurobiology, identifying therapeutic targets for pathological fear resulting from trauma, stress, and anxiety (Borgomaneri et al., 2021). We believe multidisciplinary integration and multiple technical approaches are inevitable trends for future basic research and clinical translation, including studies on PE updating of fear memories.

In conclusion, as a classic theme in learning and decision-making, prediction error has been revitalized in the emerging memory reconsolidation theory. Its significance lies in offering hope for fundamentally updating or eliminating fear memories characteristic of various clinical mental disorders. As a crucial variable for opening the memory reconsolidation window, prediction error will undoubtedly continue to be intensively studied and play a greater role in clinical applications.

References

Admon, R., & Pizzagalli, D. A. (2015). Dysfunctional reward processing in depression. *Current Opinion in Psychology*, 4, 114–118. doi:10.1016/j.copsyc.2014.12.011

Alberini, C. M., Milekic, M. H., & Tronel, S. (2006). Mechanisms of memory stabilization and de-stabilization. *Cellular & Molecular Life Sciences*, 63(9), 999–1008. doi:10.1007/s00018-006-6002-9

Amadi, U., Lim, S. H., Liu, E., Baratta, M. V., & Goosens, K. A. (2017). Hippocampal processing of ambiguity enhances fear memory. *Psychological Science*,

28(2), 168–180. doi:10.1177/0956797616674055

Barron, H. C., Auksztulewicz, R., & Friston, K. (2020). Prediction and memory: A predictive coding account. *Progress in Neurobiology*, 192, 101821. doi:10.1016/j.pneurobio.2020.101821

Beckers, T., & Kindt, M. (2017). Memory reconsolidation interference as an emerging treatment for emotional disorders: Strengths, limitations, challenges, and opportunities. *Annual Review of Clinical Psychology*, 13, 99–121. doi:10.1146/annurev-clinpsy-032816-045209

Belova, M. A., Paton, J. J., Morrison, S. E., & Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron*, 55(6), 970–984. doi:10.1016/j.neuron.2007.08.004

Borgomaneri, S., Battaglia, S., Sciamanna, G., Tortora, F., & Laricchiuta, D. (2021). Memories are not written in stone: Re-writing fear memories by means of non-invasive brain stimulation and optogenetic manipulations. *Neuroscience and Biobehavioral Reviews*, 127, 334–352. doi:10.1016/j.neubiorev.2021.04.036

Bucci, D. J., & Macleod, J. E. (2007). Changes in neural activity associated with a surprising change in the predictive validity of a conditioned stimulus. *European Journal of Neuroscience*, 26(9), 2669–2676. doi:10.1111/j.1460-9568.2007.05902.x

Cao, Y., Li, J., Chen, W., Yang, Y., Hu, Y., & Zheng, X. (2019). The retrieval-extinction paradigm and its neural mechanisms in fear memory extinction. *Advances in Psychological Science*, 27(2), 268–278. doi:10.3724/sp.J.1042.2019.00268

Chen, W., Li, J., Cao, Y., Yang, Y., Hu, Y., & Zheng, X. (2018). The role of prediction error in compound fear memory retrieval-extinction. *Acta Psychologica Sinica*, 50(7), 739–749. doi:10.3724/sp.J.1041.2018.00739

Chen, W., Li, J., Lin, X., Zhang, X., & Zheng, X. (2020). Behavioral intervention in emotional memory reconsolidation: From laboratory to clinical translation. *Advances in Psychological Science*, 28(2), 240–252. doi:10.3724/sp.J.1042.2020.00240

Chen, W., Li, J., Xu, L., Zhao, S., Fan, M., & Zheng, X. (2020). Destabilizing different strengths of fear memories requires different degrees of prediction error during retrieval. *Frontiers in Behavioral Neuroscience*, 14, 598924. doi:10.3389/fnbeh.2020.598924

Chen, W., Li, J., Zhang, X., Dong, Y., Shi, P., Luo, P., & Zheng, X. (2021). Retrieval-extinction as a reconsolidation-based treatment for emotional disorders: Evidence from an extinction retention test shortly after intervention. *Behaviour Research and Therapy*, 139, 103831. doi:10.1016/j.brat.2021.103831

Colombo, M. (2014). Deep and beautiful. The reward prediction error hypothesis of dopamine. *Studies in History and Philosophy of Biological and Biomedical*

Sciences, 45, 57–67. doi:10.1016/j.shpsc.2013.10.006

Cowansage, K. K., Shuman, T., Dillingham, B. C., Chang, A., Golshani, P., & Mayford, M. (2014). Direct reactivation of a coherent neocortical memory of context. *Neuron*, 84(2), 432–441. doi:10.1016/j.neuron.2014.09.022

Das, R. K., Gale, G., Hennessy, V., & Kamboj, S. K. (2018). A prediction error-driven retrieval procedure for destabilizing and rewriting maladaptive reward memories in hazardous drinkers. *Journal of Visualized Experiments*, (131). doi:10.3791/56097

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

Diederen, K. M. J., & Fletcher, P. C. (2021). Dopamine, prediction error and beyond. *The Neuroscientist*, 27(1), 30–46. doi:10.1177/1073858420907591

Dillon, D. G., & Pizzagalli, D. A. (2018). Mechanisms of memory disruption in depression. *Trends in Neurosciences*, 41(3), 137–149. doi:10.1016/j.tins.2017.12.006

Duvarci, S., & Nader, K. (2004). Characterization of fear memory reconsolidation. *The Journal of Neuroscience*, 24(42), 9269–9275. doi:10.1523/JNEUROSCI.2971-04.2004

Elsey, J. W. B., & Kindt, M. (2017). Tackling maladaptive memories through reconsolidation: From neural to clinical science. *Neurobiology of Learning and Memory*, 142(Pt A), 108–117. doi:10.1016/j.nlm.2017.03.007

Ergo, K., De Loof, E., & Verguts, T. (2020). Reward prediction error and declarative memory. *Trends in Cognitive Sciences*, 24(5), 388–397. doi:10.1016/j.tics.2020.02.009

Faliagkas, L., Rao-Ruiz, P., & Kindt, M. (2018). Emotional memory expression is misleading: Delineating transitions between memory processes. *Current Opinion in Behavioral Sciences*, 19, 116–122. doi:10.1016/j.cobeha.2017.12.018

Fernandez, R. S., Bavassi, L., Forcato, C., & Pedreira, M. E. (2016). The dynamic nature of the reconsolidation process and its boundary conditions: Evidence based on human tests. *Neurobiology of Learning and Memory*, 130, 202–212. doi:10.1016/j.nlm.2016.03.001

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

Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V. A., Pedreira, M. E., & Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learning & Memory*, 14(4), 295–303. doi:10.1101/lm.486107

Furlong, T. M., Cole, S., Hamlin, A. S., & McNally, G. P. (2010). The role of prefrontal cortex in predictive fear learning. *Behavioral Neuroscience*, 124(5), 574–586. doi:10.1037/a0020739

Gershman, S. J., Monfils, M. H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *Elife*, 6. doi:10.7554/eLife.23763

Hernandez, X. I., Vogel, P., Betz, S., Kalisch, R., Sigurdsson, T., & Duvarci, S. (2018). Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. *Elife*, 7. doi:10.7554/eLife.38818

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

Kim, H. F., Ghazizadeh, A., & Hikosaka, O. (2014). Separate groups of dopamine neurons innervate caudate head and tail encoding flexible and stable value memories. *Frontiers in Neuroanatomy*, 8, 120. doi:10.3389/fnana.2014.00120

Kumaran, D., & Maguire, E. A. (2006). The dynamics of hippocampal activation during encoding of overlapping sequences. *Neuron*, 49(4), 617–629. doi:10.1016/j.neuron.2005.12.024

Lapish, C. C., Seamans, J. K., & Chandler, L. J. (2006). Glutamate-dopamine cotransmission and reward processing in addiction. *Alcoholism: Clinical and Experimental Research*, 30(9), 1451–1465. doi:10.1111/j.1530-0277.2006.00176.x

Lee, J. L. (2009). Reconsolidation: Maintaining memory relevance. *Trends in Neurosciences*, 32(8), 413–420. doi:10.1016/j.tins.2009.05.002

Lee, J. L., Milton, A. L., & Everitt, B. J. (2006). Reconsolidation and extinction of conditioned fear: Inhibition and potentiation. *The Journal of Neuroscience*, 26(39), 10051–10056. doi:10.1523/JNEUROSCI.2466-06.2006

Li, J., Chen, W., Caoyang, J., Wu, W., Jie, J., Xu, L., & Zheng, X. (2017). Moderate partially reduplicated conditioned stimuli as retrieval cue can increase effect on preventing relapse of fear to compound stimuli. *Frontiers in Human Neuroscience*, 11, 575. doi:10.3389/fnhum.2017.00575

Li, J., Chen, W., Hu, Y., Cao, Y., & Zheng, X. (2021). Effects of prediction error and acute stress on retrieval-extinction of different strength fear memories. *Acta Psychologica Sinica*, 53(6), 587–598. doi:https://doi.org/10.3724/SP.J.1041.2021.00587

Long, N. M., Lee, H., & Kuhl, B. A. (2016). Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. *The Journal of Neuroscience*, 36(50), 12677–12687. doi:10.1523/JNEUROSCI.1850-16.2016

McNally, G. P., Johansen, J. P., & Blair, H. T. (2011). Placing prediction into the fear circuit. *Trends in Neurosciences*, 34(6), 283–292. doi:10.1016/j.tins.2011.03.005

Miller, R. R., & Matzel, L. D. (2006). Retrieval failure versus memory loss in experimental amnesia: Definitions and processes. *Learning & Memory*, 13(5), 491–497. doi:10.1101/lm.241006

Milton, A. L., Merlo, E., Ratano, P., Gregory, B. L., Dumbreck, J. K., & Everitt, B. J. (2013). Double dissociation of the requirement for glun2b- and glun2a-containing nmda receptors in the destabilization and restabilization of a reconsolidating memory. *The Journal of Neuroscience*, 33(3), 1109–1115. doi:10.1523/JNEUROSCI.3273-12.2013

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

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

Osan, R., Tort, A. B., & Amaral, O. B. (2011). A mismatch-based model for memory reconsolidation and extinction in attractor networks. *PLoS One*, 6(8), e23113. doi:10.1371/journal.pone.0023113

Papalini, S., Beckers, T., & Vervliet, B. (2020). Dopamine: From prediction error to psychotherapy. *Translational Psychiatry*, 10(1), 164. doi:10.1038/s41398-020-0814-x

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

Pine, A., Sadeh, N., Ben-Yakov, A., Dudai, Y., & Mendelsohn, A. (2018). Knowledge acquisition is governed by striatal prediction errors. *Nature Communications*, 9(1), 1673. doi:10.1038/s41467-018-03992-5

Raczka, K. A., Mechias, M. L., Gartmann, N., Reif, A., Deckert, J., Pessiglione, M., & Kalisch, R. (2011). Empirical support for an involvement of the mesostriatal dopamine system in human fear extinction. *Translational Psychiatry*, 1, e12. doi:10.1038/tp.2011.10

Radiske, A., Gonzalez, M. C., Conde-Ocazionez, S. A., Feitosa, A., Kohler, C. A., Bevilaqua, L. R., & Cammarota, M. (2017). Prior learning of relevant nonaversive information is a boundary condition for avoidance memory reconsolidation in the rat hippocampus. *The Journal of Neuroscience*, 37(40), 9675–9685. doi:10.1523/JNEUROSCI.1372-17.2017

Ramirez, S., Liu, X., Lin, P. A., Suh, J., Pignatelli, M., Redondo, R. L., . . . Tonegawa, S. (2013). Creating a false memory in the hippocampus. *Science*,

341(6144), 387-391. Retrieved from http://science.sciencemag.org/content/341/6144/387.abstract

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non reinforcement. *Classical Conditioning II: Current Research and Theory*, 64-69.

Rouhani, N., & Niv, Y. (2019). Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, 236(8), 2425-2435. doi:10.1007/s00213-019-05322-z

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, 17(11), 1607-1612. doi:10.1038/nn.3832

Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M. H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50), 20040-20045. doi:10.1073/pnas.1320322110

Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., Ledoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, 463(7277), 49-53. doi:10.1038/nature08637

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, 372(6537). doi:10.1126/science.abf4740

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, 1, 199-207. doi:https://doi.org/10.1038/35044563

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nature Reviews Neuroscience*, 17(3), 183-195. doi:10.1038/nrn.2015.26

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593-1599.

Sevenster, D., Beckers, T., & Kindt, M. (2012). Retrieval per se is not sufficient to trigger reconsolidation of human fear memory. *Neurobiology of Learning and Memory*, 97(3), 338-345. doi:10.1016/j.nlm.2012.01.009

Sevenster, D., Beckers, T., & Kindt, M. (2013). Prediction error governs pharmacologically induced amnesia for learned fear. *Science*, 339, 830-833. doi:10.1126/science.1231357

Sevenster, D., Beckers, T., & Kindt, M. (2014). Prediction error demarcates the transition from retrieval, to reconsolidation, to new learning. *Learning & Memory*, 21(11), 580-584. doi:10.1101/lm.035493.114

Sevenster, D., Visser, R. M., & D' Hooge, R. (2018). A translational perspective on neural circuits of fear extinction: Current promises and challenges. *Neurobiology of Learning and Memory*, 155, 113-126. doi:10.1016/j.nlm.2018.07.002

Shipton, O. A., & Paulsen, O. (2014). Glun2a and glun2b subunit-containing nmda receptors in hippocampal plasticity. *Philosophical Transactions of the Royal Society B*, 369(1633), 20130163. doi:10.1098/rstb.2013.0163

Sinclair, A. H., & Barense, M. D. (2018). Surprise and destabilize: Prediction error influences episodic memory reconsolidation. *Learning and Memory*, 25, 369–380. doi:10.1101/lm.046912.117

Sinclair, A. H., & Barense, M. D. (2019). Prediction error and memory reactivation: How incomplete reminders drive reconsolidation. *Trends in Neurosciences*, 42(10), 727–739. doi:10.1016/j.tins.2019.08.007

Solis, C. A. d., Gonzalez, C. U., Galdamez, M. A., Perish, J. M., Woodard, S. W., Salinas, C. E., . . . Ploski, J. E. (2019). Increasing synaptic glun2b levels within the basal and lateral amygdala enables the modification of strong reconsolidation resistant fear memories. *bioRxiv*, 537142. doi:10.1101/537142

Spoormaker, V. I., Andrade, K. C., Schroter, M. S., Sturm, A., Goya-Maldonado, R., Samann, P. G., & Czisch, M. (2011). The neural correlates of negative prediction error signaling in human fear conditioning. *Neuroimage*, 54(3), 2250–2256. doi:10.1016/j.neuroimage.2010.09.042

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, 20(4), 581–589. doi:10.1038/nn.4520

Thiele, M., Yuen, K. S. L., Gerlicher, A. V. M., & Kalisch, R. (2021). A ventral striatal prediction error signal in human extinction learning. *NeuroImage*, 227, 117709. doi:10.1016/j.neuroimage.2020.117709

Vinogradova, O. S. (2001). Hippocampus as comparator: Role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus*, 11(5), 578–598. doi:10.1002/hipo.1073

Walker, R. A., Wright, K. M., Jhou, T. C., & McDannald, M. A. (2020). The ventrolateral periaqueductal grey updates fear via positive prediction error. *European Journal of Neuroscience*, 51(3), 866–880. doi:10.1111/ejn.14536

Waung, M. W., Margolis, E. B., Charbit, A. R., & Fields, H. L. (2019). A midbrain circuit that mediates headache aversiveness in rats. *Cell Reports*, 28(11), 2734–2745. doi:10.1016/j.celrep.2019.08.009

Wideman, C. E., Jardine, K. H., & Winters, B. D. (2018). Involvement of classical neurotransmitter systems in memory reconsolidation: Focus on destabilization. *Neurobiology of Learning and Memory*, 156, 68–79. doi:10.1016/j.nlm.2018.11.001

Yaple, Z. A., Tolomeo, S., & Yu, R. (2021). Abnormal prediction error processing in schizophrenia and depression. *Human Brain Mapping*. doi:10.1002/hbm.25453

Zuccolo, P. F., & Hunziker, M. H. L. (2019). A review of boundary conditions and variables involved in the prevention of return of fear after post-retrieval extinction. *Behavioural Processes*, 162, 39–54. doi:10.1016/j.beproc.2019.01.011

Note: Figure translations are in progress. See original paper for figures.