Computational Investigations on Uncertainty-Dependent Extinction of Fear Memory

Yuzhe Li

Table of contents

Abstract

Uncertainty of fear conditioning is crucial for the acquisition and extinction of fear memory. Fear memory acquired through partial pairings of a conditioned stimulus (CS) and an unconditioned stimulus (US) is more resistant to extinction than that acquired through full pairings; this effect is known as the partial reinforcement extinction effect (PREE). Although the PREE has been explained by psychological theories, the neural mechanisms underlying the PREE remain largely unclear. Here, we developed a neural circuit model based on three distinct types of neurons (fear, persistent and extinction neurons) in the amygdala and medial prefrontal cortex (mPFC). In the model, the fear, persistent and extinction neurons encode predictions of net severity, of unconditioned stimulus (US) intensity, and of net safety, respectively. Our simulation successfully reproduces the PREE. We revealed that unpredictability of the US during extinction was represented by the combined responses of the three types of neurons, which are critical for the PREE. In addition, we extended the model to include amygdala subregions and the mPFC to address a recent finding that the infralimbic cortex of the mPFC (IL) is required for consolidating extinction memory but not for memory retrieval. Furthermore, model simulations led us to propose a novel procedure to enhance extinction learning through re-conditioning with a stronger US; strengthened fear memory up-regulates the extinction neuron which in turn, further inhibits the fear neuron during re-extinction. Thus, our models increased the understanding of the functional roles of the amygdala and IL in the processing of uncertainty in fear conditioning and extinction.

Chapter 1
Introduction

How do we learn? We learn from experiences, which form memories. Then over time much of these memories will be forgotten.

The process of forgetting is involved with all memories. It is the view of many theories and models that forgetting is not simply a process of erasing memories, but another type of learning process that inhibits old memories [1]. The rate at which memories are forgotten varies greatly depending on the nature of the memories and is affected by a number of factors. There is a well-known proverb that "Soon learn, soon forgotten", suggesting the hypothesis that one factor affecting how quickly a memory is forgotten is the learning process itself.

In support of this hypothesis, we can point to a phenomenon in classical conditioning, known as the "partial reinforcement learning affect" (PREE) [2–4]. Associative fear memory acquired through partial reinforcement conditioning, where a conditional stimulus(CS) (such as a sound tone) which is probabilistically paired with an unconditional stimulus(US) (such as an electric shock) is more resistant to extinction (forgetting of conditioned fear memory of CS) than those that are acquired through full reinforcement conditioning, in which the CSs are always followed by a US. Therefore, by comparing the rate of memory extinction between the partial reinforcement conditioning and full reinforcement regimes, we can deduce that the uncertainty of CS-US pairing in the partial reinforcement is a key factor leading the resistance to extinction of the conditioned fear memory, and causing the PREE.

Uncertainty is a ubiquitous and unavoidable factor in the learning process. Behavioral studies on partial reinforcement fear conditioning have been conducted since the 1970s, however for simplicity most of the neurophysiology research on fear conditioning up to now either avoids the uncertainty of CS-US pairing in the conditioning paradigms, or has ignored the effect of partial reinforcement on neural and behavioral performance. As a result, how the brain processes the uncertainty in fear conditioning and generates PREE largely remains unclear.

This thesis presents a computational model for explaining PREE, and reveals how the uncertainty in associative fear learning affects the extinction process in terms of the interactions of neural activities. I also extended this model to adapt to more complicated conditioning paradigms. The extended model captures the behavior of the additional conditioning paradigms and also suggests a method to accomplish full extinction of the residual fear memory after PREE.

1.1 Fear conditioning and extinction

The focus of my model is on *fear conditioning* which is a form of classical conditioning. Classical conditioning is a type of associative learning famously pioneered by Ivan Pavlov

in the 1920s [5], in which a subject learns the association between a neutral stimulus and an reward/punishment stimulus. Fear conditioning is a subset of classical conditioning and focuses on the neutral stimulus and punishment stimulus connection.

The process of fear conditioning is outlined here: A subject is presented with a neutral stimulus (or conditional stimulus(CS)) followed by an aversive stimulus (or unconditional stimulus(US)). This is repeated a number of times and the repetition results in the subject learning the association between the CS-US pair. The result of this is that the subject will express a fear response to the previously neutral stimulus. This part is referred to as the acquisition phase (or conditioning phase) (see left panel in Fig. 1.1A). In general we can identify the fear response through changes in behavior or neural activity of the subject. After learning the association between the CS and US, and acquiring a fear response to the CS, repeated presence of the CS without the US leads a decrease of the fear response to the CS, this is referred to as the extinction phase (the right panel in Fig. 1.1A).

1.2 Extinction: Not unlearning but new learning

Extinction is not simply "erasing" or "unlearning" the CS-US association, but a newly learned inhibition of a conditioned fear memory. Facts in support of this view are that:

- extinguished fear responses can spontaneously recover after a delay, known as *spontaneous recovery* [5–7];
- extinguished fear responses can reappear after presenting the US even without the CS. Sudden application of the US can then provoke the fear response again, this is known as *reinstatement*" [8, 9];
- extinguished fear responses in a context can recover if tested in a context different from the one where extinction occurred, known as *renewal* [10, 11].

A key conclusion is that fear extinction creates a new "safe", inhibiting memory that co-exists with the original fear memory. This conclusion has been very important to both theories of classical conditioning and to its applications in behavioral therapy.

1.3 Partial reinforcement extinction effect

One of the first reports of the partial reinforcement effect is from Humphrey et al. [12] 's research on human eyeblink conditioning, in which a light signal is followed by an air puff to the cornea. In his research, the subjects were divided into three groups: group I and group

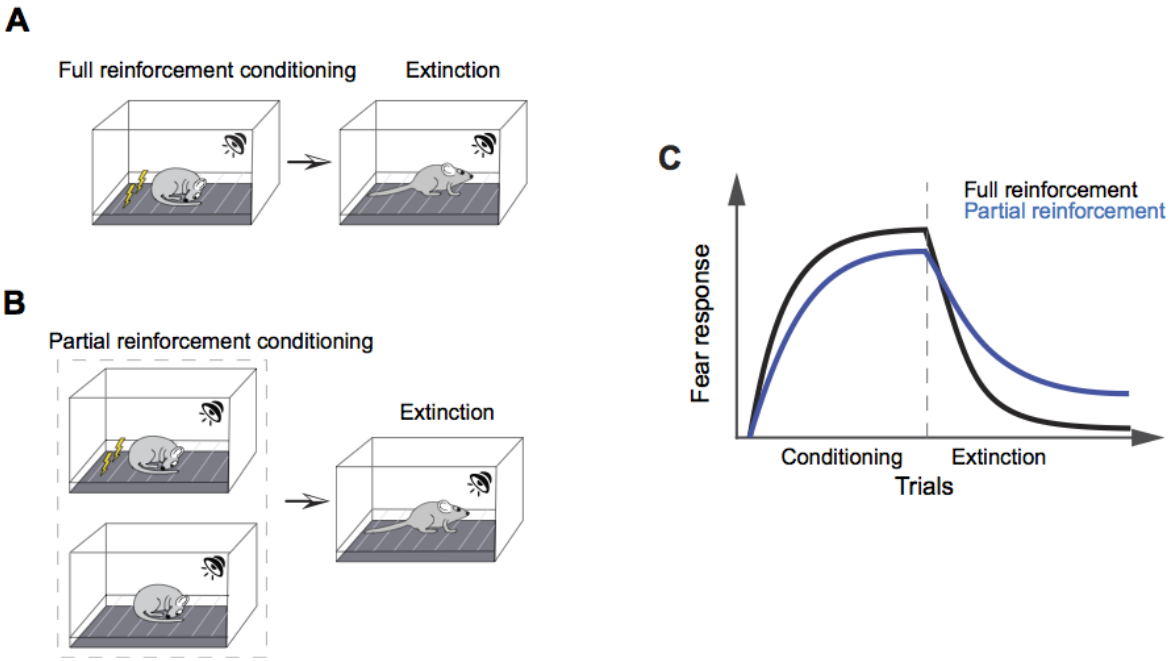

Fig. 1.1 Partial reinforcement extinction effect (PREE). (A, B) During fear conditioning, a CS, e.g., a tone, is fully (in the full reinforcement schedule (A)) or partially (in the partial reinforcement schedule (B)) followed by a US, e.g., electric foot shock (left panels). The fear memory acquired in the fear conditioning can be extinguished by the extinction training, during which the CSs are presented without the US (right panels). (C) Conditional responses to the CS, which are usually measured as the degree of behavioral freezing responses, are depicted during fear conditioning and extinction. The conditional fear memory, which can be measured as the conditional response to the CS, acquired during a partial reinforcement conditioning with $P(US|CS) < 1$ exhibits a resistance to extinction (PREE) (blue line), comparing with that acquired during a partial reinforcement conditioning with $P(US|CS) = 1$ (black line).

III were given full reinforcement, i.e., all light signals were followed by an air puff to the cornea, while group II was given 50% partial reinforcement, i.e., only 50% of the trials in conditioning phase were reinforced by the air puff; the number of the CS-US pairs in group I is twice as much as that in group II, which is the same as that in group III.

His result shows that there was no big difference in the conditioning responses acquired in the conditioning trials of the three groups, but group II responded at a significantly higher level after extinction trials than the other two groups.

This result indicates that in certain conditioning regimes *less is more*: fewer reinforcements could result in an apparently more durable association.

Early research on PREE was mainly focused on examining the behavioral effects of different patterns of reinforced (CS-US) and non-reinforced (CS-only) trials . During their

studies, it was found that behaviors that are partially reinforced (Fig. 1.1B) are more resistant to extinction than those reinforced every time (full reinforcement) (Fig. 1.1A) (see Fig. 1.1C). This is found to be true regardless of the schedules used during the conditioning phase.

1.4 Neurophysiology background

The neural substrates associated with fear conditioning and extinction are the amygdala and the medial prefrontal cortex (mPFC), respectively . The amygdala is a major region for the acquisition and expression of fear memory [16–18]. In contrast, the mPFC plays an important role in the extinction of fear memory [19–21]. Although both the amygdala and mPFC function during partial reinforcement fear conditioning [22–27], their roles in the PREE have seldom been examined [28]. Thus, the connection between the amygdala and mPFC during PREE remains elusive.

1.4.1 Amygdala

The amygdalae are two regions of highly interconnected nuclei positioned in the temporal lobes of the brain. There is one located in each hemisphere. The amygdalae are associated with memory, emotional decision making, and fear response. Each amygdala is an almond shaped assemblage of nuclei complexes, including the basolateral complex(BLA), which consists of the lateral and the basal nuclei (LA and BA); the central nucleus (CEA), which consists of the lateral and medial components (CEI and CEM); and the intercalated cell masses (ITCs) [29]. These three components of the amygdala play a crucial role in processing fear stimuli and responses, as well their inhibition.

The role each region plays in fear response is as follows: The LA receives sensory information, including the CS and the US, from various sources such as thalamic and cortical inputs. These signals are then passed to the CEA, which sends appropriate output signals to other parts of the nervous system as a fear response. The LA also sends signals to the ITCs, which in turn inhibits the fear response.

1.4.2 Medial Prefrontal Cortex

The medial prefrontal cortex (mPFC) is a region in the frontal lobe of the brain. It is associated with decision making and memories, and is involved in the fear conditioning and extinction circuits. In humans, the neural circuits underlying fear conditions and extinction in mPFC encompass the dorsal anterior cingulate cortex (dACC) and the ventromedial prefrontal

cortex(vMPFC), which have been proposed to regulate the expression and suppression of fear, respectively [30].

The mPFC in rodents is split into three subdivisions: the anterior cingulate (ACC), the prelimbic (PL) and the infralimbic cortexes (IL) [31]. The PL and the IL are involved in the fear conditioning and extinction. They both receive excitatory inputs from BLA, hippocampus and contralateral mPFC, and their output projects to the BLA and ITCs. The PL is critical for expression of fear-related behavior, whereas, the IL is important for the suppression of fear.

Common view of the IL function

The IL is widely considered to be a primary inhibition source to the fear memory because it inhibits the amygdala through activating the GABAergic ITC [32, 33]. In fact, activation of the IL leads to the suppression of CS-evoked fear memory [34, 35].

Revision of the IL function

Nevertheless, in a recent optogenetic study in rats [36] , it was shown that activating the IL during extinction reduces fear expression and enhances extinction retrieval the next day, indicating that the IL regulates fear suppression and influences the extent of retrieval, which is consistent with the common view of the IL function; However, it was also shown that silencing the IL during extinction and retrieval had no within-session effect, but silencing the IL during extinction impaired retrieval the following day.

Taken together, the IL is necessary for the formation but not the expression of the extinction memory, suggesting that inhibitory sources other than the IL could also suppress the fear memory [36]. Thus, the functional role of the IL remains controversial.

1.4.3 Fear, extinction, persistent neurons

Recently, the electrophysiological properties of neurons in the amygdala and mPFC have been extensively investigated; interestingly, three different types of neurons in BA have been identified and have been shown to have the following basic properties [37]:

- *'fear neurons'* exhibit CS-evoked activity (spike firing) after fear conditioning and abolished activity after subsequent extinction.
- *'persistent neurons'* also exhibit CS-evoked activity after fear conditioning but are resistant to subsequent extinction and display sustained activity.

- *'extinction neurons'* are silent after fear conditioning but display CS-evoked activity following extinction.

Neural populations that match the definitions of these three-types of neurons are not localized to specific regions; instead, they are distributed over the amygdala and mPFC: fear neurons have been found in the basal nuclei of the amygdala (BA) [38, 37], lateral nuclei of the amygdala (LA) [37, 39–41] and central nuclei of the amygdala (CEA) [42] ; persistent neurons have been found in the BA [38, 37] and LA [37, 40, 41]; and extinction neurons have been found in the BA [38, 37], the group of intercalated cells (ITC) [42] and the vmPFC [43–45].

The following questions arise: How do these three neural populations interact? Furthermore, how do their interactions process both CS and uncertainly generated US inputs during partial reinforcement fear conditioning and generate an extinction-resistant fear response as output?

1.5 Organization of this thesis

Based on these neural findings, this study sought a possible explanation of the PREE by hypothesizing that a combination of fear, persistent and extinction neurons plays an important role in creating the PREE. To test this hypothesis, I first introduce two classical mathematical models of the classical conditioning (Chapter 2), and inspired by which, I propose a mathematical model of a neural circuit based on three neural units with the basic properties of the fear, persistent and extinction neurons (Chapter 3). I then present an extension of the model providing a plausible explanation for the controversial role of the mPFC in the formation of extinction memory (Chapter 4). I then made some interesting predictions using the two neural circuit models (Chapter 5). A statistical inference model also is proposed to apply to the PREE, and compared with the neural circuit models (Chapter 6). At the end, I conclude this dissertation with summarizing my proposals and comparing with previous models (Chapter 7).

Chapter 2

Classical Models for Classical Conditioning

2.1 Framework of a model for conditioning

A learning theory should predict how the associations between CSs and USs change. Most theories propose a simple linear relationship between CS processing and US processing for determining the change in the associative strength, V:

$$\Delta V = (\text{level of US processing }) \times (\text{level of CS processing}). \tag{2.1}$$

In this chapter, I will introduce two famous classical models of the classical conditioning using the framework above.

2.2 The Rescorla-Wagner model

The Rescorla-Wagner model is a mathematical model of classical conditioning, proposed by Rescorla and Wagner in 1972 [46]. This model describes the amount of learning, the trial-by-trial change of the associative strength (V) between a CS and a US, as a result of a conditioning trial. The key point of this model is that the difference between the expectation of the US and its actual value, the prediction error or surprise, drives learning. This surprise can be positive, which drives excitatory learning, or negative, which drives inhibitory learning.

When a trial does not match the expectation, learning will occur and the associative strength will change. The change of the associative strength in the model is determined by

$$\Delta V_X = \alpha_X \beta (\lambda - \sum V_i), \tag{2.2}$$

where ΔV_X is the amount of learning (the change in the associative strength, which is also the predictive value of the US, V) for input X. λ is the actual value of the US, usually set to a value of 1 when the US is present, and 0 when it is absent, but a value other than 1 also might be used if you want to model a larger or smaller US. $\sum V_i$ is the sum of the predictive values of the US based on each conditional stimulus, i.

2.2.1 Successes of the Rescorla-Wagner model

The Rescorla-Wagner model explains the basic acquisition and extinction behaviors. At the beginning of an acquisition phase, the associative strength between a CS and a US is weak, the predicted value of the US is small, which causes a big surprise when US is encountered, and drives a quick learning of the CS-US association; Later in acquisition, as

learning more and more about the CS-US association, the CS predicts a bigger value of US, and the appearance of the US only causes a little, this drives slow learning. Eventually, learning will stop when the CS predicts with certainty that the US will come. Similarly in the extinction phase, the absence of the US results in a negative surprise in each trial, and reduce the predictive value of the US [47].

The Rescorla-Wagner Model makes some unexpected predictions, and many of them have been demonstrated in experiments. For instance,

- *Blocking and Unblocking*

  When a well-established CS (CS1, conditioned by the US) and a novel stimulus (CS2) are both paired with the US, the new learning on CS2 will be blocked, because CS1 itself has already completely predicted the US, and no new learning occurs on either CS1 or CS2.

  Unblocking the learning on CS2 can be achieved by pairing the well-established CS1 and the novel CS2 with a stronger US (US value bigger than 1), so that the unexpected stronger US results in learning of both CS1 and CS2.

- *Conditioned Inhibition*

  If a novel stimulus (CS2) is presented along with well-established CS (CS1), the absence of the US results in a negative surprise, causing both the CS1 and CS2 reduce their predictive value of the US (V). Because the initial V is 0, CS2 then predicts a negative value of the US, becoming a conditioned inhibitor. Repeating the presentation of CS1 with no US, the predictive value of CS2 reaches a $V = -1$, becoming a predictor of the absence of the US.

- *Protection from Extinction*

  If a well-established CS is presented without the US but with a well-established conditioned inhibitor (CI), that is, a stimulus that predicts the absence of the US ($V = -1$), the Rescorla-Wagner model predicts that the CS will not undergo extinction, due to the lack of surprise, because the CI itself is enough to well predict the absence of the US.

- *Over Expectation*

  Suppose there are two CSs (CS1 and CS2) already conditioned with the same US separately, and each by itself can fully predict the US, i.e., each with a $V = 1$. If the two CSs are presented together, and paired with the original US, together they will predict a $V = 1 + 1 = 2$, which over expects the actual value of the US, that is $\lambda = 1$.

Here, the surprise will be negative, and will cause both of the two CSs' predictive values decline.

If the two CSs are presented along with a novel stimulus (CS3), and followed with the original US, CS1 and CS2 together will predict a stronger US, giving rise to a negative surprise to all the CSs including CS3. The predictive value V for CS1 and CS2 declines on the basis of the negative surprise, while the predictive value V for CS3 goes from a initial V = 0 to a V that is negative, which makes CS3 act as a conditioned inhibitor.

2.2.2 Limitations of the Rescorla-Wagner model

Although as we discussed above, the Rescorla-Wagner model successfully explains the basic acquisition and extinction behaviors, and predicts some unexpected but experimentally provable behaviors, there are still significant limitations of the model [47]. As we introduced in the Chapter 1, extinction is not a simple reversal of the conditioned learning, but a newly learned inhibition of the previously learned memory. In the Rescorla-Wagner model, extinction is achieved by reducing the predictive value V, which cannot be recovered once extinguished unless new conditioning is applied. Therefore, the Rescorla-Wagner model fails in explaining the issues that related with the 'new learning' feature of extinction.

- *Spontaneous Recovery*

  Spontaneous recovery is a phenomenon that the conditioned response that is extinguished will recur when a CS is presented after a rest period.

  The Rescorla-Wagner model explains extinction as a reduction of the predictive value V of the CS, which makes V reduce to 0 after extinction. A rest period is incapable to change the value of V, so that the CS no longer predict the US again.

- *Rapid reacquisition*

  If a well-established CS has undergone extinction, pairing the CS again with the US will result in the reacquisition of the conditioned response to the CS more rapidly than the initial acquisition.

  In the Rescorla-Wagner model, extinction makes V reduce to 0, so that the reacquisition starts from $V = 0$ again, same as the initial condition of the original acquisition, therefore, the reacquisition should proceed at the same rate as it did in the original acquisition.

- *Latent Inhibition*

If a CS has been presented alone prior to its pairing with a US, the acquisition of the conditioned response to the CS will be slower than that without the CS-alone presentation before conditioning. Hence, the CS that prior to the conditioning works as a inhibitor to the subsequent conditioning.

However, according to the Rescorla-Wagner model, the CS presentation prior to conditioning will not affect the changing rate of the predictive value V of the CS, therefore, has no effect on the acquisition speed in the subsequent conditioning.

2.3 The TD model

The temporal-difference model (TD model) is an realtime extension of the Rescorla-Wagner model, and is a reinforcement learning algorithm proposed by Sutton & Barton [48] in artificial intelligence and robotics field for real-time learning.

The TD model utilizes the estimation of the value function (V) in the context of reinforcement learning theory to represent learning amount, which is equivalent to the associative strength in the Rescorla-Wagner model.

$$V_t(x) = w_t^T x = \sum_{i=1}^{n} w_t(i) x_t(i), \tag{2.3}$$

where V is the value function, which is regarded as the US prediction in the TD model; $w(i)$ is a modifiable weight corresponding to $x(i)$, the elements of the input stimulus.

Same as the Rescorla-Wagner model, the TD model achieves associative learning through updating the weight w on the basis of the prediction error δ,

$$\Delta w_t = \alpha \delta_t e_t, \tag{2.4}$$

where α is a learning-rate parameter and e_t is eligibility trace levels for each of the stimulus elements, which is a term in the context of reinforcement learning theory, representing the memory parameters associated with the event as eligible for undergoing learning changes.

2.3.1 Differences between the Rescorla-Wagner model and TD model

Compared with the Rescorla-Wagner model, applying the TD model to classical conditioning problem has several key differences:

- *Real-time prediction*

As opposed to the trial-by-trial calculation basis in the Rescorla-Wagner model, the TD model is a realtime algorithm, it includes a time-step in the calculation, which allows the model to predict the US in real-time, and deal with some effects of the CS-US inter-stimulus interval in the conditioning [49].

- *Representation of prediction error*

  In the Rescorla-Wagner model, the prediction error is represented by the difference of the US prediction (associative strength) from the actual US value, while in the TD model, prediction error δ is represented by

  $$\delta_t = r_t + \gamma V_t(x_t) - V_t(x_{t-1}), \tag{2.5}$$

  where r_t is the sum of the US intensity that associatived with all inputs x_t; $V_t(x_t)$ is the new US prediction from the current time step t, γ is the discount factor, and $V_t(x_{t-1})$ is the old US prediction from the previous time step $t-1$. Note that, the US prediction $V_t(x_t)$does not refer to as the prediction of the next US occurrence, but actually as the prediction of accumulative future punishments, with the discount rate γ.

  So instead of comparing the US prediction directly with the US intensity in the Rescorla-Wagner model, the TD model compares the US prediction with the sum of the US intensity and an influence from the new prediction.

  The prediction error δ can be alternatively interpreted as the difference between the US intensity and the change in US prediction:

  $$\delta_t = r_t - [V_t(x_{t-1}) - \gamma V_t(x_t)]. \tag{2.6}$$

  The prediction error is then used to update the weights w.

- *Representation of the conditioned response*

  In the Rescorla-Wagner model, the US prediction, or associative strength, is directly interpreted as the conditioned response. The TD model adopts a slightly more complex response rule that only US prediction above a certain threshold can be used to generate the conditioned response.

  $$a_t = \nu a_{t-1} + [V_t(x_t)]_\theta, \tag{2.7}$$

  where a_t represents the conditioned response at time step t, ν is a small decay constant, showing a small consistency with the previous conditioned response a_{t-1}, and θ is the threshold for US prediction to be integrated with the conditioned response.

2.3.2 Successes of the TD model

As a real-time model, the TD model learns whenever prediction does not match the actual observation , capturing subtle differences that can make empirical predictions beyond the possibility of the Rescorla-Wagner model.

The TD model is also able to capture subtle changes in inter-trial behaviours, such as the inter-stimulus-interval dependency of the conditioning behaviors [50].

- *Inter-stimulus-interval dependency: Delay conditioning*

  The inter-stimulus interval (ISI) is the temporal interval between the onset of the CS and the onset of the US in one pairing of CS-US.

  Based on the relationship between the timing of the CS offset and the timing of the US onset, classical conditioning patterns fall into two categories: delay conditioning, in which the offset of CS is the same time as the onset of the US, i.e., the CS duration is equal to the ISI; and fixed-CS conditioning, in which the CS duration is fixed, and independent of ISI.

  The TD model successfully accounts for the empirical results that the effectiveness of conditioning reduces as the ISI increases in delay conditioning. However, as shown later, the TD model fails predict the inter-stimulus-interval dependency in the fixed-CS conditions.

- *Intratrial effects*

  One of the best-known demonstrations of the intratrial effects is an experiment by Egger-Miller [1962] in which two overlapping CSs conditioned by a US, and ISI of CS1 is longer than that of CS2 . Although CS2 is in a better temporal relationship with the US, the presence of the CS1 reduces conditioning to CS2 as compared to controls in which CS1 is absent.

  The simulation using the TD model under the same conditions reproduces the empirical results above.

- *Second-order conditioning*

  The procedure of the second-order conditioning is similar to that of the conditional inhibitor experiments discussed in the section of the Rescorla-Wagner model above, in which a novel stimulus (CS2) and a well-established CS (CS1, conditioned by the US) are presented without the US. As with the Rescorla-Wagner model, the TD model predicts the negative association of CS2 and US.

2.3.3 Limitations of the TD model

Although the TD model provide a real-time understanding of the classical conditioning, there are also some limitations exist, including most of the phenomena under discrimination, preexposure, and recovery [50, 49].

- *Inter-stimulus-interval dependency: Fixed-CS conditioning*

  Fixed-CS conditioning includes trace conditioning, in which the ISI is greater than the CS duration, but also includes shorter and backward intervals.

  Empirically, backward conditioning (the US occurs before the CS) and simultaneous conditioning (the CS and the US presented together) have occasionally been found to produce weak inhibitory conditioning (associative strength is negative), but more often they produce weak excitatory conditioning (associative strength is positive).

  However, inconsistent with the empirical data, the TD model predicts strong inhibitory conditioning whenever the CS and US overlaps in fixed-CS conditioning .

- *Extinction issues*

  Same as the Rescorla-Wagner model, the lack of representation of the parallel extinction learning, the TD model fails in explaining these phenomenons that related with the *new learning* feature of extinction, such as spontaneous recovery, reinstatement and renewal.

Chapter 3

The Basic Neural Circuit Model

3.1 Lessons from the previous models

As we discussed in Chapter 2, the Rescorla-Wagner model and the TD model cannot explain PREE, because they cannot account for the key feature of extinction, that is, extinction is a 'new learning' of a inhibitory memory, paralleling the learning process of conditioned fear.

Both of the RW and TD models only contain one learning process, either going forward, to increase the associative strength of the US prediction, representing the learning of the conditioned fear, or going backward, to decrease the associative strength or the US prediction, representing extinction. Therefore, the fear conditioning and extinction are not able to be processed parallel in the RW and TD models. The learning process undergone previously does not count for the following extinction speed, which makes whether the effect comes from the full reinforcement or from the partial reinforcement can not be distinguished by the models.

There are also many studies of neural-circuit-based models that addressed the 'new learning' feature of the extinction. A representative work is done by Morén et. al. [51]. Although their model contains both the excitatory and inhibitory learning modules, which allows the fear conditioning and extinction to be processed separately, it does not involve another feature of the fear memory, that is, fear conditioning leaves a indelible trace of the fear memory. In contrast, in Morén et. al.'s model, the inhibitory memory is kept by the inhibitory learning module, while the fear memory can be erased by the inhibitory memory. In partial reinforcement conditioning, the fear memory can be fully erased by the increasing inhibitory learning from the no-US trials even before extinction phase begins.

Taking into consideration the successes and limitations of the previous models, a good model of fear conditioning that is able to account for the partial reinforcement extinction effect needs to include the following features:

- A representation of the conditioned fear response
- A storage that keeps the fear memory
- A parallel extinction process

The existence of the three types of the neurons in the amygdala just meet this requirements. Therefore, the basic model is built on the basis of the three units, as discussed below.

3.2 The basic neural circuit model

3.2.1 Components of the basic model

The basic neural circuit model consists of three units (Fig. 3.1): the fear neural unit (F), which represents the population of the fear neurons in the amygdala; the persistent neural unit (P), which represents the persistent neurons population in the amygdala; and the extinction neural unit (E), which corresponds to the population of extinction neurons in the mPFC.

Note that, the extinction neurons are also interspersed in the amygdala, receiving input from the extinction neurons in the mPFC. In other words, the extinction neurons in the mPFC interact with the fear and persistent neurons in the amygdala via the extinction neurons in the amygdala. In the basic model, the effect of the extinction neurons is equivalent to that of those in the mPFC, thus the extinction neurons in the amygdala are assimilated into the extinction neural unit.

Corresponding to their neural behaviors, the three neural units meet the requirement of a good model we discussed above:

- *Fear neural unit*

  Fear neurons are activated by the CS in the conditioning phase, and gradually become silent in the extinction phase, which agrees with the conditional fear behavior [37], indicating that fear neural unit is an appropriate candidate to represent the conditional fear response.

- *Persistent neural unit*

  Persistent neurons keep the activation to the CS acquired in the conditioning even during the extinction phase [37], suggesting that the persistent neural unit stores the short-term fear memory trace.

- *Extinction neural unit*

  Extinction neurons are silent to the CS in the conditioning phase, but starts to be activated by the CS during the extinction phase [37], indicating the extinction neural unit undergoes inhibitory learning during the trials without the presentation of the US.

3.2.2 Two types of circuits

In the basic model, two types of circuits are composed by the three neural units, being regulated by either neural activities or learning signals.

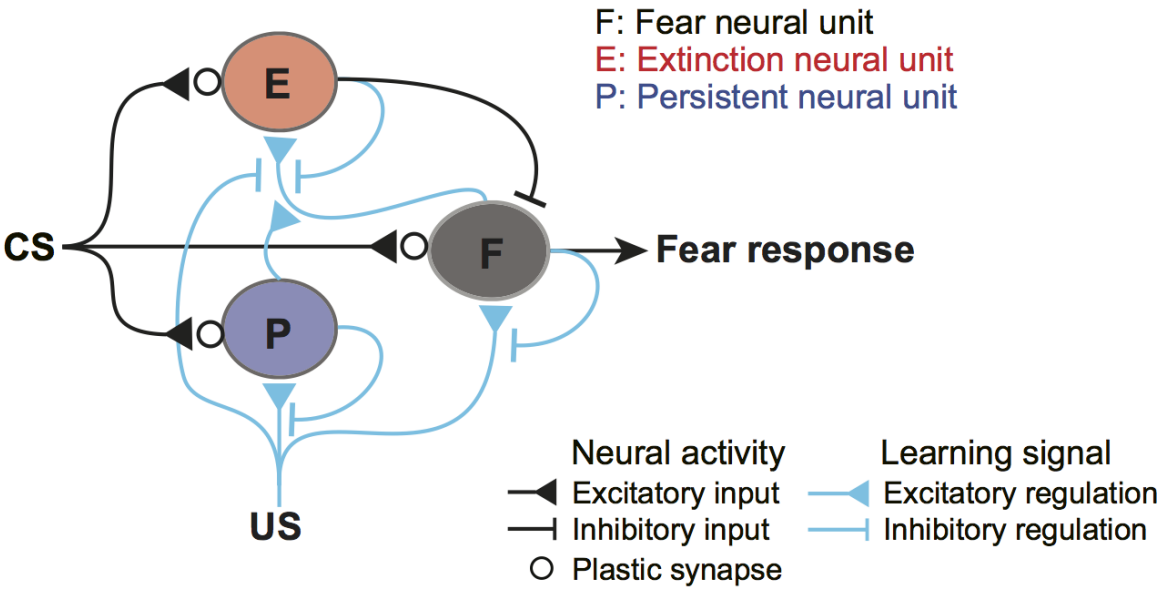

Fig. 3.1 The basic neural circuit model. The basic model consists of the fear (F), persistent (P) and extinction (E) neural units. The CS-related input activates all the neural units; the activity of the fear neural unit determines the fear expression, and is inhibited by the extinction neural unit (black lines) (equations (3.1-3.3)). The CS-related synaptic weights of the fear, persistent and extinction neural units are mediated by the learning signals (blue lines) (equations (3.4-3.6)).

Neural-activity-regulating circuit

In the neural-activity-regulating circuit, all three of the units are activated by the CS. The fear neural unit, which represents the conditioned fear response [52], is inhibited by the extinction neural unit (black line in Fig 3.1D). The activities of the three neural units at trial t are described by:

- *Fear neural unit (F)*

$$F(t) = w_F CS(t) - w_{F,E} E(t), \tag{3.1}$$

  where w_F denotes the synaptic weight of fear neural unit; $w_{F,E}$ is the inhibitory synapse weight from the extinction to the fear neural unit.

  The fear neural unit receives two inputs: the excitatory input CS, and the inhibitory input from the extinction neural unit. It is activated by the CS according to w_F , and inhibited by the extinction neural unit according to $w_{F,E}$.

- *Persistent neural unit (P)*

$$P(t) = w_P CS(t), \tag{3.2}$$

where w_P denotes the synaptic weight of the persistent neural unit.

The persistent unit receives the excitatory input from the CS, and conditionally responds to the CS according to the synaptic weight w_P .

- *Extinction neural unit (E)*

$$E(t) = w_E CS(t), \tag{3.3}$$

where w_E denotes the synaptic weight of the extinction neural unit.

The extinction unit is activated by the CS according to the synaptic weight w_E .

Learning signal-regulating circuit

Learning in the basic model is accomplished through changing the synaptic weight, which refers to the connection strength between the input from the presynaptic neuron and the activity of postsynaptic neuron, playing a essential role in representing learning and memory. Similar to the associative strength in the RW model and the US prediction in the TD model, the synaptic weights are updated on the basis of the prediction errors, which are referred to as learning signals to the corresponding neural units. Each neural unit represents a unique prediction, forming a learning-signal-regulating circuit to drive the change of the synaptic weight.

In the learning-signal-regulating circuit, the synaptic weights of the fear (w_F), persistent (w_P) and extinction (w_E) neural units are updated after each the presentation of the CS-US pairing trial-by-trial, according to the synaptic plasticity rules as follows:

- *Fear neural unit*

$$\Delta w_F = \alpha_F CS(t)\big[US(t) - F(t)\big]_+, \tag{3.4}$$

where α_F is the learning rate of the fear neural unit; and the bracket $[]_+$ is a rectifier that treats the negative values as zero, i.e., the learning rule does not take the negative prediction errors into account.

The model assumes the activity of the fear neural unit predicts the US, i.e., the severity. The prediction error of the severity is then represented by $\big[US(t) - F(t)\big]_+$, forming the learning signal to drive the synaptic plasticity of the fear neural unit.

- *Persistent neural unit*

$$\Delta w_P = \alpha_P CS(t)\big[US(t) - P(t)\big]_+, \tag{3.5}$$

where the α_P is the learning rate of the persistent neural unit.

Basic model	
α_F	0.4
α_P	0.4
α_E	0.4
$w_{F,E}$	2

Table 3.1 Parameters using in the basic model

The persistent neural unit stores the fear memory trace, which also stores the intensity of the US presentation. Therefore, the activity of the persistent neural unit can be interpreted as the prediction of the US intensity, and the corresponding prediction error is $\left[US(t)-P(t)\right]_+$, which forms the learning signal regulating the synaptic plasticity of the synaptic weight w_P.

- *Extinction neural unit*

$$\Delta w_E = \alpha_E CS(t)\left[F(t)\{P(t)-US(t)-E(t)\}\right]_+, \tag{3.6}$$

where the α_E is the learning rate of the extinction neural unit.

The extinction neural unit inhibits the activity of the fear neural unit, comforting the conditioned fear response, which indicates that the activity of the extinction neural unit can be interpreted as the prediction of safety. The prediction error of the safety then is described by $\left[P(t)-US(t)-E(t)\right]_+$.

3.3 Simulation result

3.3.1 Simulation conditions

To ensure that the chance to get reinforced by the US remains the same throughout , the simulation fixes the numbers of trials that CS is followed by the US to 10 trials during the full and partial reinforcement conditioning. In the full reinforcement conditioning , the conditional probability $P(US|CS)=1$, meaning all 10 trials in the conditioning are reinforced by the US. In the partial reinforcement conditioning, I performed simulations on the scenario where the conditional probability is $P(US|CS)=0.5$, in which, the CS is followed by the US in random half trials, and is presented alone in the rest trials.

Both full and partial reinforcement conditioning are followed by an extinction phase with 20 trials without US presence. In the simulation, both CS and US are set to be 1 when present is presented and 0 when absent. The parameters of the basic model used in simulation are shown in Table 3.1.

3.3.2 Full reinforcement

In the simulation of the full reinforcement conditioning and the subsequent extinction (Fig. 3.2A), the CS-evoked activities of the fear, persistent and extinction neural units are consistent with the respective properties of fear, persistent and extinction neurons in the amygdala (Fig. 3.2C) [28, 37, 42] . In addition, the activity of the fear neural unit well represented the behavioral freezing rate as observed in fear conditioning and extinction [37]. Thus, the basic model reproduced the behaviors of the fear, persistent and extinction neurons.

3.3.3 Partial reinforcement

In the partial reinforcement simulation (Fig. 3.2B), during the fear conditioning with partial pairing of the US, the activity of the fear neural unit increased when the US was presented and decreased when it was not (no-US), but the overall activity tended to increase; in contrast, the activities of the persistent and extinction neural units increased only when US and no-US were presented, respectively (Fig. 3.2D). During the subsequent extinction phase, we observed the PREE (Fig. 3.2D): the activity of the fear neural unit slowly decreased with residual activity, in contrast to what was observed in the full reinforcement case (Fig. 3.2C). Residual neural firing has been consistently observed in the amygdala after the extinction training of partially reinforced fear memory [28].

3.3.4 The role of the learning signals

What are the reasons for the difference between the extinction of the fear memory acquired in full and partial reinforcement conditioning?

After full reinforcement conditioning, the CS-evoked activity of the fear and persistent neural units converge at the peak of US prediction and intensity, respectively, i.e., $E = 1, P = 1$, whereas the extinction neural unit is still silent, i.e., $E = 0$ (Fig. 3.2C). Thus, the absence of the US (no-US, $US = 0$) at the beginning of the extinction causes the maximal level of the learning signal to the extinction neural unit, $F(P - US - E)$, leading the extinction neural unit to learn the extinction memory at a high rate.

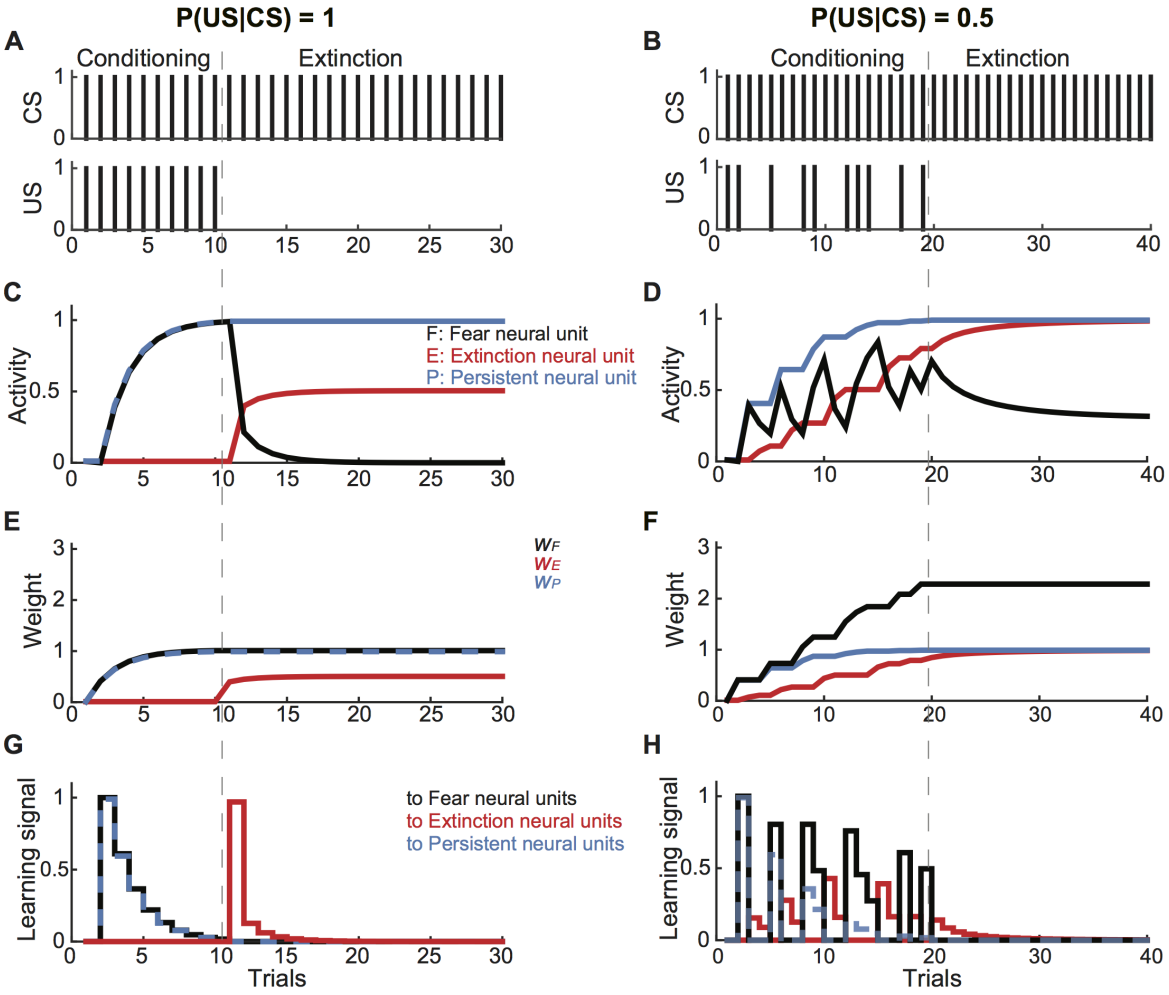

Fig. 3.2 Reproducing PREE using the basic model. Simulation results using the basic model with full ($P(US|CS) = 1$) and partial reinforcement ($P(US|CS) = 0.5$) schedules are presented in the left (A, C and E) and right (B, D and F) columns, respectively. (A, B) The CS and US schedules during fear conditioning and extinction. Same number of the CS-US pairings are presented in both the full (A) and partial reinforcement conditioning (B). (C, D) The black, blue and red lines represent the activities of the fear, persistent and extinction neural units, respectively. (E, F) The black, blue and red lines represent the CS-related synaptic weights of the fear, persistent and extinction neural units, respectively. (G, H) The black, blue and red lines represent the learning signals that regulate the CS-related synaptic weights of the fear, persistent and extinction neural units, respectively. Note that overlapping lines are changed to dashed lines to make them visible.

On the other hand, after partial reinforcement conditioning, although the CS-evoked activity of the persistent neural units also reaches the US intensity, $P = 1$, the fear neural activity cannot predict the definite appearance of the incoming US, $F < 1$, and the extinction neural already shows a certain level of activity, $E \geq 0$ (Fig. 3.2D). Thus, the $US = 0$ at the

beginning of the extinction makes the learning signal to the extinction unit, $F(P-US-E)$, weaker than that in the extinction after full reinforcement conditioning, and causes the slower extinction speed, giving rise to the PREE.

Therefore, the difference between the learning signals to extinction neural unit that are regulated by the different process in the full and partial reinforcement conditioning causes the partial reinforcement effect.

3.3.5 Inhibitory synapse weight $w_{F,E}$

The strength of $GABA_A$ receptor-mediated inhibition in the BLA, which receives inputs from the mPFC, controls extinction of conditioned fear memory [53].

There is a popular belief that the plasticity of brain mainly relies on the plasticity of the glutamatergic excitatory synapses, whereas the GABAergic inhibitory synapses are generally regarded as relatively invariant. Although some recent reports have demonstrated the existence of the plasticity of the GABAergic synapses [54], in this model, I still assume the plasticity of the inhibitory synapses to be negligible compared with the predominant plasticity of the excitatory synapses in the fear conditioning.

Although the plasticity of the inhibitory synaptic is ignored in this model, the potency of inhibition is still controllable by physiological properties of the GABAergic neurons and synapses, as well as extracellular factors. Thus in the simulation, the inhibitory synapse weight $w_{F,E}$ is set to be a constant (Table 3.1). To investigate how the inhibitory potency of extinction to the fear neural unit affects PREE, $w_{F,E}$ also is manually modulated to represent individual differences in physiological properties (Fig.3.3).

The simulation first demonstrates the ability of the inhibitory synapse weight from the extinction neural unit to the fear neural unit in regulating the extinction after partial and full reinforcement (Fig.3.3A, B).

The simulation also shows how the changes of the the inhibitory synapse weight $w_{F,E}$ affects extinction: as $w_{F,E}$ increases, the time constant (Fig.3.3C) and the residual fear memory (Fig.3.3D) of the extinction decreases.

Taken together, the simulation results indicate that the resistance to extinction is aggravated as the inhibitory synapse weight $w_{F,E}$ decreases, and alleviated as $w_{F,E}$ increases. Therefore, the inhibitory synapse weight from the extinction neural unit to the fear neural uint plays a critical role on the PREE .

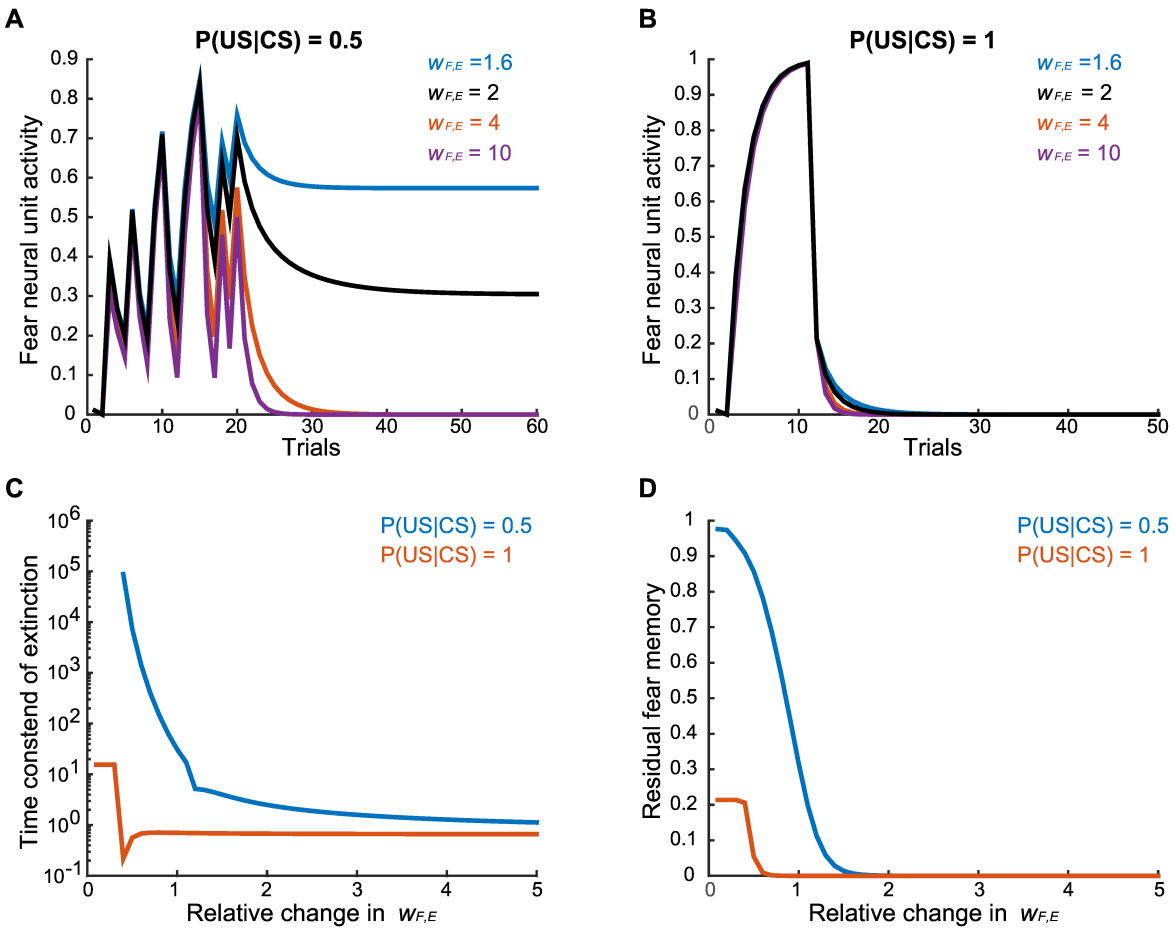

Fig. 3.3 The inhibitory synaptic weight from the extinction to fear neural units affects PREE. (A, B) The activity of the fear neural unit during the partial ($P(US|CS) = 0.5$) and full reinforcement conditioning ($P(US|CS) = 1$) and subsequent extinction is shown with various changes in $w_{F,E}$. Note that α_E is also changed accordingly so that $\alpha_E w_{F,E} = const.$. (C) The blue and red lines represent the time constant of extinction after the partial and full reinforcement conditioning, respectively, with changes in $w_{F,E}$. (D) The blue and red lines indicate the final activity level of the fear neural unit after the extinction following the partial and full reinforcement conditioning, respectively, with changes in $w_{F,E}$.

3.3.6 Uncertainty affects PREE

To test how uncertainty affects PREE, I first evaluate the uncertainty in conditioning using the Shannon entropy. The uncertainty of the next US is defined as the the Shannon entropy of the probability distribution of the waiting time (number of trials: $n \in \{1,2,...\}$) until the next US observation,

$$P(n) = P(US = 1|CS = 1)(1 - P(US = 1|CS = 1))^{n-1}. \tag{3.7}$$

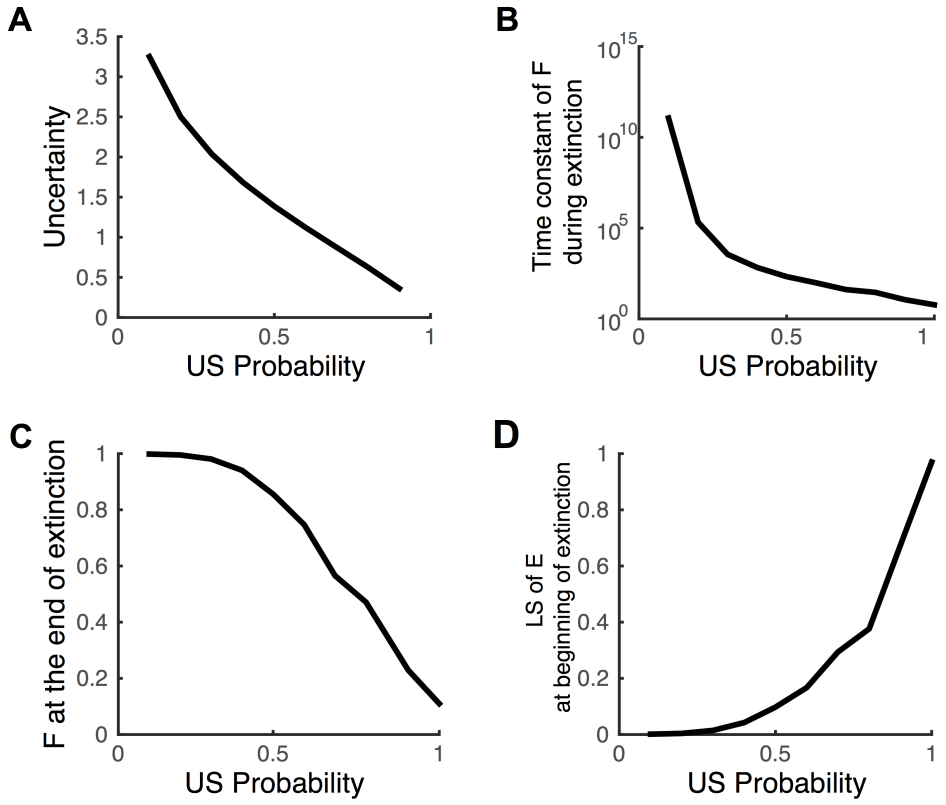

Fig. 3.4 Uncertainty affects PREE in the basic model. (A) The black line represents the uncertainty of the subsequent US occurrence varying the US probability. (B) The time constant of the extinction, which is measured by the decline of the fear neural unit activity, decreases as the US probability increases. (C)The final activity level of the fear neural unit decreases as the US probability increases. (D)The learning signal to the extinction neural unit increases as the US probability decreases.

Next, to evaluate the extinction behavior, I fit the activity of the fear neural unit to the following function:

$$F = F_0 + F_1 e^{-\frac{t}{\tau}}. \tag{3.8}$$

Here, I assume the probability distribution of the fear neural unit activity is belong to the exponential family, based on a common view that the neural activity can be regarded as a Poinsson process with variable means. And the time constant τ is referred to as the τ in the equation above (equation (3.8)), and is used to measure the degree of increase/decline of neural activities.

The simulation result first demonstrates that the uncertainty $P(n)$ is a monotonic decreasing function of the US probability $P(US|CS)$ (Fig. 3.4A).

As US probability varies in the conditioning phase, both the time constant of the extinction, which is measured as the decline of the fear neural unit activity (Fig. 3.4B), and the

final activity level of the fear neural unit (Fig. 3.4B), which represents the final level of fear memory, decrease as the US probability increases, suggesting that the resistance to extinction is alleviated by the increase of the US probability, i.e., the decrease of the uncertainty.

Finally, by investigating the changes of learning signals as the US probability varying, I found that the learning signal to the extinction neural unit at the beginning of the extinction increases as the US probability increases, indicating that the inhibitory potency of the extinction to the fear neural unit also increases to make the PREE alleviated.

Taken together, the uncertainty of the US in the conditioning affects PREE by changing the learning signal to the extinction neural unit.

3.4 Discussion

Extinction as inhibitory learning

It has been generally accepted that extinction is a form of inhibitory learning, which is the opposite of an erasure or forgetting of fear memory [1]. In fact, extinguished fear responses recover under various circumstances. For example, an extinguished fear response spontaneously recovers after a long time, e.g. several days [6], and also reappears after exposure to the US without the CS, known as reinstatement [9]. It has been reported that extinction training could inhibited the fear response but could not erase the fear memory in adult rat, although erasure of fear memory may occur during early stages of postnatal development [55, 56]. Consistently, conditioned fear memory in our model was not erased by a decrease in synaptic weights but was inhibited by extinction neurons.

Encoding

Recent physiological studies have identified neural populations with distinct firing characteristics, such as fear, persistent and extinction neurons, in the amygdala [42, 38] . These findings suggest that information processing in the amygdala may take place through these neural populations. However, the computational role of each neural population in fear conditioning and extinction has not been well studied. In this study, we presented the neural implementation based on these neural populations and proposed the types of information that are encoded and processed through interactions between these neural populations: the fear, persistent, and extinction neurons encode the prediction of net severity, of US intensity and of safety (no-US), respectively, and the weights of their synaptic inputs are modulated by the corresponding prediction errors.

Consistent with the persistent neurons in our model, a previous report showed that CS-evoked activity of the (extinction-resistant) persistent neurons in the LA does not further increase after reconditioning, suggesting that persistent neurons represent the memory of the US intensity [40]. Consistent with the extinction neuron in our model, a human fMRI study showed that the vmPFC uniquely encodes safety accompanied by the CS during extinction [24], suggesting that the CS synaptic input to the vmPFC could be plastically regulated by *prediction error of safety*. This proposed encoding mechanism could be further validated by experiments, such as electrophysiological recording of neural firing or the labeling of cFos immunoreactivity during partial reinforcement fear conditioning and extinction.

Learning signals & neuromodulators

We can speculate that the learning signals in our model could be implemented through neuromodulators such as: dopamine, serotonin, noradrenaline, acetylcholine, norepinephrine and oxytocin [57] . In general, the release of neuromodulators is associated with particular mental states including:reward, positive and negative emotions, happiness, motivation, attention and arousal [58]. Neuromodulators regulate neuronal firing and the quality of synaptic plasticity [59]. In particular, dopamine has been extensively studies, and it is widely accepted that dopamine release is a specific response to reward-related prediction error, i.e., the acquisition of a greater reward than expected [60, 61] and that dopamine release facilitates the synaptic plasticity that underlies the association between sensory input (the CS) and response (the US) [62, 63]. Based on these facts, reinforcement learning theory has suggested that animals perform temporal difference (TD) learning [64, 49] because the basal ganglia, which is involved in decision making, receives dense axonal projections from the VTA and exhibits dopamine-dependent plasticity of synaptic inputs from the cortex [65]. However, it has been known that dopaminergic neurons show firing responses not only to rewards but also to aversive stimuli [66] and, moreover, show diverse firing patterns that may encode prediction errors of other valences [67, 68]. In addition, the VTA was recently suggested to be composed of anatomically and functionally heterogeneous dopaminergic neurons whose axons project to different regions, including the amygdala and mPFC [69] . Taken together, dopamine signals to different neural populations may represent different meanings, such as the prediction error of net severity, of US intensity and of safety, as assumed in our model.

Chapter 4

The Extended Neural Circuit Model

4.1 Background

A brain subdivision that plays the same function as the extinction neural unit, being an inhibition source of the fear memory, is the infralimbic (IL) subregion of mPFC [37, 70], whose CS-evoked activity increases during fear extinction training, and activation of IL during extinction facilitates extinction of fear responses [44, 45].

In a recent optogenetic study in rats [36] , it was shown that activating the IL during extinction reduces fear expression and enhances extinction retrieval the next day, indicating that the IL regulates fear suppression and influences the extent of retrieval, which is consistent with the common view of the IL function; However, it was also shown that silencing the IL during extinction and retrieval had no within-session effect, but silencing the IL during extinction impaired retrieval the following day.

4.2 The extended neural circuit model

The result that activation of the IL enhances extinction formation and retrieval suggests that the IL plays a role in fear suppression, i.e., the IL can be interpreted as the extinction neural unit in the basic model; But silencing the IL has no within-session effect during extinction and retrieval indicating fear expression must be inhibited by an another fear suppression substrate besides the IL. i.e., another extinction neural unit should be included in the model.

The immediate enhanced fear extinction when activating the IL during extinction and the delayed effect of silencing the IL during extinction on retrieval the next day suggests multiple timescales of synaptic plasticity exist in the new extinction neural unit.

Therefore, the basic model needs to be extended by introducing two features: another extinction neural unit and multiple timescales of synaptic plasticity.

4.2.1 Another extinction neural unit

As we discussed above, besides keeping the fear and persistent neural units in the basic model, the extended neural circuit model needs to include two extinction neural units. the infralimbic (IL) subregion of mPFC can be regarded as one of the extinction neural unit in the extended model.

A candidate of the second extinction neural unit is the intercalated (ITC) cells in the amygdala, located between the basolateral (BLA) and central nuclei (CEA) of the amygdala. ITC consists of a group of GABAergic neurons, receiving inputs from the IL , and projecting

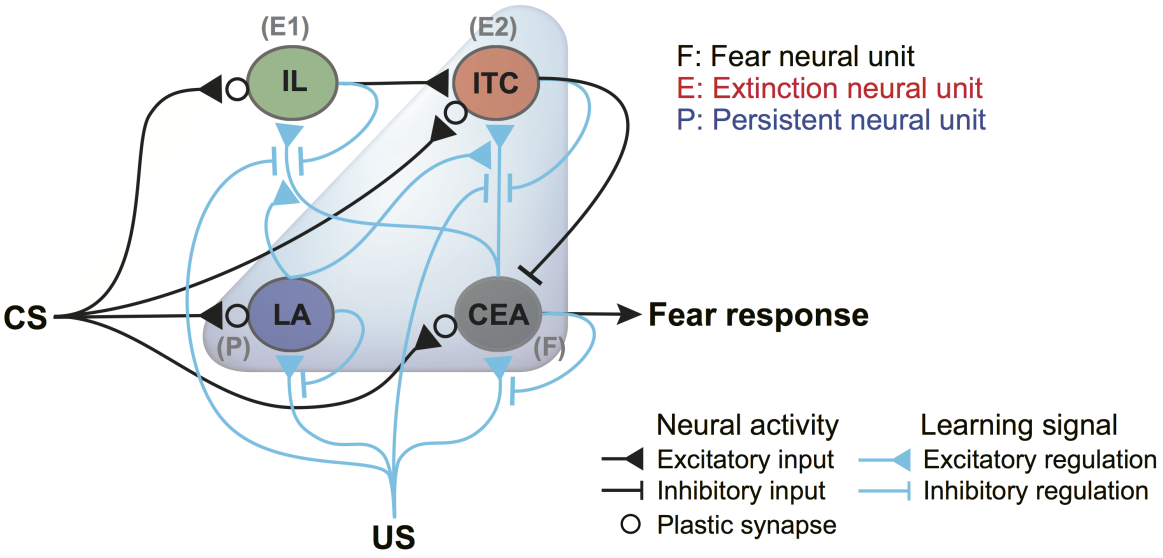

Fig. 4.1 The extended neural circuit model. Extended model includes subregions of the amygdala (the LA, CEA and ITC), as well as the IL in the mPFC. In this model, the LA and CEA correspond to the persistent and fear neural units, respectively, and the ITC and IL correspond to two extinction neural units. The CS input activates all the neural units; the IL sends excitatory input to the ITC, and the activation of the ITC inhibits the CEA (black lines) (equations (4.1-4.4)). Activity of the CEA determines the behavioral fear expression. The weights of the plastic synapses (black open circles) are modulated according to their learning signals (blue lines) (equations (4.5-4.9)).

to the BLA and CEA to suppress fear expression [71, 72]. Thus, the ITC can be considered as another extinction neural unit in the extended model.

Corresponding the rest of the neural units in the model to the brain subdivisions, fear neural unit can be compared to the CEA, due to the CEA's role in determining fear expression; and the persistent neural unit can be compared to the LA, due to the fact that persistent-firing neurons are found in the LA [73].

In the extended model, the LA, i.e., the persistent neural unit (P), is activated by the CS input, as well as the rest neural units in the model; the activity of the CEA, i.e., fear neural unit (F), is directly inhibited by the ITC, i.e., one of the extinction units ($E2$), which is regulated by the IL, i.e., the other extinction unit ($E1$), due to the fact that the IL sends projections to the ITC, then transfers to the fear and persistent neurons in the BA,

Thus, the neural-activity-regulating circuit (black lines in the Fig. 4.1) in the extended model can be described as follows:

$$F(t) = w_F(t)CS(t) - w_{F,E2}E2(t), \tag{4.1}$$

$$P(t) = w_P(t)CS(t), \tag{4.2}$$

$$E1(t) = w_{E1}(t)CS(t), \tag{4.3}$$

$$E2(t) = w_{E2}(t)CS(t) + w_{E2,E1}E1(t), \tag{4.4}$$

where w_F, w_P, w_{E1}, andw_{E2} indicate the activity-dependent plastic synaptic weights of the CEA, LA, IL and ITC, respectively; $w_{F,E2}$ indicates the constant inhibitory synapse weight of the ITC onto CEA; and $w_{E2,E1}$ indicates the constant synaptic weight projected from the IL to ITC.

4.2.2 Multiple timescales of synaptic plasticity

Same as the basic model, the synaptic weights of the CEA, LA and IL are modified by the learning-signal-regulating circuit (blue lines in the Fig. 4.1)according to the Rescorla-Wagner rule:

$$\Delta w_F = \alpha_F CS(t)\big[US(t) - F(t)\big]_+, \tag{4.5}$$

$$\Delta w_P = \alpha_P CS(t)\big[US(t) - P(t)\big]_+, \tag{4.6}$$

$$\Delta w_{E1} = \alpha_{E1} CS(t)\big[F(t)\{P(t) - US(t) - E1(t)\}\big]_+, \tag{4.7}$$

where α_F, α_P, and α_{E1} indicate the learning rates of the CEA (F), LA(P) and IL($E1$), respectively; contents in the brackets $\big[\big]_+$ indicate the learning signals to the corresponding neural units.

As for the ITC ($E2$), two different timescales are necessary to be introduced to the ITC ($E2$) to describe the early- and late-phase plasticities [74, 75] :

$$\Delta w_E2(t) = \alpha_{E2} CS(t)\big[F(t)\{P(t) - US(t) - E2(t)\}\big]_+ - \beta_{E2}(w_{E2} - w^{\infty}_{E2}), \tag{4.8}$$

where the first term represents the dynamics of the early-phase plasticity, and the second term indicates the late-phase dynamics; α_{E2} indicates the learning rates of the ITC in the early-phase plasticity; $-\beta_{E2}$ indicates the the relaxation rate of w_{E2} in the late-phase plasticity; w^{∞}_{E2} indicates the capacity of the weight w_{E2}, whose dynamics is described as follows:

$$\Delta w^{\infty}_{E2} = \alpha^{\infty}_{E2} CS(t)E1(t) - \beta^{\infty}_{E2}(w_{E2}(t) - w^{\infty}_{E2}(t)), \tag{4.9}$$

where α^{∞}_{E2} indicates the learning rates of the ITC in the late-phase plasticity; β^{∞}_{E2} indicates the the relaxation rate of w^{∞}_{E2} .

According to eqs (4.8) and (4.9), w_{E2} is consolidated to w^{∞}_{E2}, and w^{∞}_{E2} is also relaxed to w_{E2} depending on the IL activity, $E1(t)$.

Extended model	
α_F	0.4
α_P	0.4
α_{E1}	0.4
α_{E2}	0.4
α_{E2}^{∞}	0.03
β_{E2}	0.01
β_{E2}^{∞}	0.005
$w_{E2,E1}$	0.05
$w_{F,E2}$	2

Table 4.1 Parameters using in the extended model

4.3 Simulation results

4.3.1 Simulation conditions

Two simulations are implemented in this section.

First, the results of silencing the IL during extinction and retrieval is simulated, in which, the fear conditioning employs the full reinforcement conditioning paradigm (CS-US) for 10 trials, followed by 20 CS-only trials for extinction, and a long resting phase (1 day; approximately represented by 100 trials in simulation) with the absence of both the CS and US, then at the end, 10 CS-only trials are presented to test the retrieval of the fear and extinction memories.

The partial reinforcement extinction effect is also simulated using the exactly same setting as that used in the basic model simulations in Chapter 3.

The parameters of the extended model used in the simulations are shown in Table 4.1.

4.3.2 Effects of silencing IL during extinction and retrieval

The extended model consistently reproduced the experimental results observed when activating and silencing of the IL during extinction and retrieval (Fig. 4.2C-F).

This extended model first shows the essential behaviors of the acquisition and extinction of fear memory through full reinforcement conditioning paradigm (Fig. 4.2B). At the retrieval of the extinction memory after the resting phase, spontaneous recovery of the fear memory occurred to a small extent, as commonly observed after long intervals [5, 7]. This is because

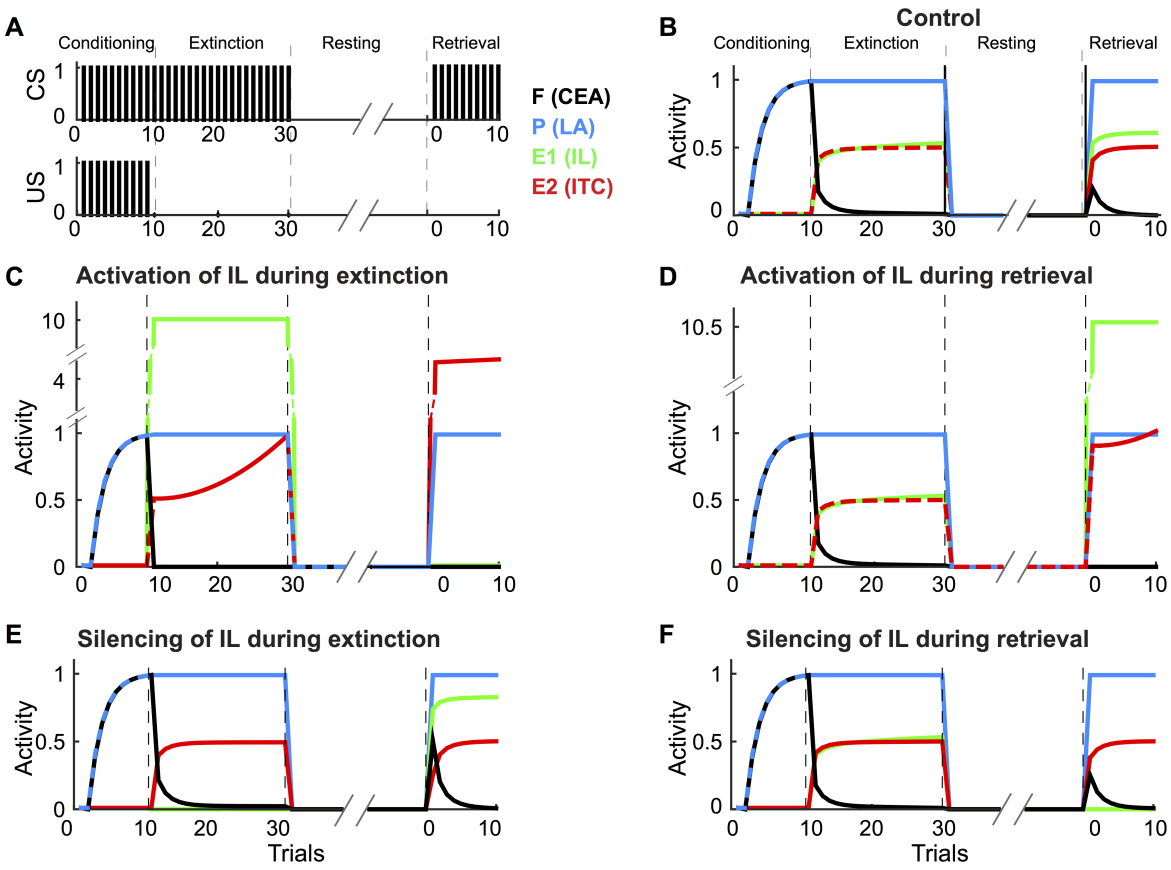

Fig. 4.2 Effects of activation and silencing the IL using the extended model. (A) The CS and US schedules; the presentation of the CS and US in the conditioning and extinction phases accords with the full reinforcement conditioning schedule, which is the same as those used in the basic model (Fig. 3.2A); followings are the resting phase, during which both the CS and the US are absent, and the retrieval phase, during which only the unreinforced CSs are presented to test the retrieval of the extinction memory. (B-F) Simulated neural activity of the CEA (black lines), LA (blue lines), IL (green lines) and ITC (red lines) using the extended model for the control condition (B), IL activation during extinction (C), IL activation during retrieval (D), IL silencing during extinction (E) and IL silencing during retrieval (F). The blue, green, red and black lines indicate the activity of the LA (persistent neural unit), IL (extinction neural unit), ITC (another group of extinction neural unit) and CEA (fear neural unit), respectively. Note that the dashed lines are used to represent overlaps to keep all the results visible. The synaptic weights of each neural unit are shown in Fig. 4.3.

.

during the resting phase, the synaptic weight to the ITC, w_{E2}, settled down to the weight capacity, w_{E2}^{∞} , due to the slow dynamics of the late-phase LTP.

Activation of the IL during both extinction (Fig. 4.2C) and retrieval (Fig. 4.2D) reduces the within-session fear expression, and the IL activation during extinction also facilitated the

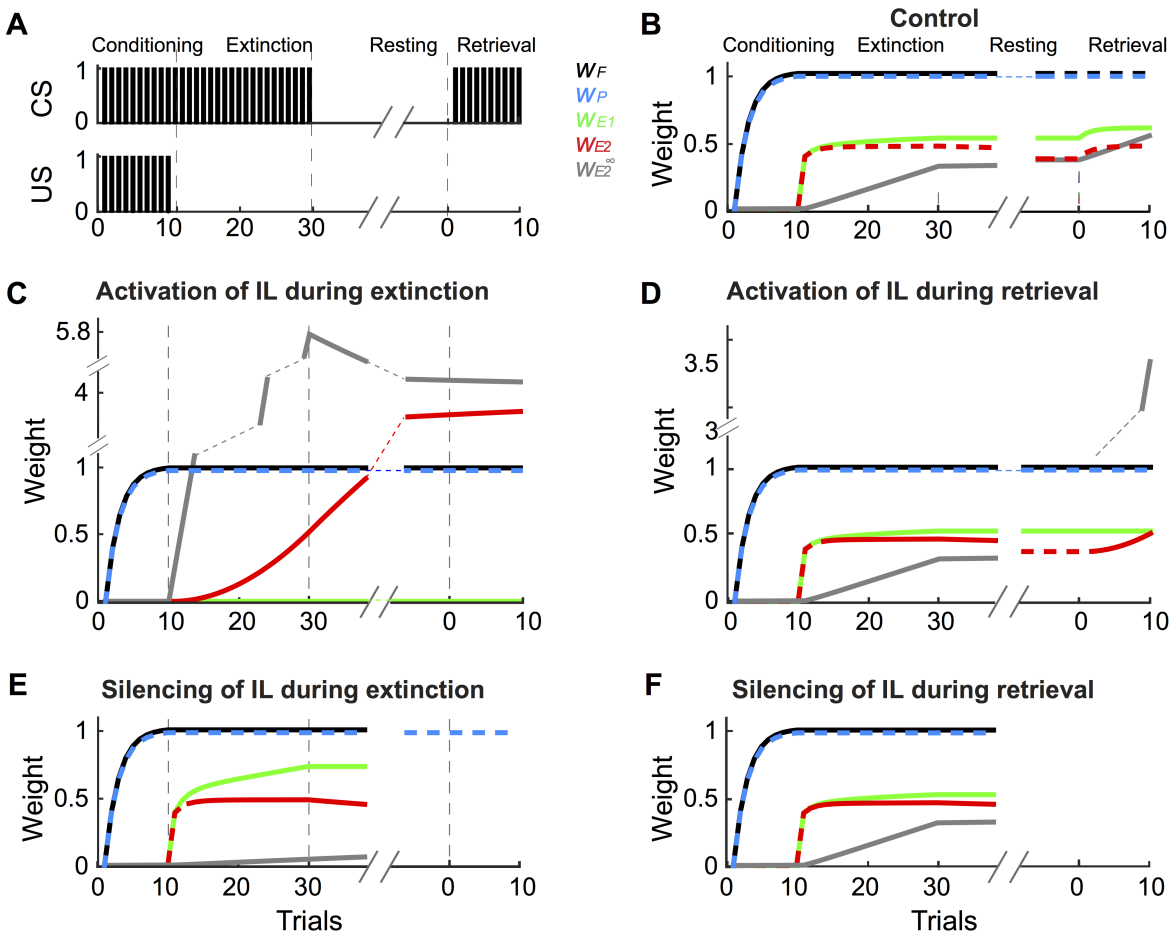

Fig. 4.3 Changes in the synaptic weights in Fig. 4.2. (A) The CS and US schedules, same as Fig. 4.2A. (B-F) The blue, green, black lines represent the CS-related synaptic weight of the LA, IL and CEA, respectively; the red and gray lines represent CS-related synaptic weights of the ITC regulated by the early- and late-phase plasticity, respectively.

subsequent consolidation of the extinction memory, impairing the following retrieval of the fear memory in the next day (Fig. 4.2C).

Silencing the IL during both extinction (Fig. 4.2E) and retrieval (Fig. 4.2F) has no effect on the within-session extinction behavior. However, silencing the IL during extinction causes the absence of the consolidation of the extinction memory during the resting phase, leading to a significant recovery of the fear memory (Fig. 4.2E).

The simulation results are consistent with recent optogenetic studies [36, 76], and also suggest that extinction learning is regulated by separate multiple timescales synaptic plasticity mechanisms consisting of fast early-phase memory formation that is independent of the IL and slow late-phase memory consolidation that depends on the IL, as assumed in our model.

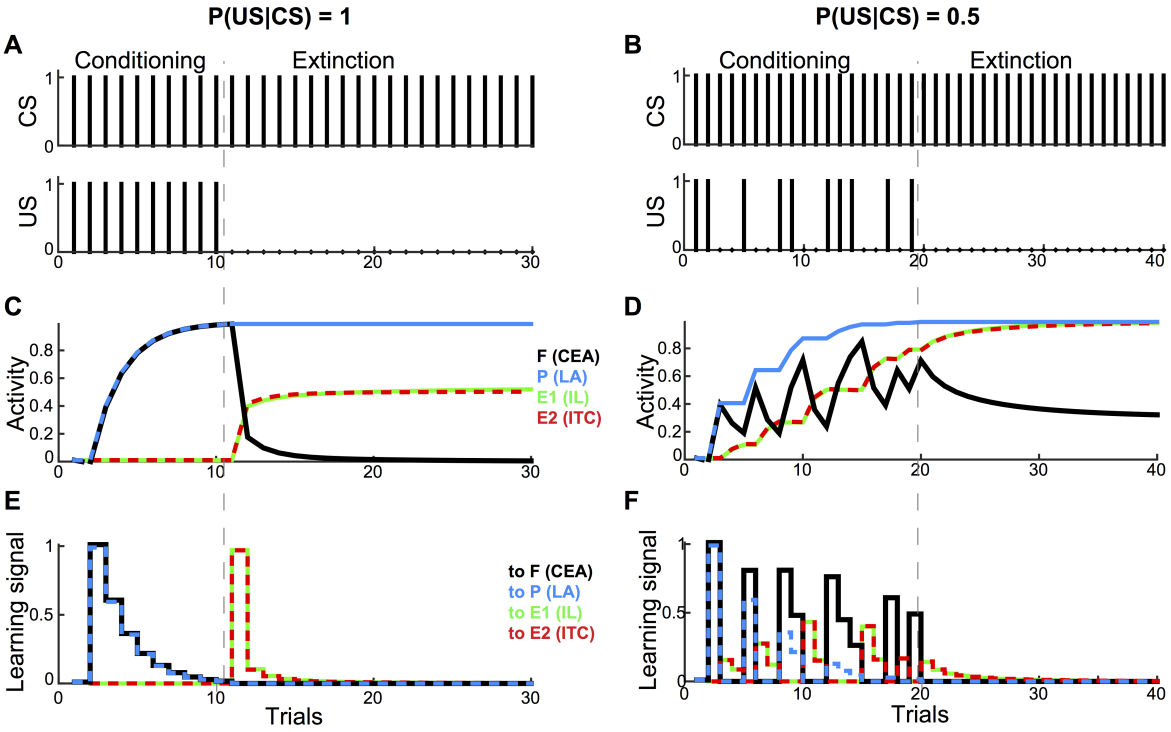

Fig. 4.4 Reproducing PREE using the extended model. Simulation results using the extended model with full and partial reinforcement schedules are presented in the left (A, C and E) and right (B, D and F) columns, respectively. (A, B) US schedules during fear conditioning and extinction, same as those settings in Fig. 3.2A, B. (C, D) The blue, green, red and black lines represent the activities of the LA (persistent neural unit), IL (extinction neural unit), ITC (another group of extinction neural unit) and CEA (fear neural unit), respectively. (E, F) The blue, green, red and black lines represent the learning signals that regulate the CS-related synaptic weights of the LA, IL, ITC and CEA, respectively.

4.3.3 Partial reinforcement extinction effect

Consistent with the basic model, the extended model also reproduces the resistant to extinction after the partial reinforcement conditioning (Fig. 4.4).

To compare with the basic model, in the simulation, the same conditioning schedules with full and partial reinforcement and the model parameters are adopted in the simulation. Without a long resting time, the ITC is mainly mediated by the early-phase plasticity rule. Therefore, the degree of the PREE generated by the extinction model are quite similar to the basic model results.

4.4 Discussion

Early- and late-phase plasticity

What are the molecular substrates of early- and late-phase plasticity?

The early-phase long-term potentiation (LTP) is regulated by Ca2+ signaling-regulated phosphorylation of AMPA-R on endosomes, which induces the exocytosis and membrane accumulation of AMPA-R [77]. In contrast, late-phase LTP is thought to be regulated by gene expression with slow dynamics, in which proteins are newly synthesized in the soma, actively transported to spines and inserted into the postsynaptic density (PSD) [78, 79] . Thus, w_{E2} and w_{E2}^{∞} in equations correspond to the total number of AMPA-Rs on membrane and the size of PSD, i.e., AMPA-R capacity, respectively. Then, each term in equations (4.8, 4.9) can be interpreted as the following biological processes: The first term in equation (4.8) represents the early-phase LTP, i.e., an increase in the total AMPA-Rs, regulated by the learning signal. The first term in equation (4.9) represents an increase in the AMPA-R capacity, regulated by the IL. The second term in equation (4.8) represents spontaneous cycling (i.e., exocytosis and endocytosis) of the total AMPA-Rs, converging to the AMPA-R capacity. The second term in equation (4.9) represents spontaneous cycling of the AMPA-R capacity (i.e., synthesis and degradation of proteins in the PSD) depending on the total AMPA-Rs.

Chapter 5

Model Prediction

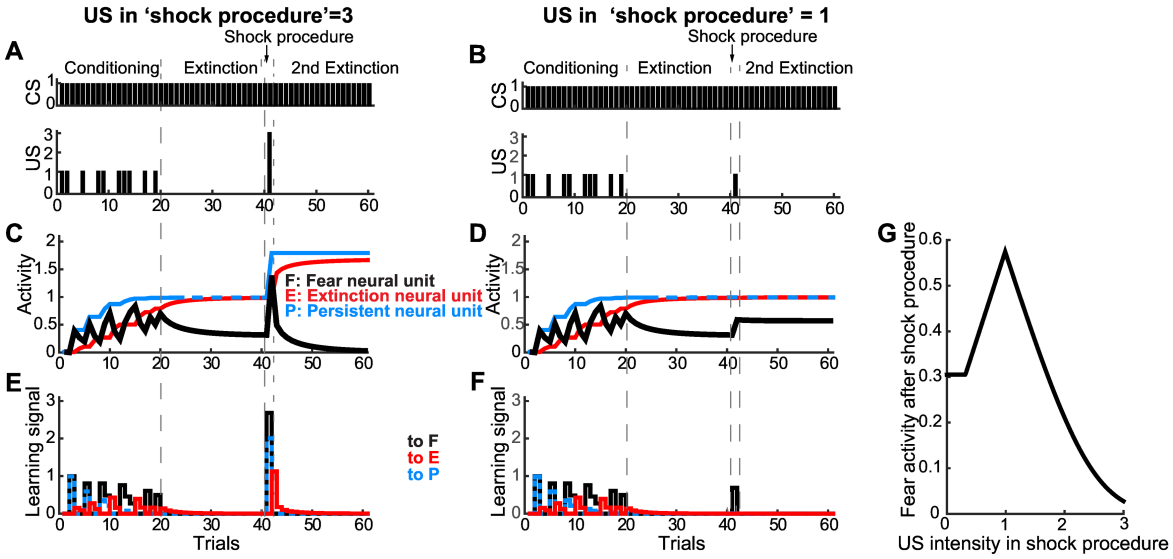

Fig. 5.1 Shock procedure using the basic model. (A, B) The CS and US schedules; a paradigm of partial reinforcement conditioning and extinction is first presented, the schedule of which is set the same as those used in the PREE simulations using the basic and extended model, Fig. 3.2A and 4.4A; following is an additional presentation of the CS, reinforced by a three times stronger US (A) and a US with original intensity (B), and another extinction phase. (C, D) The black, red and blue lines represent the activities of the fear, persistent and extinction neural units, respectively. (E, F) The black, red and blue lines represent the learning signals to the fear, persistent and extinction neural units, respectively. (G) The final neural activity level of the fear neural unit after re-extinction training is mediated by the intensity of the US used to reinforce the additional CS in the shock procedure.

5.1 Shock procedure extinguishes residual fear memory suffers PREE

The resistance to extinction after partial reinforcement conditioning leaves certain level of residual fear memory. To fully extinguish the fear memory that undergoes PREE, I propose a *shock procedure* following the partial reinforcement conditioning and extinction.

Extinguishing the fear memory needs to strengthen the extinction memory, which is regulated by the learning signal to the extinction neural uint, $\left[F(t)\{P(t)-US(t)-E(t)\}\right]_+$. We therefore need to find a way to increase the learning signal. A bigger learning signal to the extinction neural uint implies a bigger surprise when meet the absence of the US, suggesting that the subject need to experience a stronger US to induce a bigger expectation of US before the trial.

Following the partial reinforcement conditioning and extinction that induces PREE, in the shock procedure, a one-trial full reinforced conditioning is presented, in which the CS reinforced by a stronger US, followed by an another extinction training.

We then tested this shock procedure by using the basic model. The activity of the fear neural unit is first rapidly elevated due to high intensity of the US (Fig. 5.1A) and then rapidly decreased to almost 0 during the subsequent extinction (black line in Fig. 5.1C), indicating that the extinction-resistant fear memory has been completely inhibited. When the shock procedure employing an US of the same intensity is applied (Fig. 5.14B), the fear memory is conversely reinforced (Fig. 5.1D), suggesting that the intensity of the US is a critical determinant for the effectiveness of the shock procedure. The differences in the outcomes of these cases are due to different levels of learning signals to two extinction neural units ($E1$ and $E2$) at the beginning of the second extinction (Fig. 5.1E and F).

By testing the shock procedure using various US intensities, the effectiveness of the shock procedure, which is measured by the final activity level of the fear neural unit after re-extinction training, is shown in Fig. 5.1G, by (Fig. 5.1G). The result shows that changing the intensity of the US used to reinforce the additional CS affects the result of the shock procedure: a weaker US in the shock produce actually enhances the fear memory, while a stronger US strengthens the extinction memory, and are able to extinguish the residual fear memory.

The success of the shock procedure is also confirmed in the extended model (Fig. 5.2). The fear memory is represented by the CEA activity (black line in Fig. 5.2C), and shows a overt reduction after shock procedure. The extended model also demonstrated that the intensity of the US used in the shock procedure is crucial for its effectiveness in extinguishing the fear memory.

5.2 Repeating full reinforcement conditioning and extinction causes PREE-like results

In addition, the neural circuits model predict a PREE-like effect during successive full reinforcement conditioning and extinction, which is equivalent to partial reinforcement conditioning overall (Fig.5.3 and 5.4). As conditioning and extinction repeat, the residual activity of the fear neural unit accumulates and becomes saturated. In fact, it has been seen in the literature that the re-conditioned fear memory exhibited substantial resistance to re-extinction [80–83]. The inhibitory synaptic weight from the extinction neural unit to the fear neural unit also plays a crucial role in the PREE-like effect.

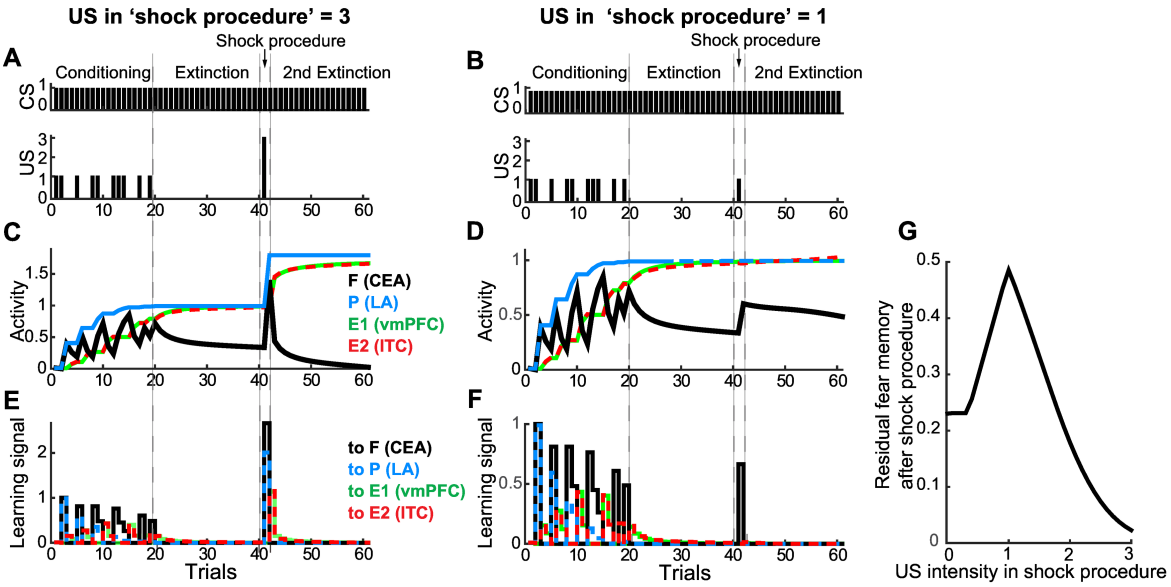

Fig. 5.2 Shock procedure using the extended model. (A, B) The CS and US schedules; same as that used in Fig. 5.2. (C, D) The black, blue, green and red lines represent the neural activities of the CEA (fear neural unit), LA (persistent neural unit), IL (extinction neural unit), and ITC (another group of extinction neural unit), respectively. (E, F) The black, blue, green and red lines represent the learning signals that regulates the CS-related synaptic weights of the CEA, LA, IL and ITC, respectively. Note that overlapped lines are changed to dashed lines to make them visible. (G) The final neural activity levels of the CEA (fear neural unit) after re-extinction training are mediated by the intensity of the US used to reinforce the additional CS in the shock procedure.

5.3 Discussion

Anxiety disorders

Post-traumatic stress disorder (PTSD) is a disorder that occur in some people who have experienced a shocking, scary, or horror events. PTSD patients have difficulty in forgetting the strong fear memory acquired in the event, and undergo a number of symptoms, such as nightmares, insomnia, irritability, impaired concentration, and unconscious avoidance the situation related with the event. Nearly 8% of the general population suffers the symptoms of PTSD [84, 85], and this number increases to roughly 30% in the veterans [86].

The avoidance symptom indicates that PTSD can be regarded as a Pavlovian classical conditioning, during which patients acquire the conditioned fear memory to the original neutral situation.

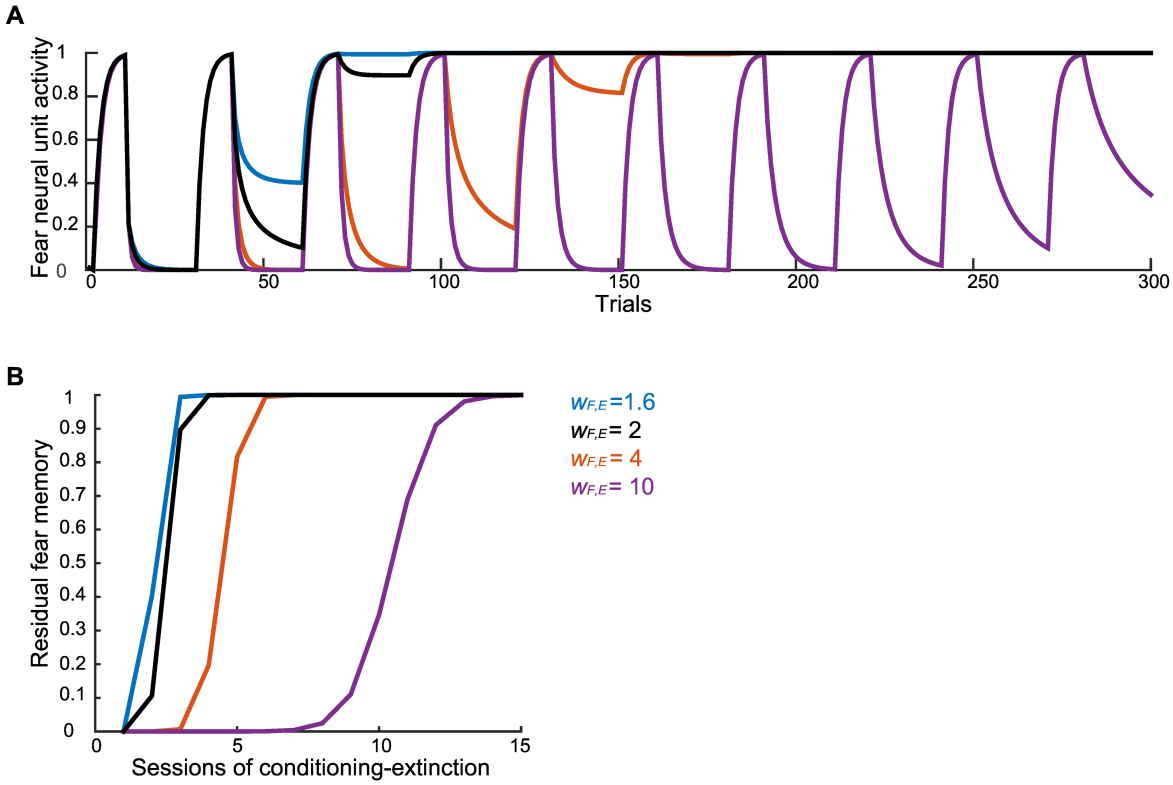

Fig. 5.3 Successive conditioning and extinction induces PREE using the basic model. (A) The activity of the fear neural unit during successive conditioning and extinction with changes in $w_{F,E}$. Note that α_E is also concurrently changed so that $\alpha_E w_{F,E} = const$. (B) The residual fear memory represented by the final activity level of the fear neural unit after each extinction session is plotted with changes in $w_{F,E}$.

Exposure therapy is considered as an effective treatment on PTSD. It exposes the situation that reminds of the fear memory to the patients. It is similar as the extinction phase used in the fear conditioning.

Fear conditioning has been used as a model system for anxiety disorders such as panic disorder, PTSD and obsessive-compulsive disorder (OCD) [87–89]. Traditional exposure therapy, which corresponds to extinction training, is an effective cure for anxiety disorders in some patients [90], but some severe patients also show strong resistance to exposure therapy [88, 89]. Moreover, anxiety disorders may be worsened by occasionally experienced negative social reactions [91, 92], which are akin to partial reinforcement experiences, and become strongly resistant to exposure therapy, similar to the PREE. Thus, we think that the widely used fear conditioning with full reinforcement, in which the acquired fear memory can be easily diminished by extinction training, is rare in real life and thus is not a good model for understanding anxiety disorders; instead, fear conditioning with partial reinforcement, which

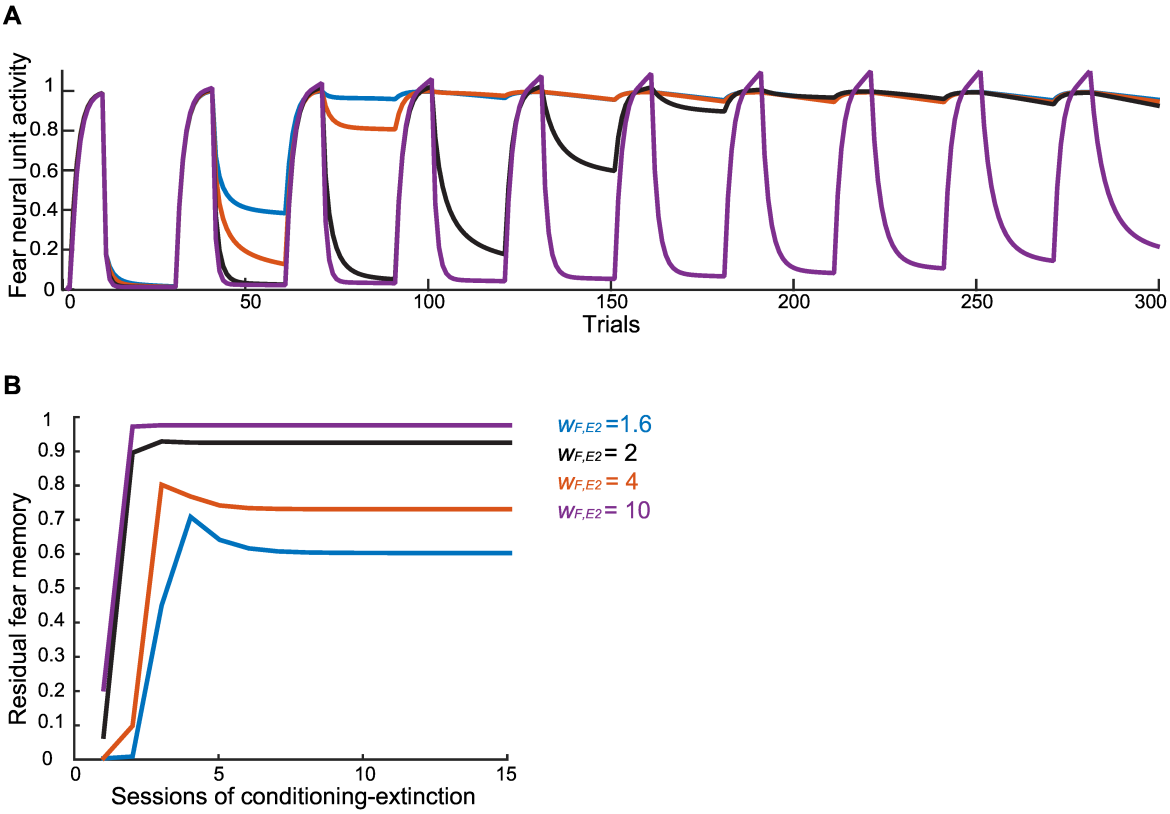

Fig. 5.4 Successive conditioning and extinction induces PREE using the extended model. (A) The activity of CEA during successive conditioning and extinction with changes in $w_{F,E2}$. Note that α_{E2} is also changed accordingly so that $w_{F,E2}\alpha_{E2} = const$. (B) The residual fear memory, which is represented by the final activity level of the CEA, after each extinction session is plotted with changes in $w_{F,E2}$.

results in an extinction-resistant fear memory, is a more realistic model for neuroscience research on anxiety disorders and should improve the translatability of results [20, 28].

To relieve extinction-resistant fear memory, we proposed a *shock procedure* based on our models. The fear memory was diminished by extinction if a stronger US was paired with the CS before the extinction procedure (Fig. 5.1 and 5.2). In the shock procedure, although the fear memory is temporarily strengthened by the stronger US, the subsequent extinction training becomes effective, suggesting that an increase in activity in the amygdala (persistent and fear neurons) or in the learning signal to the vmPFC is key for effective extinction training. The shock procedure can be tested in animal experiments, but employing a stronger US as part of the shock procedure may raise ethical concerns for humans. It can be intuitively interpreted that the animal cannot comprehend the rule change to extinction after fear conditioning with partial reinforcement, whereas after the shock procedure with a strong US, in contrast, the animal can internalize the pronounced rule change to extinction, thus

allowing the fear memory to be extinguished. The proposed *shock procedure* may provide insight not only for the development of new therapies but also for understanding the neural mechanisms of fear memory extinction.

Chapter 6

Statistical Inference Model

6.1 Introduction

What type of statistical processing do animals perform during fear conditioning with partial reinforcement and subsequent extinction?

The fact that extinction learning largely depends on the US probability during fear conditioning (Fig. 3.4), which reflects the uncertainty, suggests that the uncertainty must be encoded in the brain. Thus, animals may process the statistical properties of sequential US observations.

6.2 Statistical Inference model

To quantify the degree of surprise from the perspective of inferring whether the US would occur, I developed a statistical model based on logistic regression with sequential Bayesian updating [93]. We assumed that the animals inferred the probability that the US would occur based on the previous observation of the US as

$$\pi_t = \sigma(w_{t,0} + w_{t,1} US_{t-1}), \tag{6.1}$$

and sequentially updated $w_{t+1} = [w_{t+1,0}, w_{t+1,1}]$ to predict the probability based on a new instance of the US, where US_t denotes a binary variable, i.e.,$US = 0$ and $US = 1$ US=0 represent the presence and absence of the US, respectively; $w_{t,0}$ and $w_{t,1}$ are internal variables in animals, representing the effects of inferring the probability that the US would occur independent of and dependent on the previous US, respectively; $\sigma(\cdot)$is the logistic function$\sigma(x) = \frac{1}{1+e^{-x}}$; π_t represents the probability of the US occurrence that the animals infer based on the previous US, US_{t-1}. We also assumed that the degree of surprise (S) that animals feel in response to the US or no-US is quantified by the amount of information [94]:

$$S(\text{US}) = -\log\pi_t, \tag{6.2}$$

$$S(\text{no-US}) = -\log(1-\pi_t). \tag{6.3}$$

We considered w_{t+1} to be updated by sequential Bayesian updating as

$$P(w_t|US_{1:t}) \propto P(US_t|w_t, US_{1:t-1} P(w_t|US_{1:t-1}) \tag{6.4}$$

$$= P(US_t|w_t, US_{t-1} \int P(w_t|w_{t-1}) P(w_{t-1}|US_{1:t-1}) dw_{t-1}, \tag{6.5}$$

where $P(w_t|US_{1:t-1})$ and $P(w_t|US_{1:t})$ represent the prior and posterior distributions, respectively. The animals were assumed to think that instances of the US were simply generated by a Bernoulli process as

$$P(US_t|w_t, US_{t-1}) = \pi_t^{US_t}(1-\pi_t)^{1-US_t}. \tag{6.6}$$

In addition, the animals were assumed to believe that $w_t$remained almost constant with a small degree of noise as $w_t = w_{t-1} + \varepsilon_t$, where ε_t represents independent and identically distributed Gaussian noise with a mean of zero and a low variance, s. Then,

$$P(w_t|w_{t-1} = N(w_t|w_{t-1}, sI), \tag{6.7}$$

where I represents the unit matrix.

The Bayesian belief update was implemented as an extended Kalman filter, in which the prior and posterior distributions are approximated by Gaussian distributions. $P(w_{t-1}|US_{1:t-1})$ was represented by

$$P(w_{t-1}|US_{1:t-1}) = N(w_{t-1}|\mu_{t-1}, \Sigma_{t-1}), \tag{6.8}$$

where N represents the function of the Gaussian distribution parameterized by Σ_t and μ_t, which are the variance-covariance matrix and mean vector of w_t, respectively. Then, equation (6.6) can be transformed into

$$P(w_t|US_{1:t}) \propto P(US_t|w_t, US_{1:t-1})N(w_t|\mu_{t-1}, s+\Sigma_{t-1}). \tag{6.9}$$

Note that the prior and posterior distributions in equation (6.9) do not have conjugate relationship. Thus, the posterior distribution, $P(w_t|US_{1:t})$, can be approximated by a Gaussian distribution using Laplace approximation as

$$P(w_t|US_{1:t}) = N(w_t|\mu_t, \Sigma_t), \tag{6.10}$$

where μ_t and Σ_t are updated as

$$\mu_t = \arg_{w_t}\max E(w_t), \tag{6.11}$$

$$\Sigma_t^{-1} = -\Delta\Delta E(w_t) = (s+\Sigma_{t-1})^{-1} + \sigma(\mu_t^T\phi_{t-1})(1-\sigma(\mu_t^T\phi_{t-1}))\phi_{t-1}\phi_{t-1}^T, \tag{6.12}$$

where ϕ indicates $(1, US_{t-1})^T$ and

$$E(w_t) = -\frac{1}{2}(w_t - \mu_{t-1})^T (s + \Sigma_{t-1})^{-1} (w_t - \mu_{t-1}) \tag{6.13}$$

$$+ US_t \log\sigma(w_t^T \phi_{t-1}) + (1 - US_t)\log(1 - \sigma(w_t^T \phi_{t-1})) + const. \tag{6.14}$$

To calculate μ_t requires numerical optimization because $E(w_t)$ is a non-linear function of w_t.

6.3 Simulation results

6.3.1 Fear memory as a statistical inference

Because continuous and discontinuous presentation of the US during full and partial reinforcement paradigms leads to fear memories that differ in their resistance to extinction, the effect of US continuity is incorporated into the model; the probability that the US would occur is inferred based on the previous observation of the US (equation (6.1)). Then, the simulation demonstrated that, as the basic model, the sequentially predicted US probability reproduced, (Fig. 3.2) the characteristics of the extinction of fear memory acquired through both the full and partial reinforcement schedules (Fig. 6.1C adn D) in a manner consistent with biological fear neurons.

6.3.2 Uncertainty affects PREE

This statistical model also generated several properties that depended on the US probability, as obtained in the neural circuit model. These properties included the time constant of extinction (Fig. 6.2B), the amount of residual fear memory that remained after extinction (S4C Fig), and the degree of surprise to the no-US (Fig. 6.2D), which was quantified as the amount of information (Fig. 6.1E and F). Moreover, we compared the measure of surprise in the neural circuit model (the learning signals) with the measure of surprise in the statistical model (amount of information) when the same US pattern was applied, and we found a high degree of correlation, providing additional evidence that the learning signal in the neural circuit model represents the degree of surprise in terms of statistics (Fig. 6.2E and F). Taken together, these commonalities between the two models suggest that the neural circuit model that consisted of fear, persistent and extinction neurons effectively processed the statistical property of the occurrence of the US through sequential updating of Bayesian logistic regression.

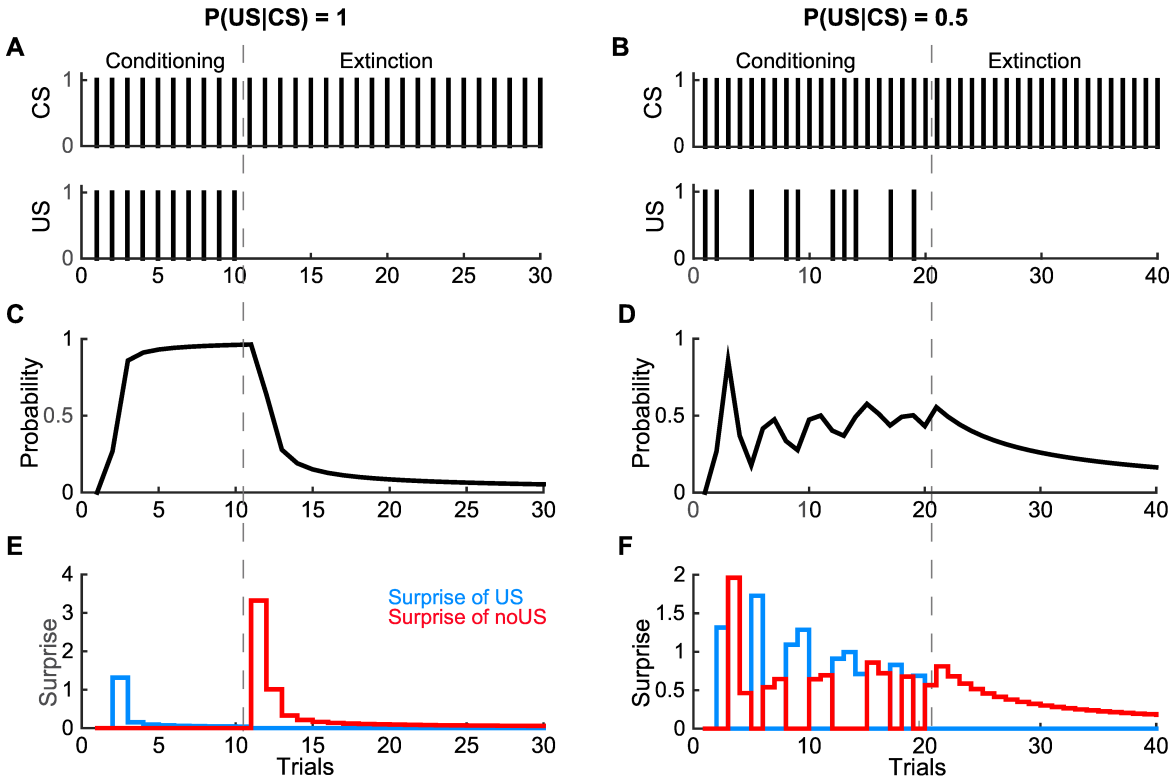

Fig. 6.1 Estimated US probability in the full and partial conditioning using the statistical inference model. Simulation results using the statistical inference model with full ($P(US|CS) = 1$and partial ($P(US|CS) = 0.5$ reinforcement schedules are presented in the left (A, C, and E) and right (B, D, and F) columns, respectively. (A, B) The CS and US schedules; same as those used in Fig. 3.2A,B and Fig. 4.4A,B. (C, D) The black lines present the estimated US probability by the statistical model. (E, F) The blue and red lines present the degree of surprises to the US and no-USs (i.e., the absence of the US), respectively. The surprises are measured by the amount of information gained when encountering a US and no-US input, according to $logP(US)$ and $log(1-P(US))$, respectively.

6.3.3 Relations between statistical surprise and learning signals

In addition, we found that the learning signal to the extinction neural unit was correlated with the degree of *surprise* from a statistical standpoint; after fear conditioning with full reinforcement, the no-US input at the beginning of the extinction phase was unpredictable, leading to a relatively large degree of surprise (red line in Fig. 3.2G). In contrast, after fear conditioning with partial reinforcement, the no-US input at the beginning of the extinction phase was predictable, leading to a relatively small degree of surprise (red line in Fig. 3.2H).

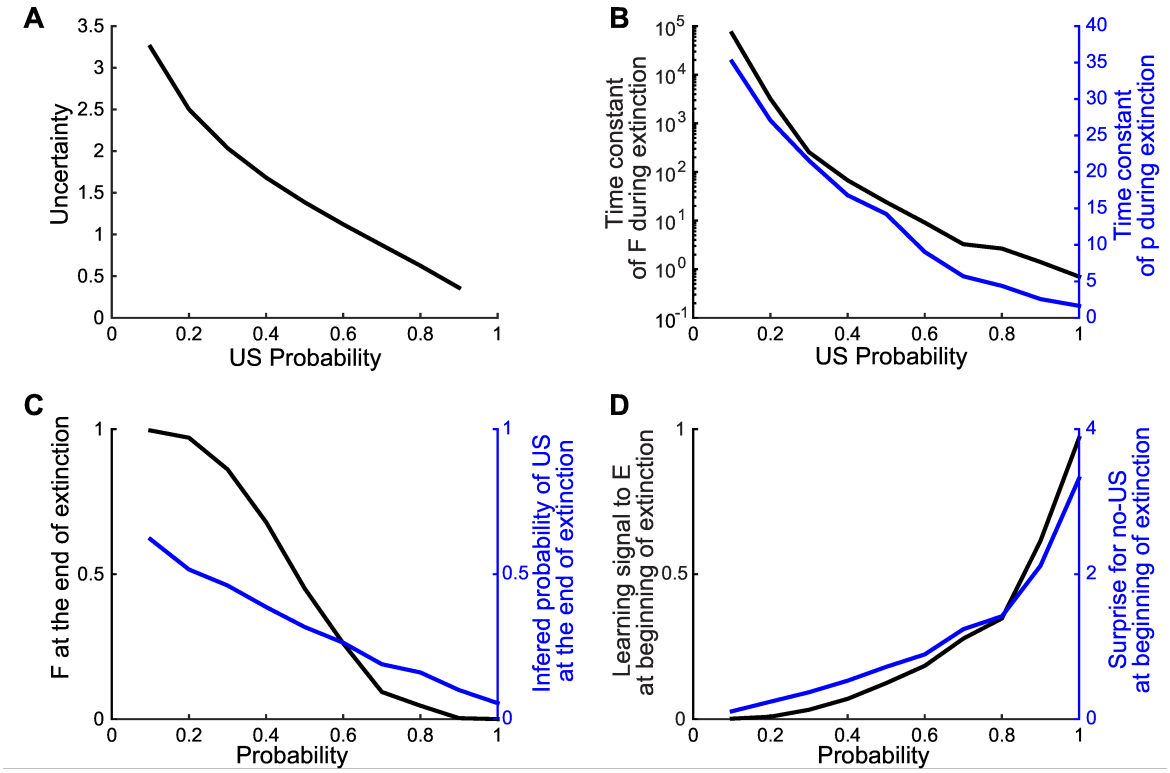

Fig. 6.2 Uncertainty-depended PREE using both the neural circuit model and the statistical inference model (A) The black line represents the uncertainty of the subsequent US occurrence varying the US probability. (B) The black and blue lines represent the time constant of the extinction varying the US probability, in the forms of the decline of the fear neural unit activity in the basic model and the estimated US probability in the statistical inference model, respectively. (C) The black and blue lines represent the final level of the fear memory varying the US probability, in the forms of the fear neural unit activity and the estimated US probability in the statistical inference model, respectively. (D) The black and blue lines indicate the surprise associated with the no-US, i.e., the learning signal to the extinction neural unit in the basic neural circuit model and the amount of information associated with a no-US observation in the statistical inference model, respectively, as a function of US probability during fear conditioning.

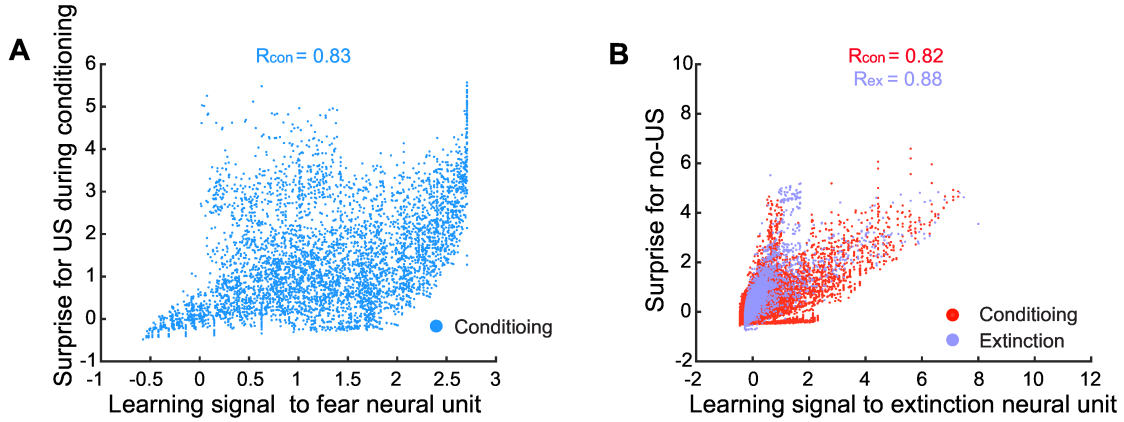

Fig. 6.3 Correlations of the learning signals and the statistical surprises. The same CS and US schedules are applied in the both model. (A) Each dot represents a relationship between the surprise for US in the statistical inference model and the learning signal to the fear neural unit in the basic model at each trial in the conditioning phase. (B) Each dot represents a relationship between the surprise for no-US in the statistical inference model and the learning signal to the extinction neural unit in the basic model at each trial in the extinction phase.

Chapter 7

Conclusion

7.1 Summary

In this dissertation, I mainly focus on providing an explanation of the neural mechanisms of uncertainty-dependent resistance to extinction after partial reinforcement fear conditioning, which is known as the partial reinforcement extinction effect (PREE).

In Chapter 1, I introduced the background of classical conditioning, fear response, PREE, and the brain regions that are related to the fear neural circuit. I also introduced a finding of three types of neurons in the amygdala: fear, extinction and persistent neurons, which constitute the basis of the models I proposed.

In Chapter 2, I reviewed two classical models of classical conditioning, which are the Rescorla-Wagner model, and the TD model. Both of these models can explain learning and extinction of fear responses but lack the ability to explain the inhibition memory that is observed in experiments. The models have other strengths and weaknesses in explaining the properties of classical learning, and provide a good framework for developing the models proposed in this thesis.

In Chapter 3, I proposed a basic neural circuit model, which consists of the fear, extinction and persistent neural units. The three components are regulated by the synaptic-weight-updating rules based on the Rescorla-Wagner model, and their activities predict the US occurrence, US intensity and the safety. The simulation results of this model demonstrate its success in reproducing the PREE, and reveal that the learning signal to the extinction neural unit gained in the no-US trials in the partial reinforcement conditioning is the key factor causing PREE. The simulation results also indicate that the inhibitory synaptic weight from the extinction neural unit to the fear neural unit and the uncertainty during the conditioning, which is correlated with the US probability, are also crucial to PREE.

In Chapter 4, I first interpreted the three neural units in the basic model to correspond to the subdivisions of the brain: the fear neural unit corresponds to the central nucleus of amygdala (CEA), the persistent neural unit corresponds to the lateral nucleus of amygdala (LA), and the extinction neural unit corresponds to the infralimbic cortex (IL) of the mPFC. A recent research shows that silencing IL does not affect the within-session behaviors, but changes the retrieval of the extinction memory after a long resting phase, suggesting that the IL is not the only inhibiting complex in the brain. Therefore, I extended the basic model to include another extinction unit, which corresponds to the intercalated cells (ITC) in the amygdala. The new extinction unit (ITC) has different synaptic regulating rules from the other three units; the CS-related input synaptic weight is regulated by the early- and late-phase plasticity. The simulation results of the extended model first reproduces the effects of

silencing and activating IL, consistent with the experiment reports, and also demonstrated the PREE, due to the regulating rule of the learning signals as the basic model.

In Chapter 5, I made some predictions based on the basic and extended neural circuit models proposed in Chapter 3 and Chapter 4. To fully extinguished the residual fear memory that suffers PREE, I proposed a *shock procedure*, which is presented after the partial reinforcement and extinction that still leaves a certain level of fear memory. In the shock procedure, a CS paired with a stronger US is presented, and followed by another extinction phase. The simulation results in both basic and extended neural circuit model show the success of extinguishing the residual fear memory by using shock procedure, and also pointed out that the US intensity in the shock procedure affects the effectiveness of extinguishing the residual fear memory. Using the two models, I also predict a PREE-like effect during successively repeating the full reinforcement conditioning and extinction.

In Chapter 6, I developed a statistical inference model to estimated the probability of US occurrence, implementing a Bayesian logistic regression algorithm. Using this statistics model, I defined the statistical surprise when encountering a US or no-US (the absence of the US) input, which is highly correlated with the learning signals to the fear and extinction neuron units, respectively. The simulation indicates that the neural-circuit based learning signals encode the statistical surprise.

7.2 Comparison with previous theoretical models

Classical conditioning has been computationally modeled in a number of ways. In the field of behavioral psychology, *former learning theory*, which defines mathematical embodiments to describe learning and behavioral phenomena, has been tested [86]. The Rescorla-Wagner model was a seminal former learning theory that described an association between CS and US controlled by prediction error as a learning signal [43]. Since then, many alternative models have been proposed to reproduce many observed phenomena in classical conditioning and extinction [95–99] . However, these models failed to explain the PREE. Moreover, these models did not fully describe their neural mechanisms, although several models can be implemented using neural networks [100, 101].

Reinforcement learning was proposed as an extension of the Rescorla-Wagner model; in this system, which animals explore optimal behavioral strategies by interacting with their environment to maximize the accumulated reward over time [48]. Sutton and Barto proposed temporal difference (TD) learning, in which the prediction of expected cumulative future reward was updated by its prediction error, called TD error [48]. The framework of TD learning reproduced classical conditioning and extinction [49] but not the PREE.

To account for the PREE, TD learning was extended by two models [102, 103]. Redish et al. [102] introduced a categorization process for inexperienced observations into new latent states, whereas Song et al. [103] introduced arousal signal-dependent learning to the existing TD learning model [104]. Although these TD learning-based models were successful in reproducing the PREE, how neural computation is performed by fear, persistent and extinction neurons has remained unclear.

Another aspect of computational modeling is *statistical decision theory* [105]. The PREE has been addressed by Bayesian estimation of the US probability per trial or unit time [106–108]. This framework was extended to introduce latent causes [109]. Related to latent causes, Gershman et al. [110] developed a Bayesian inference model based on a categorization process of contexts [102]. This model was further extended by introducing a hidden Markov model (HMM), in which a particular context tended to persist over time, and this model successfully generated the PREE [111] . Although these models provided important concepts in light of statistical decision making, they did not describe the underlying neural mechanisms.

There have also been two types of approaches to computational models of neural circuits consisting of the amygdala and other brain regions. One approach is the firing-rate model, in which neural units represent the average firing rate of neurons, neural populations or brain regions [112–114]. Balkenius et al. [114] first developed a mathematical model in which fear memory was extinguished by inhibition of amygdalar activity by the orbitofrontal cortex, which is subdivision of the vmPFC, but their model did not focus on and hence failed to reproduce the PREE. Moustafa et al. [113] developed a neural circuit model consisting of the BLA, CEA, ITC and vmPFC (i.e., IL), combined with TD learning, in which synaptic plasticity was regulated by the TD error as a learning signal. In their study, however, the PREE was not explicitly addressed, though it was mentioned that their model exhibited the PREE only when extensive training trials were performed in the acquisition phase, with no detailed results.

The other approach is based on a spiking-neuron model, in which action potentials are simulated based on membrane voltage dynamics. Nair's group has extensively developed biophysically realistic conductance-based models that express several types of firing patters that have been experimentally observed [115–118] . These studies addressed fear conditioning only with full reinforcement, not with partial reinforcement, while investigating the roles of synaptic input from the vmPFC to the ITC [115], interaction between prelimbic cortex (PC) in the mPFC and the BA [116], and synaptic inputs from thalamus and cortex to the LA [117]. On the other hand, Vlachos et al. [119] first proposed a large-scale neural network

model of the BA by introducing populations of fear, persistent and extinction neurons, but that work did not address the PREE.

7.3 Conclusion

Compared with these previous models, the basic and the extended neural circuit models proposed in this dissertation are the first neural network models of the amygdala/mPFC circuits that could satisfactorily explain the PREE.

The key unique feature of our models is the constitution of the fear, extinction, and the persistent neural units. Especially, the proposal of the inclusion of the persistent model. To my knowledge, it is the first time a neural circuit model that addresses fear conditioning is able to keep the memory trace of the US intensity, which is measured by the activity of the persistent neural unit and the actual received US input.

Also, although the concept that the safety prediction is not original, our proposal of the *degree of safety* , which is represented by the difference between memory of the US intensity and the US that actually observed, $P - US$, is the first time to be applied in a model that addresses fear conditioning.

Based on crosstalk between the fear, persistent and extinction neurons, the basic and extended model gain the ability to process the probabilistic nature of the CS-US pairing, the uncertainty.

References

[1] Karyn M Myers and Michael Davis. Behavioral and neural analysis of extinction. *Neuron*, 36(4):567–584, 2002.

[2] Dale W Leonard. Partial reinforcement effects in classical aversive conditioning in rabbits and human beings. *Journal of comparative and physiological psychology*, 88(2):596–608, 1975.

[3] Robert A Rescorla. Partial reinforcement reduces the associative change produced by nonreinforcement. *Journal of Experimental Psychology: Animal Behavior Processes*, 25(4):403–414, 1999.

[4] Nicholas John Mackintosh. *The psychology of animal learning.* Academic Press, 1974.

[5] Ivan P Pavlov. *Conditioned reflexes: An Investigation of the physiological activity of the cerebral cortex.* 1927.

[6] Gregory J Quirk. Memory for extinction of conditioned fear is long-lasting and persists following spontaneous recovery. *Learning & Memory*, 9(6):402–407, 2002.

[7] Robert A Rescorla. Spontaneous recovery. *Learning & Memory*, 11(5):501–509, 2004.

[8] Robert A Rescorla and C Donald Heth. Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 1(1):88, 1975.

[9] R. Frederick Westbrook, Mihaela Iordanova, Gavan McNally, Rick Richardson, and Justin A. Harris. Reinstatement of fear to an extinguished conditioned stimulus: two roles for context. *Journal of Experimental Psychology: Animal Behavior Processes*, 28(1):97–110, 2002.

[10] Mark E Bouton and Robert C Bolles. Contextual control of the extinction of conditioned fear. *Learning and Motivation*, 10(4):445–466, 1979.

[11] Mark E Bouton and David A King. Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes*, 9(3):248, 1983.

[12] Lloyd G Humphreys. The effect of random alternation of reinforcement on the acquisition and extinction of conditioned eyelid reactions. *Journal of Experimental Psychology*, 25(2):141–158, 1939.

[13] Abram Amsel. Frustrative nonreward in partial reinforcement and discrimination learning: some recent history and a theoretical extension. *Psychological review*, 69(4):306–328, 1962.

[14] Abram Amsel. Frustration theory: Many years later., 1992.

[15] E J Capaldi, Dick Hart, and Larry R Stanley. Effect of intertrial reinforcement on the aftereffect of nonreinforcement and resistance to extinction. *Journal of Experimental Psychology*, 65(1):70–74, 1963.

[16] Joseph E LeDoux. Emotion circuits in the brain. *Annual Review of Neuroscience*, 23(1):155–184, mar 2000.

[17] Stephen Maren and Gregory J Quirk. Neuronal signalling of fear memory. *Nature Reviews Neuroscience*, 5(11):844–852, 2004.

[18] Joshua P. Johansen, Christopher K. Cain, Linnaea E. Ostroff, and Joseph E. LeDoux. Molecular mechanisms of fear learning and memory. *Cell*, 147(3):509–524, oct 2011.

[19] Francisco Sotres-Bayon, David E A Bush, and Joseph E LeDoux. Emotional perseveration: an update on prefrontal-amygdala interactions in fear extinction. *Learning & Memory*, 11(5):525–535, 2004.

[20] Mohammed R Milad and Gregory J Quirk. Fear extinction as a model for translational neuroscience: ten years of progress. *Annual review of psychology*, 63:129–151, 2012.

[21] Sevil Duvarci and Denis Pare. Amygdala microcircuits controlling learned fear. *Neuron*, 82(5):966–980, 2014.

[22] Elizabeth A. Phelps, Mauricio R. Delgado, Katherine I. Nearing, and Joseph E. Ledoux. Extinction learning in humans: role of the amygdala and vmPFC. *Neuron*, 43(6):897–905, 2004.

[23] Mohammed R Milad, Christopher I Wright, Scott P Orr, Roger K Pitman, Gregory J Quirk, and Scott L Rauch. Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological psychiatry*, 62(5):446–454, 2007.

[24] Daniela Schiller, Ifat Levy, Yael Niv, Joseph E LeDoux, and Elizabeth A Phelps. From fear to safety and back: reversal of fear in the human brain. *The journal of neuroscience*, 28(45):11517–11525, 2008.

[25] Mohammed R Milad, Roger K Pitman, Cameron B Ellis, Andrea L Gold, Lisa M Shin, Natasha B Lasko, Mohamed A Zeidan, Kathryn Handwerger, Scott P Orr, and Scott L Rauch. Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biological psychiatry*, 66(12):1075–1082, 2009.

[26] Daniela Schiller, Jonathan W Kanen, Joseph E LeDoux, Marie-H Monfils, and Elizabeth A Phelps. Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, 110(50):20040–20045, 2013.

[27] Pan Feng, Yong Zheng, and Tingyong Feng. Spontaneous brain activity following fear reminder of fear conditioning by using resting-state functional MRI. *Scientific reports*, 5:16701, 2015.

[28] Uri Livneh and Rony Paz. Amygdala-prefrontal synchronization underlies resistance to extinction of aversive memories. *Neuron*, 75(1):133–142, 2012.

[29] P Sah, E S L Faber, M Lopez DE Armentia, and J Power. The amygdaloid complex: anatomy and physiology. *Physiological Reviews*, 83(3):803–834, jul 2003.

[30] Amit Etkin, Tobias Egner, and Raffael Kalisch. Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in cognitive sciences*, 15(2):85–93, feb 2011.

[31] Thomas F Giustino and Stephen Maren. The Role of the Medial Prefrontal Cortex in the Conditioning and Extinction of Fear. *Frontiers in Behavioral Neuroscience*, 9:298, nov 2015.

[32] Ekaterina Likhtik, Daniela Popa, John Apergis-Schoute, George a Fidacaro, and Denis Paré. Amygdala intercalated neurons are required for expression of fear extinction. *Nature*, 454(7204):642–645, 2008.

[33] Cornelia Strobel, Roger Marek, Helen M. Gooch, Robert K.P. Sullivan, and Pankaj Sah. Prefrontal and auditory input to intercalated neurons of the amygdala. *Cell Reports*, 10(9):1435–1442, 2015.

[34] Brittany M Thompson, Michael V Baratta, Joseph C Biedenkapp, Jerry W Rudy, Linda R Watkins, and Steven F Maier. Activation of the infralimbic cortex in a fear context enhances extinction learning. *Learning & memory (Cold Spring Harbor, N.Y.)*, 17(11):591–599, 2010.

[35] Gregory J Quirk, Ekaterina Likhtik, Joe Guillaume Pelletier, and Denis Paré. Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. *The Journal of Neuroscience*, 23(25):8800–8807, 2003.

[36] Fabricio H Do-Monte, Gabriela Manzano-Nieves, X Kelvin Quiñones-Laracuente, Liorimar Ramos-Medina, and Gregory J Quirk. Revisiting the role of infralimbic cortex in fear extinction with optogenetics. 35(8):3607–3615, 2015.

[37] Cyril Herry, Stephane Ciocchi, Verena Senn, Lynda Demmou, Christian Müller, and Andreas Lüthi. Switching on and off fear by distinct neuronal circuits. *Nature*, 454(7204):600–606, 2008.

[38] Taiju Amano, Sevil Duvarci, Daniela Popa, and Denis Pare. The fear circuit revisited : contributions of the basal amygdala nuclei to conditioned fear. *J Neurosci.*, 31(43):15481–15489, 2011.

[39] G J Quirk, C Repa, and J E LeDoux. Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: parallel recordings in the freely behaving rat. *Neuron*, 15(5):1029–1039, 1995.

[40] B. An, I. Hong, and S. Choi. Long-term neural correlates of reversible fear learning in the lateral amygdala. *Journal of Neuroscience*, 32(47):16845–16856, 2012.

[41] J C Repa, J Muller, J Apergis, T M Desrochers, Y Zhou, and J E LeDoux. Two different lateral amygdala cell populations contribute to the initiation and storage of memory. *Nature neuroscience*, 4(7):724–731, 2001.

[42] Taiju Amano, Cagri T Unal, and Denis Paré. Synaptic correlates of fear extinction in the amygdala. *Nature Neuroscience*, 13(4):489–494, 2010.

[43] E. Santini, G. J. Quirk, and J. T. Porter. Fear conditioning and extinction differentially modify the intrinsic excitability of infralimbic neurons. *Journal of Neuroscience*, 28(15):4028–4036, 2008.

[44] M.R. Milad and G.J. Quirk. Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature*, 420(6911):70–74, 2002.

[45] Ivan Vidal-Gonzalez, Benjamín Vidal-Gonzalez, Scott L Rauch, and Gregory J Quirk. Microstimulation reveals opposing influences of prelimbic and infralimbic cortex on the expression of conditioned fear. *Learning & Memory*, 13(6):728–733, 2006.

[46] Robert A; Rescorla and Allan R Wagner. A theory of pavlovian conditioning : variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory*, pages 64–69. Appleton Century Crofts, 1972.

[47] Ralph R Miller, Robert C Barnet, and Nicholas J Grahame. Assessment of the Rescorla-Wagner model. *Psychological bulletin*, 117(3):363–386, 1995.

[48] Richard S Sutton and Andrew G Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[49] Richard S Sutton and Andrew G Barto. Time-derivative models of Pavlovian reinforcement. In *Learning and computational neuroscience: Foundations of adaptive networks*, pages 497–537. MIT Press, 1990.

[50] Elliot A Ludvig, Richard S Sutton, and E James Kehoe. Evaluating the TD model of classical conditioning. *Learning {&} Behavior*, 40(3):305–319, 2012.

[51] J Moren and Christian Balkenius. A computational model of emotional learning in the amygdala. *From animals to animats 6 proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, 32:383, 2000.

[52] Stephane Ciocchi, Cyril Herry, François Grenier, Steffen B. E. Wolff, Johannes J. Letzkus, Ioannis Vlachos, Ingrid Ehrlich, Rolf Sprengel, Karl Deisseroth, Michael B. Stadler, Christian Müller, and Andreas Lüthi. Encoding of conditioned fear in central amygdala inhibitory circuits. *Nature*, 468(7321):277–282, 2010.

[53] Karyn M Myers and Michael Davis. Mechanisms of fear extinction. *Molecular psychiatry*, 12(2):120–150, 2007.

[54] Andrea Barberis and Alberto Bacci. Editorial: Plasticity of GABAergic synapses. *Frontiers in Cellular Neuroscience*, 9:262, 2015.

[55] Jee Hyun Kim and Rick Richardson. Immediate post-reminder injection of gamma-amino butyric acid (GABA) agonist midazolam attenuates reactivation of forgotten fear in the infant rat. *Behavioral Neuroscience*, 121(6):1328–1332, 2007.

[56] Jee Hyun Kim and Rick Richardson. A developmental dissociation of context and GABA effects on extinguished fear in rats. *Behavioral Neuroscience*, 121(1):131–139, 2007.

[57] Kenji Doya. Modulators of decision making. *Nature Neuroscience*, 11(4):410–416, 2008.

[58] Trevor W. Robbins. Arousal systems and attentional processes. *Biological Psychology*, 45(1):57–71, 1997.

[59] Eve Marder. Neuromodulation of neuronal circuits: back to the future. *Neuron*, 76(1):1–11, oct 2012.

[60] W Schultz, P Apicella, and T Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 13(3):900–913, 1993.

[61] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, 299(5614):1898–1902, 2003.

[62] J. N J Reynolds and Jeffery R. Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4-6):507–521, 2002.

[63] P Read Montague, Peter Dayan, and Terrence J Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience*, 16(5):1936–1947, 1996.

[64] Saori C Tanaka, Kenji Doya, Go Okada, Kazutaka Ueda, Yasumasa Okamoto, and Shigeto Yamawaki. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature neuroscience*, 7(8):887–893, 2004.

[65] S. Yagishita, A. Hayashi-Takagi, G. C. R. Ellis-Davies, H. Urakubo, S. Ishii, and H. Kasai. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345:1616–1620, 2014.

[66] Masamoto Yokoyama, Eiji Suzuki, Taku Sato, Shuji Maruta, Shigeru Watanabe, and Hitoshi Miyaoka. Amygdalic levels of dopamine and serotonin rise upon exposure to conditioned fear stress without elevation of glutamate. *Neuroscience Letters*, 379(1):37–41, 2005.

[67] Wolfram Schultz. Multiple dopamine functions at different time courses. *Annual review of neuroscience*, 30:259–288, 2007.

[68] Ethan S. Bromberg-Martin, Masayuki Matsumoto, and Okihide Hikosaka. Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, 68(5):815–834, 2010.

[69] Stephan Lammel, Andrea Hetzel, Olga Häckel, Ian Jones, Birgit Liss, and Jochen Roeper. Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron*, 57(5):760–773, 2008.

[70] Robert P Vertes. Differential projections of the infralimbic and prelimbic cortex in the rat. *Synapse*, 51(1):32–58, 2004.

[71] Sébastien Royer, Marzia Martina, and Denis Paré. An Inhibitory Interface Gates Impulse Traffic between the Input and Output Stations of the Amygdala. *The Journal of Neuroscience*, 19(23):10575 LP – 10583, dec 1999.

[72] Anne Marowsky, Yuchio Yanagawa, Kunihiko Obata, and Kaspar Emanuel Vogt. A Specialized Subclass of Interneurons Mediates Dopaminergic Facilitation of Amygdala Function. *Neuron*, 48(6):1025–1037, dec 2005.

[73] Alexei V Egorov, Klaus Unsicker, and Oliver Von Bohlen und Halbach. Muscarinic control of graded persistent activity in lateral amygdala neurons. *European Journal of Neuroscience*, 24(11):3183–3194, 2006.

[74] Cyril Herry and Nicole Mons. Resistance to extinction is associated with impaired immediate early gene induction in medial prefrontal cortex and amygdala. *European Journal of Neuroscience*, 20(3):781–790, 2004.

[75] U Frey, Y Y Huang, and E R Kandel. Effects of cAMP simulate a late stage of LTP in hippocampal CA1 neurons. *Science*, 260(5114):1661–1664, jun 1993.

[76] Gregory J Quirk, Gregory K Russo, Jill L Barron, and Kelimer Lebron. The role of ventromedial prefrontal cortex in the recovery of extinguished fear. *The Journal of Neuroscience*, 20(16):6225–6231, aug 2000.

[77] Victor A Derkach, Michael C Oh, Eric S Guire, and Thomas R Soderling. Regulatory mechanisms of AMPA receptors in synaptic plasticity. *Nature Reviews Neuroscience*, 8(2):101–113, 2007.

[78] Raphael Lamprecht and Joseph LeDoux. Structural plasticity and memory. *Nature Reviews Neuroscience*, 5(1):45–54, 2004.

[79] John Lisman, Ryohei Yasuda, and Sridhar Raghavachari. Mechanisms of CaMKII action in long-term potentiation. *Nature reviews neuroscience*, 13(3):169–182, 2012.

[80] Julia M Langton and Rick Richardson. The temporal specificity of the switch from NMDAr-dependent extinction to NMDAr-independent re-extinction. *Behavioural brain research*, 208(2):646–649, 2010.

[81] David Anglada-Figueroa and Gregory J Quirk. Lesions of the basal amygdala block expression of conditioned fear but not extinction. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(42):9680–9685, 2005.

[82] Mouna Maroun, Alexandra Kavushansky, Andrew Holmes, Cara Wellman, and Helen Motanis. Enhanced extinction of aversive memories by high-frequency stimulation of the rat infralimbic cortex. *PLoS One*, 7(5):e35853, 2012.

[83] Vitor De Castro Gomes, Laura Andrea León, Daniel Mograbi, Fernando Cardenas, and Jesus Landeira-fernandez. Contextual fear extinction and re-extinction in Carioca high- and low-conditioned freezing rats. *World Journal of Neuroscience*, 4(June):247–252, 2014.

[84] RC Kessler, A Sonnega, E Bromet, M Hughes, and Nelson CB. Posttraumatic stress disorder in the national comorbidity survey. *Archives of General Psychiatry*, 52(12):1048–1060, dec 1995.

[85] Ronald C Kessler, Wai Tat Chiu, Olga Demler, and Ellen E Walters. Prevalence, Severity, and Comorbidity of Twelve-month DSM-IV Disorders in the National Comorbidity Survey Replication (NCS-R). *Archives of general psychiatry*, 62(6):617–627, jun 2005.

[86] Randal A Koene and Michael E Hasselmo. Consequences of parameter differences in a model of short-term persistent spiking buffers provided by pyramidal cells in entorhinal cortex. *Brain Research*, 1202:54–67, apr 2008.

[87] Michael B. VanElzakker, M. Kathryn Dahlgren, F. Caroline Davis, Stacey Dubois, and Lisa M. Shin. From Pavlov to PTSD: the extinction of conditioned fear in rodents, humans, and anxiety disorders. *Neurobiology of Learning and Memory*, 113:3–18, 2014.

[88] Cara Katz, Murray Stein, and J Don Richardson. A review of interventions for treatment-resistant posttraumatic stress d isorder. In *Different views of anxiety disorders*, pages 251–270. 2011.

[89] Steven Taylor, Dana S Thordarson, Louise Maxfield, Ingrid C Fedoroff, Karina Lovell, and John Ogrodniczuk. Comparative efficacy, speed, and adverse effects of three PTSD treatments: exposure therapy, EMDR, and relaxation training. *Journal of consulting and clinical psychology*, 71(2):330–338, 2003.

[90] Scott F Coffey, Bonnie S Dansky, and Kathleen T Brady. Exposure-based, trauma-focused therapy for comorbid posttraumatic stress disorder-substance use disorder. *Trauma and substance abuse: Causes, consequences, and treatment of comorbid disorders*, pages 127–146, 2003.

[91] Joseph A. Boscarino. Post-traumatic stress and associated disorders among Vietnam veterans: the significance of combat exposure and social support. *Journal of Traumatic Stress*, 8(2):317–336, apr 1995.

[92] Harold Kudler. Trauma and the Vietnam War Generation: Report of Findings from the National Vietnam Veterans Readjustment Study. *The Journal of Nervous and Mental Disease*, 179(10):644–645., 1991.

[93] Andrew Y Ng and Michael I Jordan. On Discriminative vs. Generative classifiers: a comparison of logistic regression and naive Bayes. In *Advances in neural information processing systems*, volume 14, page 841. 2002.

[94] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[95] John M Pearce and Geoffrey Hall. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, 87(6):532–552, 1980.

[96] N. J. Mackintosh. A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4):276–298, 1975.

[97] Anthony Dickinson, Geoffrey Hall, and N. J. Mackintosh. Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, 2(4):313–322, 1976.

[98] Nestor A Schmajuk. Computational models of classical conditioning. *Scholarpedia*, 3(3):1664, 2008.

[99] Allan R Wagner. SOP: A model of automatic memory processing in animal behavior. *Information processing in animals: Memory mechanisms*, 85:5–47, 1981.

[100] Stephen Grossberg. A neural model of attention, reinforcement and discrimination learning. *International review of neurobiology*, 18:263–327, 1975.

[101] Nestor A Schmajuk and James J DiCarlo. A neural network approach to hippocampal function in classical conditioning. *Behavioral Neuroscience*, 105(1):82–110, 1991.

[102] A David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, 114(3):784–805, 2007.

[103] Minryung R Song and Jean-Marc Fellous. Value learning and arousal in the extinction of probabilistic rewards: the role of dopamine in a modified temporal difference model. *PloS one*, 9(2):e89494, 2014.

[104] W.-X. Pan, Robert Schmidt, Jeffery R Wickens, and Brian I Hyland. Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *Journal of Neuroscience*, 28(39):9619–9631, sep 2008.

[105] Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37(1):205–220, jul 2014.

[106] C R Gallistel and John Gibbon. Time, rate, and conditioning., 2000.

[107] Sham Kakade and Peter Dayan. Acquisition and extinction in autoshaping. *Psychological Review*, 109(3):533–544, 2002.

[108] John McNamara and Alasdair Houston. The application of statistical decision theory to animal behaviour. *Journal of Theoretical Biology*, 85(4):673–690, 1980.

[109] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006.

[110] Samuel J Gershman, David M Blei, and Yael Niv. Context, learning, and extinction. *Psychological Review*, 117(1):197–209, 2010.

[111] Kevin Lloyd and David S Leslie. Context-dependent decision-making: a simple Bayesian model. *Journal of The Royal Society Interface*, 10(82):20130069, 2013.

[112] J L Armony, D Servan-Schreiber, L M Romanski, J D Cohen, and J E LeDoux. Stimulus generalization of fear responses: effects of auditory cortex lesions in a computational model and in rats. *Cerebral cortex (New York, N.Y. : 1991)*, 7(2):157–165, 1997.

[113] Ahmed A Moustafa, Mark W Gilbertson, Scott P Orr, Mohammad M Herzallah, Richard J Servatius, and Catherine E Myers. A model of amygdala-hippocampal-prefrontal interaction in fear conditioning and extinction in animals. *Brain and Cognition*, 81(1):29–43, 2013.

[114] Christian Balkenius and Jan Morén. Emotional learning: a computational model of the amygdala. *Cybernetics and Systems*, 32(6):611–636, 2001.

[115] Guoshi Li, Taiju Amano, Denis Pare, and Satish S Nair. Impact of infralimbic inputs on intercalated amygdala neurons: A biophysical modeling study. *Learning & Memory*, 18(4):226–240, mar 2011.

[116] Sandeep Pendyam, Christian Bravo-Rivera, Anthony Burgos-Robles, Francisco Sotres-Bayon, Gregory J Quirk, and Satish S Nair. Fear signaling in the prelimbic-amygdala circuit: a computational modeling and recording study. *Journal of neurophysiology*, 110(4):844–61, 2013.

[117] Dongbeom Kim, D. Pare, and Satish S Nair. Mechanisms contributing to the induction and storage of Pavlovian fear memories in the lateral amygdala. *Learning & Memory*, 20(8):421–430, jul 2013.

[118] Guoshi Li, Satish S Nair, and Gregory J Quirk. A biologically realistic network model of acquisition and extinction of conditioned fear associations in lateral amygdala neurons. *Journal of neurophysiology*, 101(3):1629–1646, 2009.

[119] Ioannis Vlachos, Cyril Herry, Andreas Lüthi, Ad Aertsen, and Arvind Kumar. Context-dependent encoding of fear and extinction memories in a large-scale network model of the basal amygdala. *PLoS Computational Biology*, 7(3):e1001104, 2011.

Acknowledgements

This dissertation was made possible by the intensive collaboration of my advisors, Prof. Honda and Prof. Matsuda. They provided just the right balance of freedom and guidance for me to thrive in Kyoto University. I would like to thank them as well as Dr. Nakae who helped me develop my statistical models, and Prof. Ishii who advised during the preparation of my manuscripts.

This thesis is based on material contained in the following scholarly paper.

Li Y, Nakae K, Ishii S, Naoki H (2016)

Uncertainty-Dependent Extinction of Fear Memory in an Amygdala-mPFC Neural Circuit Model.

PLoS Comput Biol 12(9): e1005099. doi:10.1371/journal.pcbi.1005099