A neurocomputational model of the mammalian fear conditioning circuit

by

Carter Kolbeck

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, I present a computational neural model that reproduces the high-level behavioural results of well-known fear conditioning experiments: first-order conditioning, second-order conditioning, sensory preconditioning, context conditioning, blocking, first-order extinction and renewal (AAB, ABC, ABA), and extinction and renewal after second-order conditioning and sensory preconditioning. The simulated neural populations used to account for the behaviour observed in these experiments correspond to known anatomical regions of the mammalian brain. Parts of the amygdala, periaqueductal gray, cortex and thalamus, and hippocampus are included and are connected to each other in a biologically plausible manner.

The model was built using the principles of the Neural Engineering Framework (NEF): a mathematical framework that allows information to be encoded and manipulated in populations of neurons. Each population represents information via the spiking activity of simulated neurons, and is connected to one or more other populations; these connections allow computations to be performed on the information being represented. By specifying which populations are connected to which, and what functions these connections perform, I developed an information processing system that behaves analogously to the fear conditioning circuit in the brain.

Acknowledgements

First and foremost, I would like to thank my supervisor Chris Eliasmith. His dedication to his field and passion for his work have left an indelible impression on me.

I also want to thank all the guys in the lab for making the CNRG a great place to work. I especially want to thank Terry Stewart, Trevor Bekolay, Xuan Choo, Travis DeWolf, and Daniel Rasmussen for their willingness to answer my questions, lend a hand, and offer suggestions.

Dedication

To my family and friends who have provided me with a stimulating intellectual environment throughout my life.

Table of Contents

List of Figures **ix**

1 Introduction **1**

1.1 Outline of thesis . . . 2

2 Fear conditioning **4**

2.1 Classical conditioning . . . 4

2.1.1 First-order cued fear conditioning . . . 5

2.1.2 First-order context fear conditioning . . . 6

2.2 Second-order conditioning . . . 7

2.3 Sensory preconditioning . . . 7

2.4 Blocking . . . 8

2.5 Extinction . . . 9

2.5.1 Context-dependent extinction and renewal . . . 9

2.5.2 Extinction in second-order conditioning and sensory preconditioning 11

2.6 Other phenomena related to fear conditioning . . . 15

3 Anatomy of the fear conditioning circuit **16**

3.1 Amygdala . . . 16

3.2 Periaqueductal Gray . . . 18

3.3 Hippocampus . . . 18

3.4 Sensory cortex 19

3.5 Thalamus 19

3.6 Prefrontal cortex 20

4 Other models of fear conditioning **21**

4.1 Grossberg and Levine 21

4.2 Balkenius and Morén 21

4.3 Vlachos et al. 22

4.4 Li et al. 25

4.5 Krasne et al. 26

4.6 Comparisons of the models 27

5 Neurons and plasticity **28**

5.1 Neurons 28

5.2 Plasticity 31

5.2.1 Hebbian plasticity 31

5.2.2 Neuromodulatory learning 33

6 Large-scale neural modelling with the NEF **35**

6.1 Single neuron model 35

6.2 Representation 36

6.3 Transformation 42

6.4 Learning 47

6.4.1 Error minimization 47

6.4.2 Hebbian learning 49

6.4.3 Combined learning rule 50

7 A new model of fear conditioning **52**

7.1 Implementation 52

7.2 Model description and simulations 55

7.2.1 First-order conditioning 55

7.2.2 Second-order conditioning 58

7.2.3 Sensory preconditioning 58

7.2.4 Blocking 62

7.2.5 Context conditioning 64

7.2.6 Extinction and renewal 67

8 Discussion and conclusions **75**

References **78**

List of Figures

2.1 A demonstration of a first-order cue conditioning experiment. 6

2.2 A demonstration of a context conditioning experiment. 7

2.3 A demonstration of a second-order conditioning experiment. 8

2.4 A demonstration of a sensory preconditioning experiment. 9

2.5 A demonstration of a blocking experiment. 10

2.6 A demonstration of an AAB extinction experiment. 11

2.7 A demonstration of an ABC extinction experiment. 12

2.8 A demonstration of an ABA extinction experiment. 12

2.9 A demonstration of an extinction after second-order conditioning experiment. 13

2.10 A demonstration of an extinction after sensory preconditioning experiment. 14

3.1 Anatomical regions of the fear conditioning circuit. 17

4.1 Grossberg and Levine model. 22

4.2 Balkenius and Morén model. 23

4.3 Vlachos et al. model. 24

4.4 Li et al. model. 25

4.5 Krasne et al. model. 26

5.1 Illustration of a neuron. 29

5.2 Illustration of a synapse. 30

5.3 Illustration of one of the consequences of Hebbian learning. 32

5.4 Illustration of neuromodulatory learning. 33

6.1 Circuit that models a leaky integrate and fire neuron. 37
6.2 Tuning curves for a population representing a one-dimensional input, **x**. . . 40
6.3 PSC, spikes, and filtered spikes. 43
6.4 A communication channel between two populations of neurons. 44
6.5 A neural population receiving inputs from two neural populations. 46
6.6 A general error-modulated learning circuit. 48

7.1 The proposed model. 53
7.2 Learning sub-circuit for LA. 57
7.3 Simulation results for a first-order fear conditioning experiment. 59
7.4 Simulation results for a second-order fear conditioning experiment. 60
7.5 Learning circuit for CS and CSi populations. 61
7.6 Simulation results for a sensory preconditioning fear conditioning experiment. 63
7.7 Simulation results for a blocking fear conditioning experiment. 65
7.8 Simulation results for a context fear conditioning experiment. 66
7.9 Learning circuit for context-dependent extinction. 67
7.10 Simulation results for an AAB extinction and renewal experiment. 69
7.11 Simulation results for an ABC extinction and renewal experiment. 70
7.12 Simulation results for an ABA extinction and renewal experiment. 71
7.13 Simulation results for a second-order conditioning extinction and renewal experiment. 73
7.14 Simulation results for a sensory preconditioning extinction and renewal experiment. 74

Chapter 1

Introduction

The modern mammalian brain is built upon ancient emotional subsystems [53]. These systems, which come hardwired at birth, ensure that animals are able to find food, avoid predators, reproduce, and generally meet their basic needs [53]. Recently in the evolution of mammals, the development of the cortex enabled more complex sensory processing, motor planning, language, and so on. These faculties developed, not separately from the emotional subsystems of the brain, but rather, alongside them [53].

Throughout their day to day activities, mammals answer to emotional subsystems, even though it may be through complex networks of cortical associations. These subsystems are ultimately what provide the motivational signals that guide learning and behaviour [64]. If we hope to understand how the brain works and how it developed, theoretical neuroscientists need to develop a deeper understanding of the mechanisms of emotion, and how they affect the rest of the brain.

In this thesis, I examine one of the emotional subsystems, the fear system, using a detailed neurocomputational model. The fear system has been extensively studied; the key brain regions involved have been identified, and theories have been developed regarding their functions [37]. It is a particularly interesting case study because of how well linked it is to learning and memory; fearful experiences have a strong effect on an animal's future behaviour [37]. This link provides a potential avenue of investigation into how low-level reward and punishment systems affect higher-level cognitive functions.

Building a computational model to simulate the fear system forces us to integrate a wide variety of neuroscientific and behavioural evidence to tell a coherent story about what specific neural mechanisms may be at work during fear-related behaviour. It also provides a medium for us to explore new theories, which can be incorporated into the model and eventually tested by empirical experiments.

A good model should be able to replicate data from experiments performed on animals. The fear system has a wide range of related behavioural data that comes from a collection of ‘fear conditioning’ experiments, which show how animals learn that certain environmental stimuli come to predict aversive events [43]. The primary contribution of this thesis is to demonstrate a model that can replicate a wide variety of the findings from these experiments, suggesting that the specific mechanisms proposed in this thesis are plausible candidates for those found in the biological systems under study.

I begin by providing background information on fear conditioning experiments, the anatomy of the fear conditioning circuit, the mechanisms of computation in the brain, and the modelling techniques employed. These sections provide the methods and empirical constraints used for the design of the model, and give insights into how and why it was developed.

1.1 Outline of thesis

Chapter 2, Fear conditioning, introduces the topic of fear conditioning. The results of several classic fear conditioning experiments are presented.

Chapter 3, Anatomy of the fear conditioning circuit, discusses the regions of the brain involved in the fear conditioning circuit. Their function, as it is related to known fear conditioning experiments, is outlined.

Chapter 4, Other models of fear conditioning, reviews a selection of previously developed fear conditioning models. The function and implementation of each are outlined, and comparisons are made to the model presented in this thesis.

Chapter 5, Neurons and plasticity, discusses the basic biological foundations of computation used to inform the development of the model. Neurons, synaptic weights, plasticity and their basic mathematical characterizations are reviewed.

Chapter 6, Large-scale neural modelling with the NEF, provides an overview of the computational methods used to build the model. It describes the Neural Engineering Framework: a mathematical framework that allows information to be represented and transformed in populations of spiking neurons. Learning in the context of the NEF is also discussed, as is the single neuron model that is used in the simulations.

Chapter 7, A new model of fear conditioning, presents the novel work of this thesis, the NEF model of fear conditioning. The function of the model, as well as design justifications, are detailed in the context of the fear conditioning experiments with which it was tested. The results of simulations run using the model are plotted and discussed.

Chapter 8, Discussion and conclusions, provides an analysis of the pros and cons of the model and sets the stage for future work on the project.

Chapter 2

Fear conditioning

2.1 Classical conditioning

Classical conditioning is a type of learning that involves the formation of associations between a neutral stimulus and a stimulus that has inherent relevance to an animal [55]. The neutral stimulus (NS) can be any sensory stimulus such as an auditory, visual, olfactory, or tactile cue, that does not have a significant emotional meaning to the animal. The stimulus with inherent emotional meaning to the animal is called the unconditioned stimulus (US), and can be a cue related to food, sex, danger, etc. to which the animal has a biologically coded response called the unconditioned response (UR). Unconditioned responses include behaviours such as salivation (for a cue related to food) or increased heart rate and dilated pupils (for a cue related to fear).

Associations between the NS and the US are formed through temporal pairing. An NS that is paired with a US such that the presence of the NS comes to predict a US will acquire an emotional relevance to the animal. After becoming emotionally relevant, the NS is considered a conditioned stimulus (CS), and will cause a biological response - called a conditioned response (CR) - in the animal that is related to the US to which the CS was paired. In the literature, and in this thesis, the NS is referred to as a CS both before and after association with a US.

Perhaps the most well-known examples of classical conditioning are the experiments performed by Ivan Pavlov [55]. In his experiments, Pavlov noted that when presented with the sight of food, dogs began to salivate. Here, the food is the US, and salivation is the UR. The presentation of food was then repeatedly paired with a bell (the CS); i.e., the dogs were called to their food by the sound of a bell. After several repetitions of this procedure, the dogs began to salivate at the sound of the bell even if no food was presented. Thus, after pairing of the CS (bell) with the US (food), the CS evoked a CR related to the behaviour evoked by a US.

The field of classical conditioning has been widely studied in psychology and neuroscience since Pavlov's initial experiments [46]. One experimental paradigm that has proven particularly fruitful in the study of classical conditioning is fear conditioning. Some of the experimental results of the study of fear conditioning are discussed in the following sections.

2.1.1 First-order cued fear conditioning

Fear conditioning is a subset of classical conditioning that involves the association between CSs and USs that evoke behaviours associated with fear. One well-known (and ethically controversial) fear conditioning experiment was performed by John Watson in 1919 [73]. In his experiment, Watson taught an infant (known as 'Albert B.') to fear a white rat. Initially the infant was presented with a variety of animals (including the white rat) and inanimate objects towards which he showed no signs of fear. The infant was given the opportunity to play with the rat, during which time he would reach out to touch it. This initial phase of the experiment established that the rat was not inherently frightening to the child, and could thus be considered a neutral stimulus.

Subsequently, during periods of interaction with the rat, an experimenter would make a loud sound behind the infant by striking a piece of metal with a hammer. The infant reacted to the noise by crying and showing signs that he was afraid. This established the noise as an unconditioned stimulus. After repeatedly pairing the disturbing noise with periods of interaction with the rat, the infant came to show signs that he was afraid when the rat was presented in the absence of the noise. He came to respond to the rat by crying and turning, and moving away from it.

Although some doubt the validity of this particular experiment [27], fear conditioning experiments have been replicated many times since (e.g. [11], [63], [17]).

The typical cued fear conditioning experiments - and all of those replicated by the model described in this thesis - use a rat as a subject [66]. Rats are widely used in many neuroscientific experiments; they are easy to breed, easy to handle, and have a similar neuroanatomy to humans - especially regarding the brain regions involved in the fear circuit [1]. The cue (CS) in these experiments is usually visual or auditory: for example a light or a tone. The US is an electric shock applied to the foot of the rat. The UR to

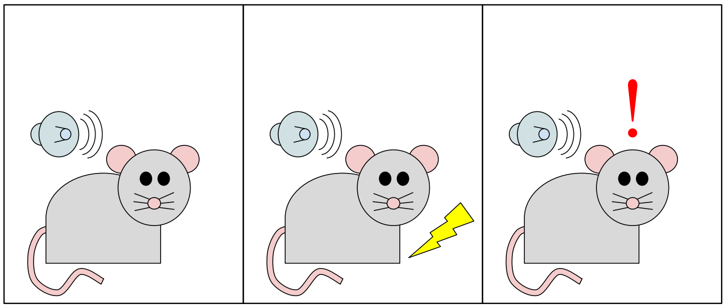

Figure 2.1: A demonstration of a first-order cue conditioning experiment. Frame 1: the tone does not elicit a fear response. Frame 2: the tone is then paired with a footshock. Frame 3: finally, the tone elicits a fear response.

a shock is typically an attempt to escape it. The CR that is typically measured as a sign that an association between the CS and US has been made is 'freezing' (see figure 2.1).

Freezing is a natural fear response seen in rats. It is characterized by a period of watchful immobility during which the rat stops moving ('freezes'), and increases alertness [12]. Dilated pupils as well as increased breathing and heart rate are also present during freezing. In the wild, freezing helps the animal go undetected by predators, and also prepares them for flight or fight responses. In fear conditioning experiments, freezing is used as the conditioned response as it is very easy to observe, and has been shown to be a reliable indicator of learned fear [11].

2.1.2 First-order context fear conditioning

As described above, humans and other animals demonstrate fear conditioning through associations between specific cues and USs; however, animals also demonstrate fear conditioning through associations between contexts and USs. Although a context might be thought of as a collection of cues, the literature on fear conditioning generally draws a distinction between a cue and a context [58].

The typical context experiment involves placing a rat in an environment, such as a box or cage, and applying an aversive US, such as a footshock [9]. The association between the US and the context can be learned either after a single trial or after multiple trials. The rat is then removed from the environment. After being placed back into the environment, the rat will demonstrate a fear response (see figure 2.2).

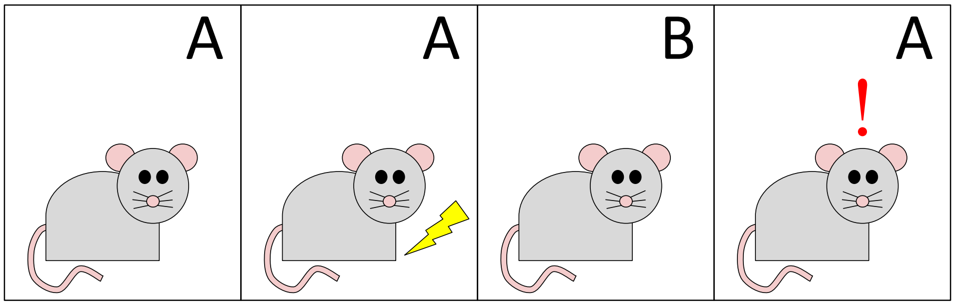

Figure 2.2: A demonstration of a context conditioning experiment. Frame 1: being in context A does not elicit a fear response. Frame 2: the rat is given footshocks in context A. Frame 3: the rat is moved to context B. Frame 4: returning to context A elicits a fear response.

2.2 Second-order conditioning

Direct pairing between CSs and contexts and USs is considered first order pairing; a neutral cue or context is associated directly with the aversive stimulus. However, once a CS (call it CS1) has been associated with a US, such that it is capable of eliciting a fear response, a second CS (CS2) that is paired with CS1 can also come to elicit a fear response in the absence of CS1 and the US [25].

A standard experiment used to demonstrate this phenomenon is performed as follows [63]. First a CS1, say a tone, is paired with with a footshock until CS1 comes to elicit the fear response in the absence of the footshock. Following that, the tone is played to the rat at the same time as a light (CS2) is presented. During the pairing of the light and the tone, the rat comes to associate the light with the fear that is associated with the tone. Subsequent presentations of the light on its own will elicit a fear response (see figure 2.3).

2.3 Sensory preconditioning

Second-order conditioning is considered higher-order conditioning: a CS gains affective significance without being paired directly with a US. Another form of higher-order conditioning is sensory preconditioning. The distinction between second-order conditioning and sensory preconditioning is that second-order conditioning involves pairing between two CSs

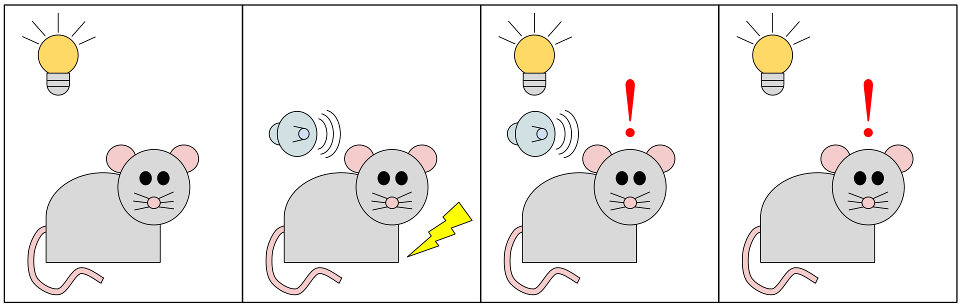

Figure 2.3: A demonstration of a second-order conditioning experiment. Frame 1: the light does not elicit a fear response. Frame 2: a neutral tone is paired with a footshock. Frame 3: the light is paired with the tone (the tone elicits a fear response). Frame 4: the light elicits a fear response on its own.

after training between a CS and US has taken place, while sensory preconditioning involves pairing between two CSs before any training with a CS and US has taken place [25].

A standard sensory preconditioning experiment is performed as follows [63] . A CS1, for example a tone, becomes associated with a CS2, for example a light, through pairing of the two stimuli. Next, the tone is paired with a footshock (US) until the tone comes to elicit a fear response when presented in the absence of the footshock. Subsequently, when the light is presented, it elicits a fear response through its association with CS1 (see figure 2.4).

2.4 Blocking

In the previous examples, pairing between a CS and a US is able to give affective significance to the CS. However, under certain conditions this association does not occur; one commonly studied phenomenon that prevents this association is called blocking. Blocking occurs when a CS that has been trained to elicit a fear response is present while a neutral stimulus is being paired with a US. The strong CS overshadows the neutral stimulus and blocks the conditioning of the weaker neutral stimulus by the US [31]. An explanation for this phenomenon is that the animal's attention is directed at the strong CS, which has affective significance, and the neutral stimulus therefore does not have the salience required for learning [62].

A standard blocking experiment is performed as follows [31]. CS1, a tone, is paired

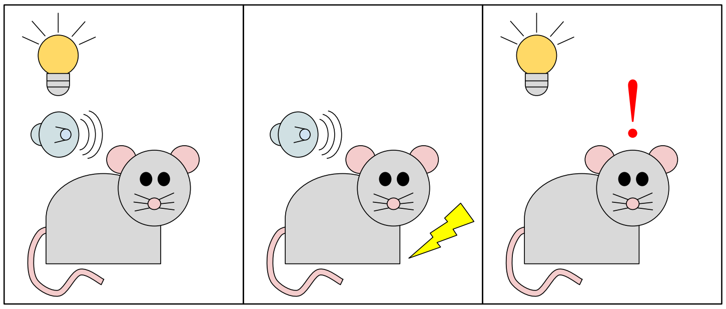

Figure 2.4: A demonstration of a sensory preconditioning experiment. Frame 1: the light and the tone are paired together. Frame 2: the tone is paired with a footshock. Frame 3: the light (through its association with the tone and the tone's association with the footshock) elicits a fear response.

with a US until CS1 is able to elicit a fear response on its own. While CS1 is present, CS2, a light, is introduced and paired with a US. In subsequent presentations of CS2, the animal will not demonstrate a fear response despite the pairing between CS2 and the US (see figure 2.5).

2.5 Extinction

In the study of fear conditioning, extinction refers to a conditioned stimulus losing its affective significance. As seen in the experiments discussed above, a CS gains the ability to elicit a fear response after being paired with a US. However, there are mechanisms in the brain that prevent a CS that once evoked a fear response from evoking future fear responses.

2.5.1 Context-dependent extinction and renewal

A CS with affective significance can lose its ability to evoke a fear response if it is repeatedly presented in the absence of a US. While conditioning can be explained as the animal learning to associate a CS with a US, this type of extinction can be explained as the animal learning to associate a CS that previously evoked a fear response with the absence of a US [61]. Critically, this process seems to be dependent on the context in which the extinction occurs. If a CS with affective significance is extinguished in one context, it

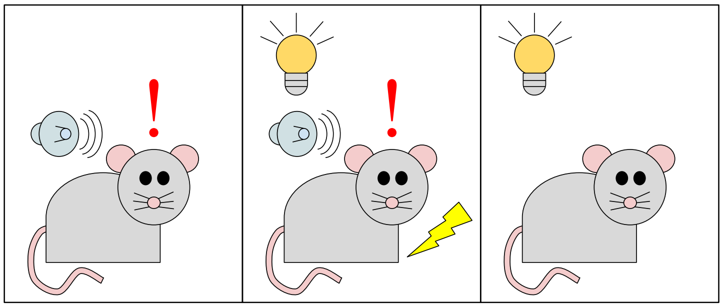

Figure 2.5: A demonstration of a blocking experiment. Frame 1: a conditioned tone elicits a fear response. Frame 2: a light is paired with a footshock in the presence of the tone. Frame 3: the light does not elicit a fear response because pairing between the light and the footshock was blocked by the tone.

can maintain its ability to elicit a fear response in another context [10]. Because of this observation, it is thought that extinction is an active learning process (between a CS and the absence of a US) as opposed to an unlearning of the association between a CS and a US. The ability for an extinguished CS to elicit a fear response in a context other than the one in which extinction occurred is called renewal.

A standard context-dependent extinction/renewal experiment is performed as follows [71]. In this AAB renewal experiment, a CS, such as a tone, is paired with a footshock so that it comes to elicit a fear response. While in the same environment, the tone is repeatedly presented to the rat in the absence of the footshock. After some time, fear responding to the tone in this context will cease. Then the rat is moved to a new environment, context B, and presented with the tone cue again. In the new environment the tone cue is again able to elicit a fear response (see figure 2.6).

Another experimental setup demonstrates ABC renewal [71]. First a CS, for example a tone, is paired with a footshock (US) so that it comes to elicit a fear response. The rat is then placed in a new environment. In this environment the tone is presented to the rat repeatedly (in the absence of a US). Fear responding to the tone will decrease during this time. Subsequently the rat is placed in another environment. When the tone is presented in this environment, it elicits a fear response. This phenomenon is referred to as ABC renewal because of the three different contexts involved (see figure 2.7).

In ABA renewal, the ‘AB’ part of the experiment proceeds as described with ABC renewal [71]. However, instead of being placed in a new context, C, after extinction in B the rat is placed back into context A, and fear responses to the tone cue are renewed (see

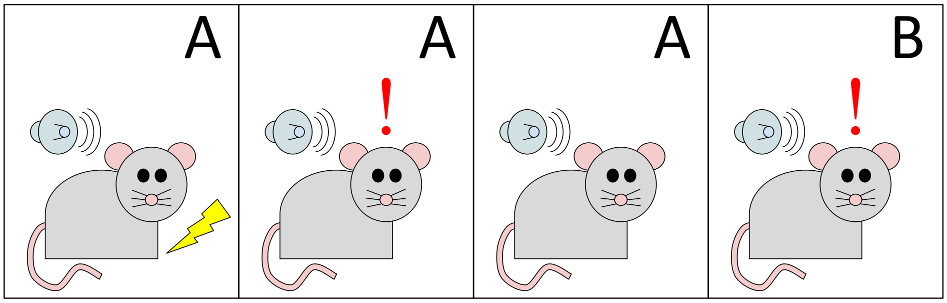

Figure 2.6: A demonstration of an AAB extinction experiment. Frame 1: a tone is paired with a footshock in context A. Frame 2: the tone elicits a fear response in context A. Frame 3: after repeatedly being played in the absence of a footshock, the tone stops eliciting a fear response in context A. Frame 4: the tone is able to elicit a fear response in a new context, B.

figure 2.8).

2.5.2 Extinction in second-order conditioning and sensory preconditioning

The experiments described above demonstrate the effect of extinction on a CS that was trained through direct pairing with a US. In the case of higher-order conditioning, extinction has an effect on both the CS1 (the CS trained through direct pairing with the US) and the CS2 (the CS trained through pairing with CS1). Extinction has a different effect on a CS2 after second-order conditioning than it does on a CS2 after first-order conditioning.

Extinction of a CS1 after second-order conditioning does not extinguish the CS2 with which it was trained [54]. This follows from the observation that, in second-order conditioning, a CS2 is associated with the fear response evoked by CS1 as opposed to CS1 directly (see figure 2.9).

On the other hand, extinction of a CS1 after sensory preconditioning can extinguish the CS2 with which it was trained [54]. This follows from the observation that, in sensory preconditioning, CS2 is associated directly with CS1, which only later becomes associated with fear (see figure 2.10).

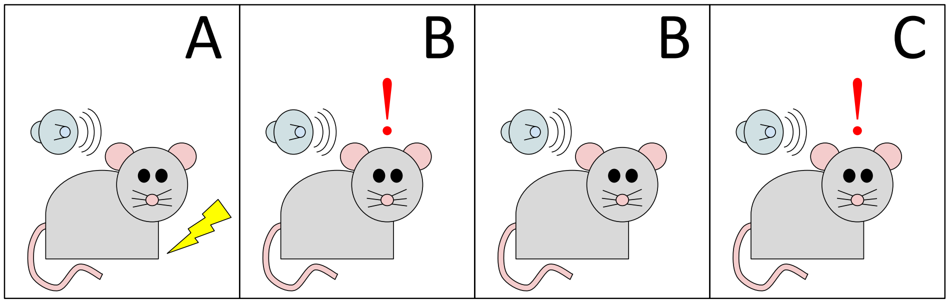

Figure 2.7: A demonstration of an ABC extinction experiment. Frame 1: a tone is paired with a footshock in context A. Frame 2: the tone elicits a fear response in context B. Frame 3: after repeatedly being played in the absence of a footshock, the tone stops eliciting a fear response in context B. Frame 4: the tone is able to elicit a fear response in a new context, C.

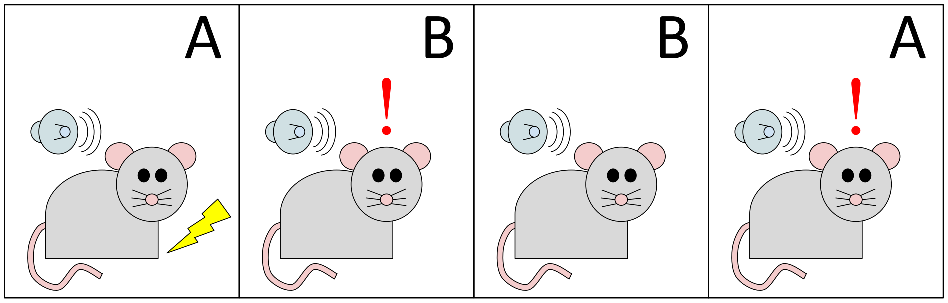

Figure 2.8: A demonstration of an ABA extinction experiment. Frame 1: a tone is paired with a footshock in context A. Frame 2: the tone elicits a fear response in context B. Frame 3: after repeatedly being played in the absence of a footshock, the tone stops eliciting a fear response in context B. Frame 4: the tone is able to elicit a fear response in the original context, A.

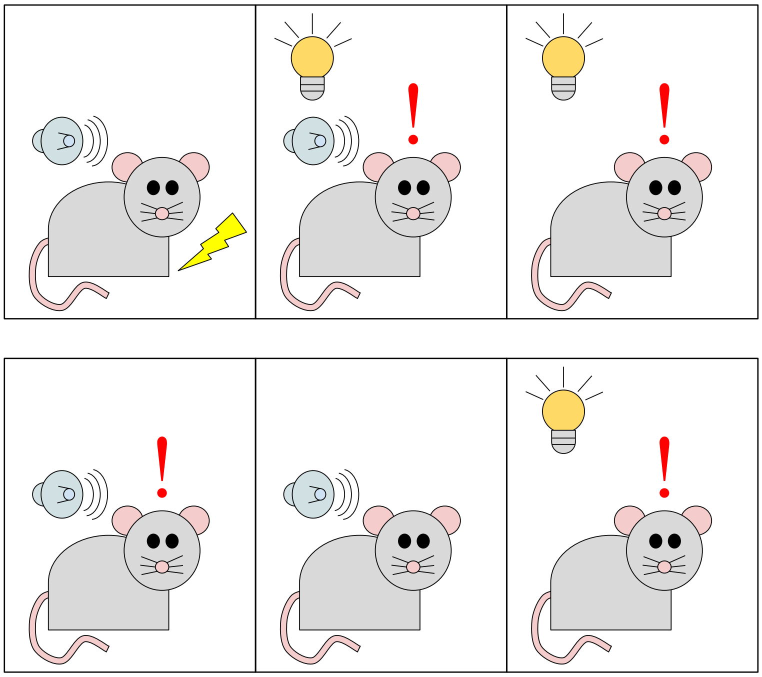

Figure 2.9: A demonstration of an extinction after second-order conditioning experiment. Frame 1: a tone is paired with a footshock. Frame 2: the tone is paired with a light. Frame 3: the light elicits a fear response. Frame 4: the tone elicits a fear response. Frame 5: after repeatedly being played in the absence of a footshock, the tone stops eliciting a fear response. Frame 6: the light still elicits a fear response.

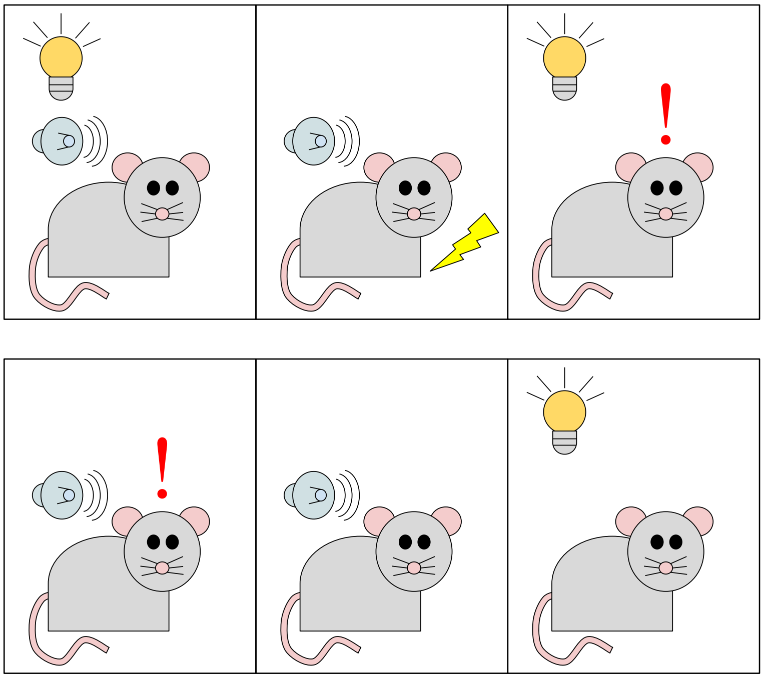

Figure 2.10: A demonstration of an extinction after sensory preconditioning experiment. Frame 1: a tone is paired with a light. Frame 2: the tone is paired with a footshock. Frame 3: the light elicits a fear response. Frame 4: the tone elicits a fear response. Frame 5: after repeatedly being played in the absence of a footshock, the tone stops eliciting a fear response. Frame 6: the light no longer elicits a fear response.

2.6 Other phenomena related to fear conditioning

The phenomena discussed in this chapter are some of the most well-known related to fear conditioning, and are thus the focus of the model discussed in this thesis. However, the experiments performed related to fear conditioning have produced many interesting results in addition to those discussed. The model presented in this thesis does not account for all of these results.

For example, as discussed above, presentation in a new context is enough to renew a response to a CS after extinction. Another form of renewal, in which the CS regains affective significance, occurs when an extinguished CS is later paired with a US [13]. This type of renewal appears to be context independent.

As well, a significant factor in how well a CS is conditioned is the order and timing of presentation of the CS relative to the US. In this chapter, a CS was said to be conditioned through 'temporal pairing' of a US. In reality, temporal pairing does not describe the requisite condition well enough. How well a CS is conditioned is dependent on the time interval between presentation of the CS and the US [70]. If a CS precedes a US by too long of a time period, conditioning will proceed slowly. Similarly, if a CS precedes a US by too short of a time period, or occurs after the presentation of the US, conditioning will proceed slowly. There appears to be a 'sweet spot' in which the CS is learned as a predictor of the US and in which conditioning proceeds the fastest. Testing the presented model to determine its detailed temporal properties is left for future work.

Chapter 3

Anatomy of the fear conditioning circuit

The fear conditioning circuit model described in this thesis includes multiple regions of the mammalian brain. A brief overview of the structure and function of these brain regions is given in this chapter. Figure 3.1 shows simplified connections between the relevant brain regions; a more detailed circuit diagram will be presented in chapter 7.

3.1 Amygdala

The amygdala is a small, almond-sized region found in each hemisphere of the brain. It is part of what would be considered the 'old brain' of mammals; analogous regions are found in distant relatives such as birds and reptiles [29]. It has long been associated with learning and memory involved in emotional processes - especially fear. Much is known about the function of three main sub-regions of the amygdala [43]: the lateral amygdala (LA), the lateral basal area (BL), and the medial central nucleus (CEm), all of which are included in this model.

The lateral amygdala has been found to respond to both conditioned and unconditioned stimuli [69]. It is this convergence of CSs and USs that suggests that the LA is involved in forming the associations required for cued fear learning. Lesion studies (in which the region is intentionally damaged) reported in [38] support the critical role of this brain region in fear learning. The type of learning that occurs in the LA is thought to involve both neuromodulatory and Hebbian processes, which will be discussed in section 5.2 [69].

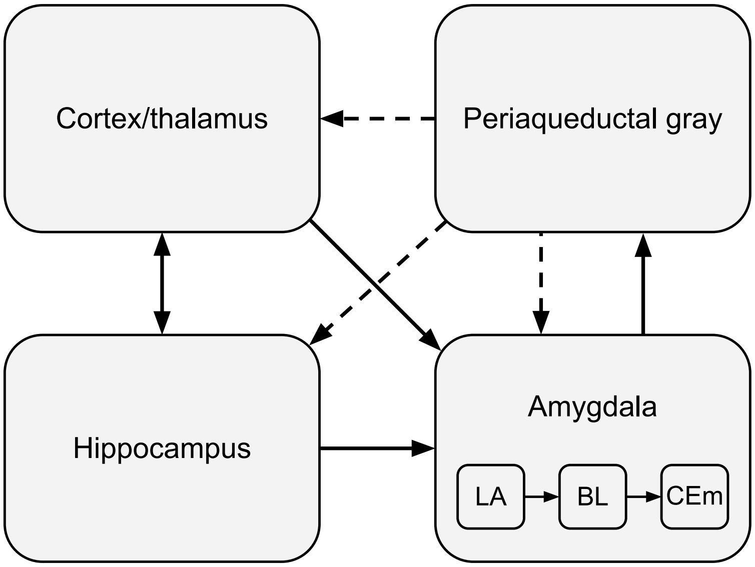

Figure 3.1: Anatomical regions of the fear conditioning circuit that are being modelled here. The dashed lines coming from the periaqueductal gray are to signify that these are generally reinforcement signals facilitating learning in the other regions. Note that these signals may project to their target areas indirectly: going first through other brain regions. The sub-regions of the amygdala are the lateral amygdala (LA), lateral basal area (BL), and the medial central nucleus (CEm).

BL has also been implicated in the formation of associations between conditioned and unconditioned stimuli. Specifically, the BL is thought to play a role in contextual learning as it receives significant projections from the hippocampus [44]. The BL also serves as a path through which information from the LA reaches the CEm. Krasne et al. also propose that there may be an alternate route to the CEm from the LA, but that route is not included in this model [35].

The CEm is thought to drive fear responses [20]. It receives input from BL and LA and its activity is correlated with fearful behaviour in animals. Lesions of the CEm interfere with nearly every measurable fear response including freezing [36]. This is likely because the output of the CEm goes on to the periaqueductal gray: a region responsible for the autonomic responses associated with many fear responses [40].

3.2 Periaqueductal Gray

There are many potential routes for the affective value of USs to reach the amygdala; one proposed route is through the periaqueductal gray (PAG) [35]. Because of its proposed access to US information, in this model PAG is responsible for the reinforcement signals that project - at least in part, indirectly [35] - to the other regions to facilitate learning. The control of these signals - also proposed to occur in PAG [35] - enables more complex learning behaviours such as blocking, extinction, and second-order conditioning.

PAG has another distinct, yet crucial, role in the fear conditioning circuit. Experiments have implicated PAG in various autonomic processes including cardiovascular control, vocalization, and those related to fear and anxiety [4]. In the context of the fear conditioning circuit, PAG is thought to be responsible for initiating fear responses such as freezing [23]. These processes are initiated by activity in the CEm.

3.3 Hippocampus

The hippocampus has long been implicated in learning and memory, especially related to contextual information; it is thought to represent relationships between various stimuli as opposed to individual cues such as tones and lights [51]. The hippocampus provides input to the BL where associations between context and unconditioned stimulus are thought to occur [43]. Studies have implicated the region in fear conditioning; in particular, in animals with hippocampal lesions, fear responding elicited by contexts is attenuated, but fear responses elicited by tones are not [68].

There is also evidence that implicates the hippocampus in the process of context-dependent extinction and renewal of cues. Cells in the rat hippocampus have been identified that respond to cues only when the rat is in a particular context [49]. Based on this evidence, the model includes a circuit in the hippocampus that associates cues with the contexts in which they were presented.

3.4 Sensory cortex

The cortex is involved in many higher-level neural processes, such as the formation of complex relationships between stimuli and concepts [33]. The cortex is an evolutionarily new brain region that experienced expansion well after limbic structures like the amygdala were established [18]. It is often classified into many sub-regions; however, for the model presented here, we are only interested in the outputs of the sensory cortex (processed sensory information) that reach the amygdala [39]. These outputs provide the amygdala with the CSs involved in fear conditioning.

3.5 Thalamus

One of the roles of the thalamus is to relay sensory information to the amygdala for processing in fear conditioning [43]. While the cortex is implicated in processing sensory information, the thalamus provides a direct pathway to the fear conditioning circuit for CSs. Joseph Ledoux has referred to the differences between these two pathways as the high road (cortex) and low road (thalamus) [37]. The cortical pathway is slower but provides more complex representations, whereas the thalamic pathway is faster and bypasses the cortical route to the amygdala. Emotionally relevant visual or auditory cues are therefore able to engage the fear conditioning circuit quickly: something that can be very advantageous for an animal upon seeing or hearing a predator for example.

For the model presented here, we do not distinguish between these two pathways; CS information comes from one neural population which can be thought of as either the cortex or thalamus.

3.6 Prefrontal cortex

The prefrontal cortex (PFC) is thought to be involved in the normal function of the conditioning circuit; however, its role may be a supporting one, facilitating proper functioning of memory consolidation and retrieval between the hippocampus and other cortical areas [35]. These higher-level functions are thought to affect the conscious processing of fear memories in humans [7]. Because of its proposed role mainly in higher-level functions, the PFC is not included in the model presented here.

Chapter 4

Other models of fear conditioning

Before discussing the theory behind the novel NEF model of the fear conditioning circuit, the following presents a short review of other models related to fear conditioning that have been previously developed. Following the summaries of these models, we will look at how the model presented in this thesis compares to the previous approaches.

4.1 Grossberg and Levine

Grossberg and Levine developed a mathematical model of the interactions of a CS1, CS2, and US [26]. The model accounts for first and second-order conditioning, and is able to reproduce experimental data showing the relationship between learning efficacy and the time between the presentation of a CS and a US. The model also accounts for the formations of two types of memories: short term memories and long term memories.

The activity of each of the nodes in figure 4.1 is represented by a differential equation. There are no biologically realistic neurons in the model; rather, it is an abstract mathematical model that accounts for the behaviours observed in the first and second-order conditioning and blocking experiments discussed earlier, as well as other lower-level results.

4.2 Balkenius and Morén

Balkenius and Morén developed a computational model of emotional learning which includes mathematical descriptions of the amygdala, orbitofrontal cortex, sensory cortex,

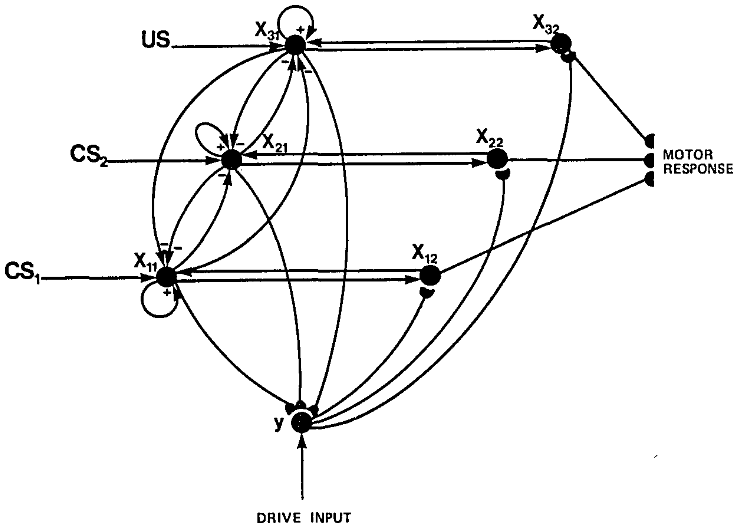

Figure 4.1: Grossberg and Levine model from [26]. x_{ij} and y denote the different computational nodes. The projections terminating with semi-disks are plastic, and allow learning to occur in the model. See cited paper for details.

and thalamus [2]. A focus of the model is the interaction between the orbitofrontal cortex and the amygdala, with the orbitofrontal cortex inhibiting incorrect emotional responses of the amygdala. The model reproduces phenomena related to fear conditioning including extinction and blocking, and is used to investigate the effects of lesions to parts of the model.

The information processing units of the model are generalized nodes (circles in figure 4.2) that are connected with excitatory, inhibitory, or plastic connections. Because it lacks neural detail, the authors say that "[t]he model should be considered at a functional rather than at a neuronal level" [2].

4.3 Vlachos et al.

Vlachos et al. developed a large-scale neural network model that explains how contextual information might affect learning in the amygdala [72]. It is primarily a model of the basal nucleus of the amygdala, receiving inputs from the lateral amygdala and the medial prefrontal cortex. The model reproduces behavioural results related to cued and context

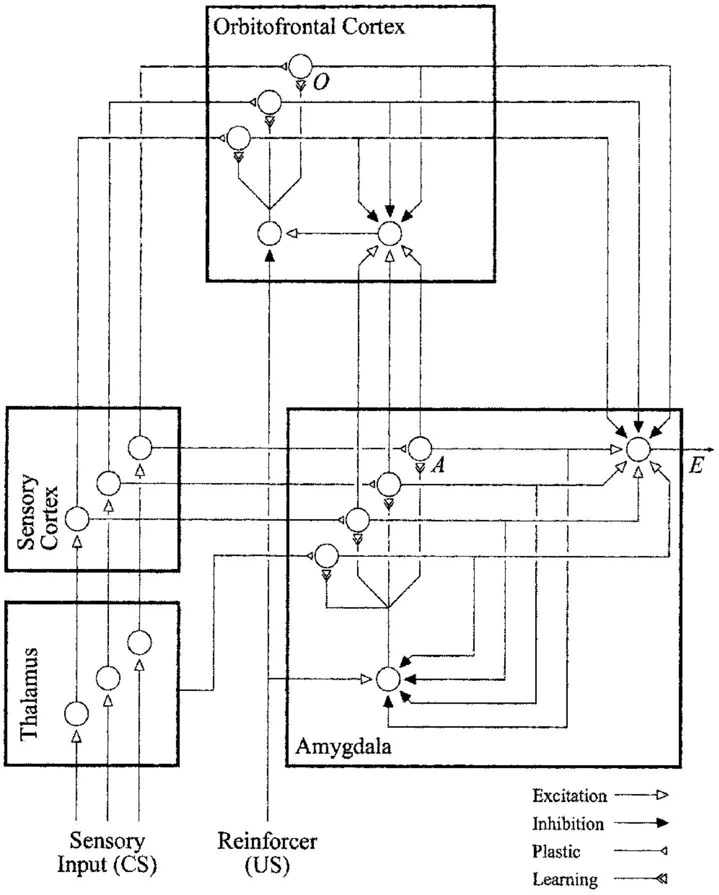

Figure 4.2: Balkenius and Morén model from [2]. The circles are computational nodes: each classified into an anatomical region of the brain. The legend shows the types of inter-nodal connections used in the model. See cited paper for details.

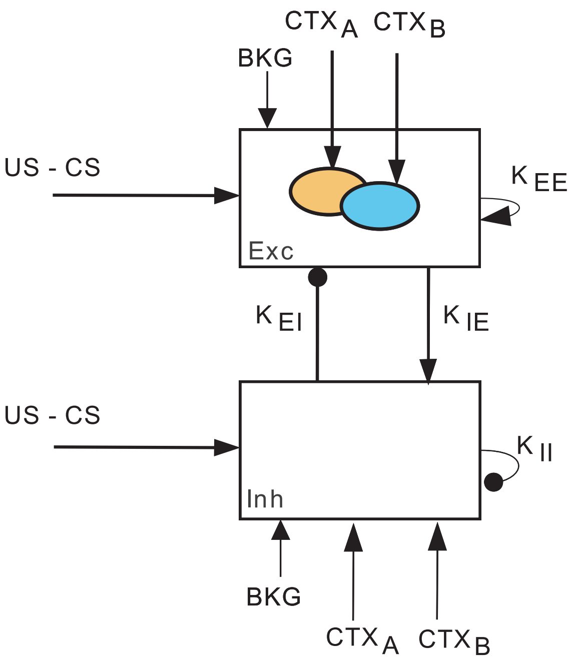

Figure 4.3: Vlachos et al. model from [72]. Excitatory (top box) and inhibitory (bottom box) neural populations receive CS/US/context information as well as background inputs (BKG). In addition, each population has recurrent connections (K_{EE} and K_{II}). See cited paper for details.

conditioning, as well as context-dependent extinction and renewal, and shows how two distinct areas of the basal nucleus of the amygdala are recruited during these processes.

It is a spiking neural network model that uses 4000 leaky-integrate-and-fire neurons: 3400 excitatory (top box in figure 4.3) and 600 inhibitory (bottom box in figure 4.3). The activity of the populations of these neurons are explained with differential equations, as are the changes in connection strengths between the populations. The learning rule in the model relies on temporal overlap between stimuli.

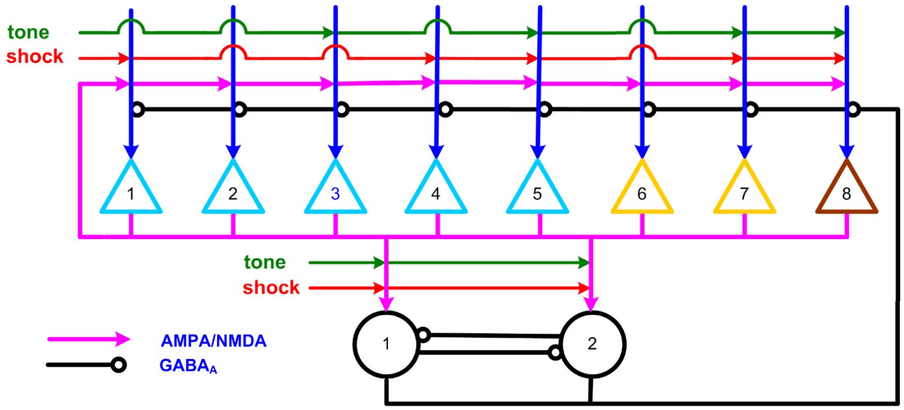

Figure 4.4: Li et al. model from [41]. Tone and shock signals project to excitatory (triangles 1-8) and inhibitory (circles 1 and 2) neurons each with recurrent connections (connections terminating with open circles are inhibitory). The neurotransmitters modelled are specified in the legend. See cited paper for details.

4.4 Li et al.

Li et al. developed a model focused on acquisition and extinction of fear in lateral amygdala neurons [41]. It is a cellular model that takes into consideration levels of neuromodulators available at synapses, and includes details such as calcium, sodium, and potassium concentrations, as well as GABA, NMDA, and AMPA receptors. The amygdala model includes different types of pyramidal cells (triangles in figure 4.4) as well as interneurons (circles in figure 4.4).

Learning in the model depends on calcium concentrations at the cell and resembles basic Hebbian learning. The model explores the cellular mechanism of plasticity in the amygdala including the role of NMDA currents in extinction learning. The stated goal of the model is to "bridge biophysical and network modelling approaches" to gain a greater understanding of fear conditioning [41]. However, the model only accounts for the first-order conditioning, extinction, and blocking experiments discussed earlier.

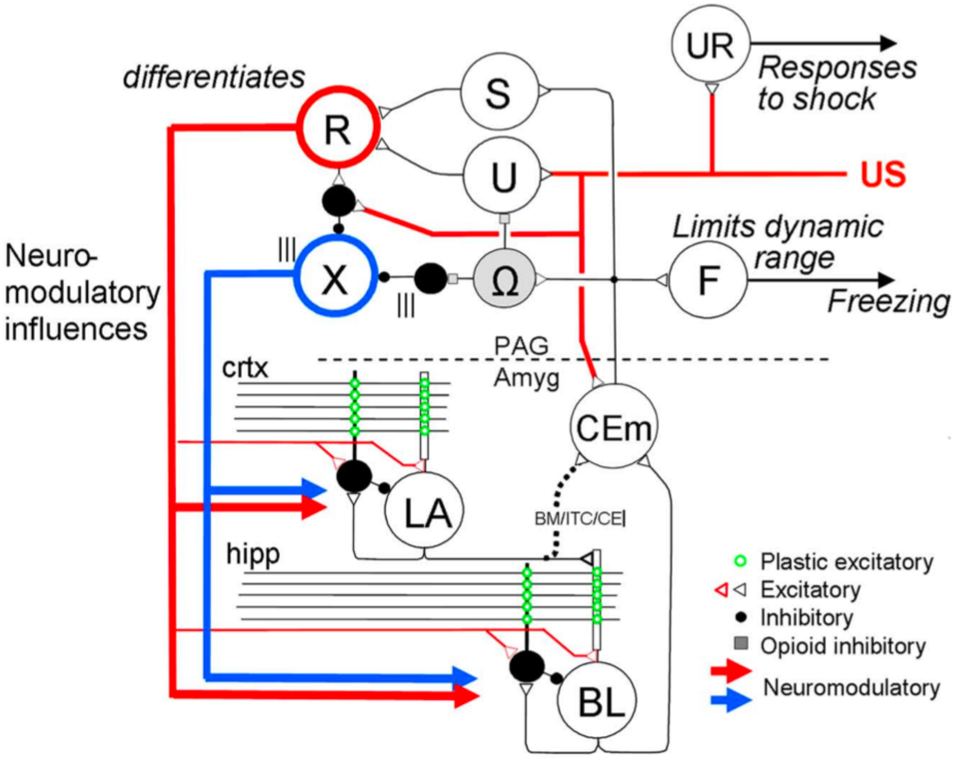

Figure 4.5: Krasne et al. model from [35]. The circles represent neuron-like computational nodes with different types of connections as shows in the legend. Learning occurs on the projections from the cortex and hippocampus to the LA and BL respectively. See cited paper for details.

4.5 Krasne et al.

Krasne et al. developed a functional model of fear conditioning that reproduces results from first-order conditioning, second-order conditioning, blocking, and extinction and renewal experiments [35]. The model includes areas of the amygdala and the periaqueductal gray and receives inputs from cortical/thalamic neurons as well as hippocampal neurons. Subpopulations of the amygdala and periaqueductal gray are connected with excitatory, inhibitory, neuromodulatory and plastic connections in order to simulate a wide range of fear conditioning related behaviours.

The model uses generalized, non-spiking neurons to represent its sub-populations (one neuron per population: circles in figure 4.5). Plasticity of connections in the model is described by a Hebbian-like rule that includes an eligibility parameter governed by reward signals as well as local cellular activity.

4.6 Comparisons of the models

There are considerable similarities and differences between the models discussed above and the model presented in this thesis.

The model presented by Li et al. includes greater biological detail than the one presented in this thesis, but is focused on a specific region of the amygdala only. Grossberg and Levine's model formulates equations governing CS/US association, but does not closely map the equations to biological function. Vlachos et al.'s model does use spiking neurons, but considers only a specific part of the fear conditioning circuit. Balkenius and Morén's model, as well as Krasne et al.'s model, takes a circuit level modelling approach but uses idealized neurons as opposed to spiking neural populations.

One of the main benefits of developing neural models with the NEF is that it is easy to construct circuit-level models that are based on biologically realistic spiking neurons and neural connections. The model presented in this thesis combines many aspects of the above models; it accounts for the widest variety of behavioural experimental results, and contains detailed biological mechanisms. In fact, the model is based on the model described by Krasne et al. with added biological detail, more flexible and general representations, and functional extensions. The important differences between this model and the one described by Krasne et al. will be covered in chapter 8.

Chapter 5

Neurons and plasticity

5.1 Neurons

The neuron is the fundamental information processing component of the animal brain. The remainder of this thesis assumes a basic understanding of the biological properties of a neuron, which will be discussed briefly here.

A neuron is a cell that contains mitochondria, ribosomes, DNA, and the other cellular components that are found across most animal cells. One of the distinguishing features of a neuron (compared to other cells) is the branches that extend from its cell body [33]. The majority of these branches are dendrites, which connect up to, and receive information from, other cells (a pyramidal neuron may have thousands of these connections [3]). In this way, dendrites can be thought of as the inputs to the cell. Each neuron typically has one axon, the output, which is a branch that sends information from the cell to other neurons via the axon terminal. See figure 5.1 for an illustration of these parts of a neuron.

The process by which input and output signals are collected and sent involves chemical as well as electrical mechanisms [33]. Along the length of a dendrite or axon, transmission can generally be thought of as an electrical process. Electrical activity at one end of a dendrite or axon propagates through its length to the other end of the branch. Where axons of one cell meet dendrites of another (a location called a synapse), communication is achieved via a chemical process. Electrical activity arriving at the end of an axon causes the release of chemicals called neurotransmitters. The neurotransmitters released from the axon are detected by nearby dendrites, and cause an increase in electrical activity at those dendrites. See figure 5.2 for an illustration of a synapse.

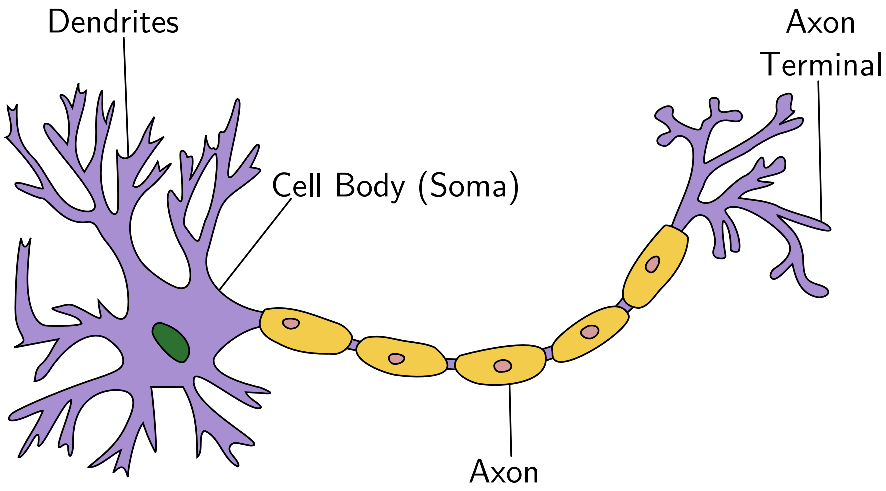

Figure 5.1: Illustration of a neuron, from [5]. The cell receives inputs from the dendrites, processes them at the cell body, and sends the output down the axon. The axon terminals transmit signals to dendrites of other neurons.

While dendrites and axons are primarily involved in the transmission of information (in the form of electrical activity) from neuron to neuron, much of the actual computation is typically taken to be performed at the cell body. Incoming dendritic currents meet at the cell body; the electrical activity on the axon (output) of a neuron can be thought of as being the sum of the electrical activity on the dendrites (input) of that neuron [45]. This computation can be generalized as follows:

$$a = \sum_{i=1}^{n} d_i \tag{5.1}$$

where a is the electrical activity on the axon, d_i is the electrical activity on dendrite i, and n is the number of dendrites on a neuron.

Critical to this summation, is that dendrites do not precisely transmit the information passed to them by an axon. Some dendrites of a cell are more or less affected by the electrical activity of an axon of another cell than are other dendrites [45]. How well each axon transmits information to a dendrite is characterized by the connection weight of a synapse. A high connection weight means that a dendrite receives and transmits to its cell body a large electrical signal when excited by a specific nearby axon. A low connection

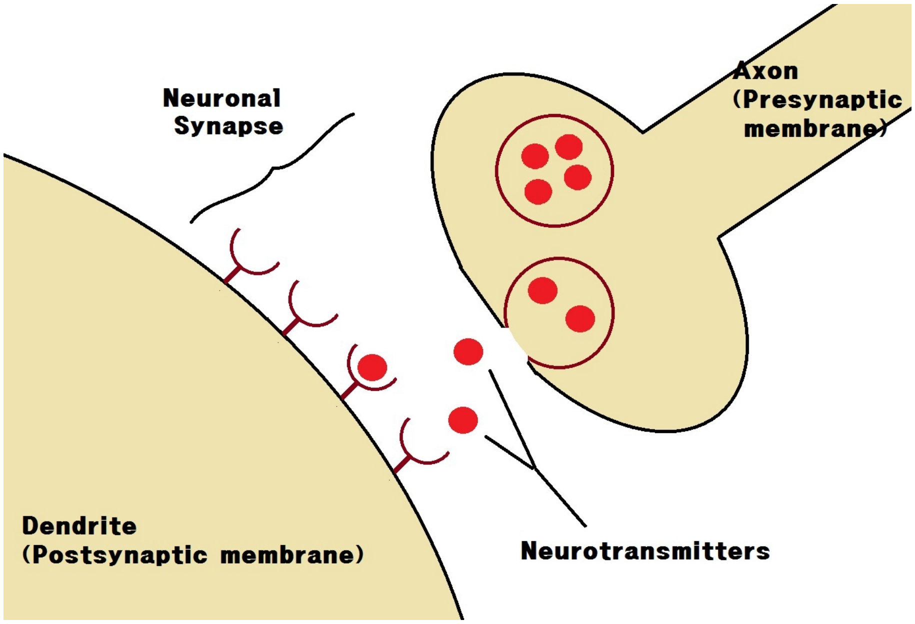

Figure 5.2: Illustration of a synapse. Electrical activity on the axon causes the release of neurotransmitters. These neurotransmitters are detected by the dendrites of the postsynaptic neuron.

weight means that a dendrite receives and transmits to its cell body a small electrical signal when excited by a specific nearby axon.

To account for this, we can then rewrite equation 5.1 as follows:

$$a = \sum_{i=1}^{n} p_i w_i \tag{5.2}$$

where w_i represents the weight for a particular dendrite, i, and p_i represents the presynaptic axon activity at that dendrite.

Equation 5.2 describes a simple, commonly used abstraction of the information processing capabilities of a neuron [45].

5.2 Plasticity

The connection weights discussed in the previous section do not always remain the same value. Connection weights are plastic (able to change); the term 'synaptic plasticity' describes the process by which a weight at a synapse increases or decreases. Synaptic plasticity is thought to be the primary mechanism by which animals learn and store memories [32].

Two models (which are employed in the work described in this thesis) have been proposed to explain various synaptic plasticity experiments.

5.2.1 Hebbian plasticity

One model that has been proposed to govern synaptic plasticity has been attributed to the Canadian neuroscientist Donald Hebb [28]. The model is described nicely in Hebb's own words:

> Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability. When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased. [28]

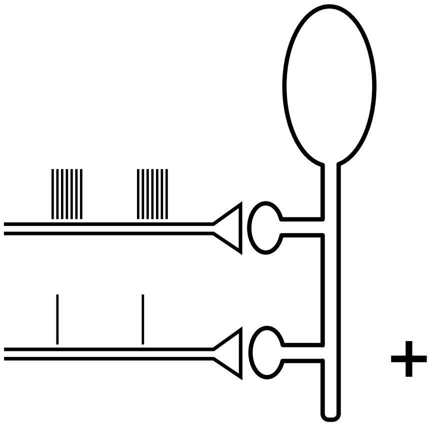

Figure 5.3: Illustration of one of the consequences of Hebbian learning, from [5]. When a weak stimulation (bottom synapse) is present at the same time as a strong stimulation (top synapse) causes the cell to fire, the weight at the synapse of the weak stimulation will be strengthened (indicated by the plus sign)

Figure 5.3 illustrates an example of Hebbian plasticity.

The popular axiom summarizing this paragraph is: "cells that fire together wire together". The following equation describes this phenomenon:

$$\Delta\omega_{ij} = a_i a_j \tag{5.3}$$

That is, the change in weights at the synapse between two neurons depends on the co-activity of the presynaptic neuron (a_i) and the postsynaptic neuron (a_j).

Hebbian learning on its own is unstable and does not allow for decreases in synaptic weights. As seen in equation 5.3, if activity is a positive quantity, the synaptic weight could grow without bound. To account for this unrealistic behaviour, modifications to Hebbian learning have been proposed. One well known modification is the Bienenstock, Cooper, and Munro (BCM) rule [8]. The BCM equation (5.4) is similar to equation 5.3, but with an added term that allows for both increases and decreases in synaptic weights. If the activity of the postsynaptic cell is above its average activity, θ, the weight between the active presynaptic and postsynaptic cells will be increased. However, if the activity of the postsynaptic cell is below its average activity, the weight between the active presynaptic and postsynaptic cells will be decreased.

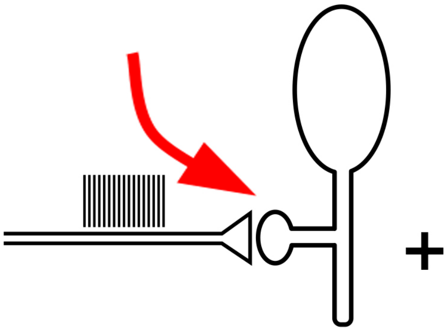

Figure 5.4: Illustration of neuromodulatory learning, modified from [5]. The presence of a neuromodulator (red arrow) at the synapse between two neurons affects the change of the weight at the synapse. The release of the neuromodulator could be a result of the activity of this neuron, or other neurons.

$$\Delta\omega_{ij} = a_i a_j (a_j - \theta) \tag{5.4}$$

In summary, if the presynaptic neuron, i, is involved in making the postsynaptic neuron, j, fire more than it usually does, weights between i and j will increase. If i is involved in making the postsynaptic neuron j fire less than it usually does, weights between i and j will decrease.

5.2.2 Neuromodulatory learning

In Hebbian learning, synaptic plasticity is governed by the activity of presynaptic and postsynaptic cells. However, synaptic plasticity can also be governed by neurotransmitters released from cells that do not directly excite the postsynaptic cell (see figure 5.4).

Certain neurotransmitters, such as dopamine, are released by brain subsystems in response to affectively significant stimuli [57] [60]. It is hypothesized that these neurotransmitters help to facilitate synaptic plasticity in the brain regions in which they are released [56]. One popular suggested mechanism of this facilitation is that the neurotransmitter carries an error signal [67]. This error signal provides information to the synapse which determines whether the weight of the synapse should be strengthened or weakened. The effects of this type of error signal will be discussed further in section 6.4, but can be summarized with the following equation.

$$\Delta\omega_{ij} \propto E \tag{5.5}$$

The error signal E can be either a positive or negative value, and may be modulated by pre and post-synaptic activity as discussed later in section 6.4.

Chapter 6

Large-scale neural modelling with the NEF

The Neural Engineering Framework developed by Eliasmith and Anderson [21] provides a method for representing and transforming information with neurons. Using the NEF, complex algorithms can be encoded in neurons to generate models such as the one presented in this thesis.

The NEF has three principles. The first is the representation principle, which details how a population of spiking neurons can represent high dimensional information (a vector). The second is the transformation principle, which details how the connections between populations of neurons can be used to perform computations on the vectors being represented; the synaptic connection weights between populations can be solved for analytically in order to efficiently compute an approximation to a wide variety of functions. In addition, these synaptic connection weights can be learned during a simulation to approximate a desired function [42]. And the third principle is the principle of dynamics, which we will not be considering in detail here.

6.1 Single neuron model

The NEF can support a wide variety of neuron models. Neuron models describe the dynamics of individual neurons; i.e., how they turn input activity from dendrites into output activity on axons. The most common neuron model used in NEF models (including the one presented in this thesis) is the leaky integrate and fire (LIF) model.

Parameter	Value
t_{ref}	.002s
RC	.02s
V_{thresh}	1

Table 6.1: Default LIF parameters used in this thesis.

For this neuron model, the activity on an axon is a train of spikes. At a synapse, the train of spikes is translated into a current injection into nearby dendrites. The generation of a cell's output spikes proceeds as follows. The voltage at the soma of an LIF neuron ($V(t)$) is affected by the sum of the currents from its dendrites ($J(t)$). Once the voltage level of the soma reaches a threshold voltage, V_{th}, a spike is generated, and the soma voltage is reset to zero. If the voltage does not reach V_{th}, it will tend towards zero over time; this is what the 'leak' term in the name refers to. After a spike has been generated, the membrane voltage will be reset to zero and remain zero for a certain length of time, t_{ref}, before it can begin to increase again.

Figure 6.1 shows how a cell membrane can be described as an electrical circuit. Equation 6.1 describes how this circuit behaves while the soma voltage is between 0 and V_{th}.

$$\frac{dV}{dt} = -\frac{1}{RC}(V(t) - J(t)R) \tag{6.1}$$

The parameters required for this LIF model, along with their corresponding default values used here, are shown in table 6.1.

There are advantages to choosing LIF neurons over other model neurons: they capture the primary computational mechanisms of biological neurons, and at the same time are computationally efficient to simulate.

6.2 Representation

At the core of the NEF is a method for representing information in neural populations. That is, a method that allows the activity of a population of neurons to represent an external stimulus, a numeric value, a concept, etc. Any of these types of information can be represented by a vector in a one-dimensional, two-dimensional, or n-dimensional space.

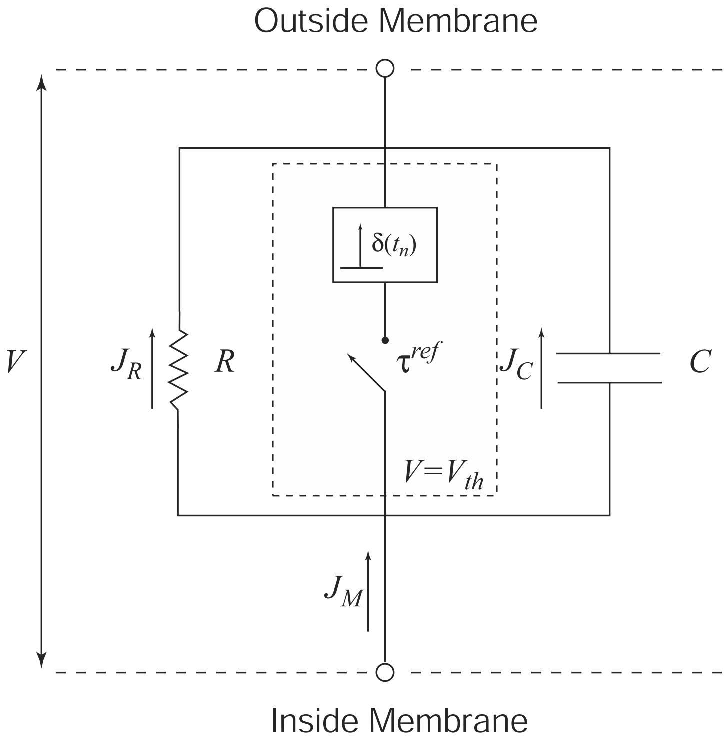

Figure 6.1: Circuit that models a leaky integrate and fire neuron, from [21]. The membrane voltage V depends on the leak resistance, R, the capacitance of the bilipid layer of the cell, C, and the components inside the dashed box. Once V reaches the threshold voltage, V_{th}, at time t_n, the switch in the dashed box is closed. This generates a spike, $\delta(t_n)$, and resets V to zero until the switch is opened again after the refractory period, τ^{ref}.

For a working example, consider that it may be useful for an animal to have access to a neural representation of the two-dimensional position of an eye; having access to such a neural representation may allow the animal to keep the eye steady while its head moves. In this case, the vector being represented by the population of neurons could be $[x\text{-}coordinate, y\text{-}coordinate]$. A single population of neurons could represent all the possible positions of the eye.

A population of neurons, as opposed to a single neuron, is required to represent information in a high-dimensional space, because a single neuron cannot accurately represent information in a high-dimensional space on its own. The output of a neuron in the NEF could theoretically represent one entire dimension (although it can represent parts of many dimensions); that is, the representational information from a neuron is taken only from its spike-rate (the frequency at which it generates spikes along its axon), which is a scalar (one-dimensional) value. It would be impossible to represent an entire two-dimensional space using only a single neuron.

The NEF provides a method for combining the scalar outputs of single neurons to represent higher-dimensional information. This is done by allowing each neuron in a population to represent a part of a higher-dimensional space. In the two-dimensional eye positioning example, one neuron could represent either the y-position of the eye, or the x-position of the eye. One population of two neurons - one representing the entire y-axis, and the other representing the entire x-axis - could represent any possible position of the eye. In this case, one neuron may not fire at all when the eye is at -1 (looking all the way down) on the y-axis, and might fire fastest when it is at 1 (looking all the way up) on the y-axis; the preferred direction of this neuron would be said to be along the y-axis, and is represented with $[0, 1]$. The other may not fire at all when the eye is at -1 (far left) on the x-axis, and might fire fastest when it is at 1 (far right) on the x-axis; the preferred direction of this neuron would be said to be along the x-axis, and is represented by the vector $[1, 0]$.

The two-neuron population is an idealized example. Animal brains are thought to utilize population coding in representing information [24], so in reality, the two-dimensional eye position may actually be represented in a population of neurons containing hundreds or thousands of neurons instead of only two. In such a population, some neurons will be more sensitive to the input (have a greater gain), and some may respond most sensitively as the input moves along a line other than the x or y axis; for example, a diagonal (such a neuron would have a preferred direction vector of $[0.707, 0.707]$ if we restrict our preferred direction vectors to the unit circle).

Every neuron in the population will respond differently to the same input depending on which part of the space (which vector) the neuron prefers. As discussed in the previous section, the spike rate of a neuron depends on the sum of the currents being received from the dendrites of the cell. In the example of the two-dimensional eye position, let us assume that each neuron in the population representing eye position is connected to the muscles of the eyeball. The amount of current, J, that each neuron receives is dependent on the position of the eye (as measured from the muscles), $\mathbf{x}$, the neuron's gain, α, the neuron's preferred direction vector, $\mathbf{e}$, and a value that accounts for background current that the neuron receives regardless of its input, J_{bias} (see equation 6.2). It is thought that a population of neurons is made up of many heterogeneous neurons with different parameters [15]. Because of this, when creating neural populations with the NEF, the parameters α, $\mathbf{e}$, and J_{bias} are often randomly chosen (within a suitable range) for every neuron in a population.

$$J = \alpha \mathbf{e} \cdot \mathbf{x} + J_{bias} \tag{6.2}$$

The dot product between eye position and the neuron's preferred direction vector determines how sensitive the neuron is to input in a particular direction. For example, let us consider two neurons with preferred direction vectors along the x-axis and y-axis. The preferred direction vectors of these neurons are $[1, 0]$ and $[0, 1]$ respectively. If the eye position is at $[0.5, 1]$, the result of the dot product for the x-axis neuron is 0.5, and the result of the dot product for the y-axis neuron is 1. Assuming other parameters being equal, the y-axis neuron would have more current injected into it; it is more sensitive to that eye position. Consider another neuron in the population that has a preferred direction vector on the diagonal: $[0.707, 0.707]$. If the eye position is at $[0.5, 0.5]$, its dot product would be 0.707, whereas both the x and y-axis neurons would have dot products of 0.5. The neuron with a preferred direction vector along the diagonal is more sensitive to that eye position, and would have more current injected into it than the other two neurons.

This discussion applies for higher-dimensional representations as well. The preferred direction vector of a neuron will always be in the same dimensionality as the information it is encoding. In any dimensionality, the result of a dot product is a scalar value that represents the similarity of the two vectors.

Having calculated the current being injected into the neuron, we turn our attention to the function of the neuron that converts the sum of its input currents to output spikes, which is described by the LIF model mentioned earlier. The non-linear function of the LIF neuron will be referred to as $G()$ from here on. Using this convention, activity of a cell can be written as

$$a(J) = G(J). \tag{6.3}$$

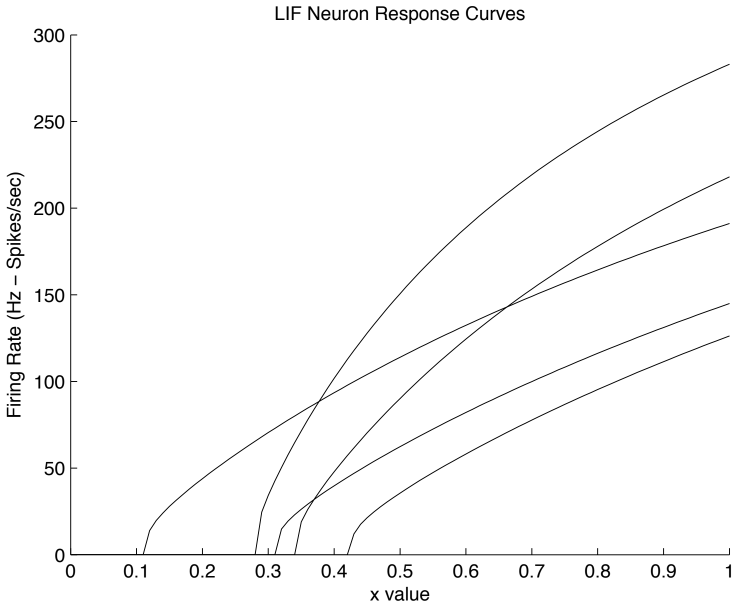

Figure 6.2: Tuning curves for a population representing a one-dimensional input, **x**, from [16]. Each line on the plot represents the response of a neuron in a population. These neurons increase their firing rate as the input increases.

Allowing for randomly generated parameters within a biologically realistic range for equation 6.2, the current injected into a neuron can be considered primarily a function of the information it is representing, **x**. As the activity (spike-rate) of a neuron is a function of J, and allowing for randomly generated parameters within a biologically realistic range for the LIF equation 6.1, the activity of a neuron can also be considered primarily a function of **x**. For one-dimensional inputs, a plot of neuron activity vs **x** yields what is called a tuning curve for a neuron (see figure 6.2).

For the case of higher-dimensional inputs, a similar tuning curve can be found for each neuron by plotting neuron activity vs the dot product between **x** and the neuron's preferred direction vector **e**. This means simply replacing the x-values on the x-axis of figure 6.2 with dot products.

It should be clear by looking at figure 6.2 that a single neuron will likely not represent a one-dimensional space well on its own. Further, it should be clear that a single neuron will likely not represent one direction in a higher-dimensional space well on its own. Because of the assumptions made in the NEF, the number of neurons in a population, N, needed to achieve a specified accuracy in a representation can be determined analytically; the mean square error of the representation decreases at $1/N$. See [21] for details. Further discussion in this thesis assumes that information is being represented in a population with enough neurons to encode the input information with a root-mean-square precision of around 1%.

We have seen how to generate neuron activities in a population of neurons based on some input, but we can go the other way as well. We can estimate the input to a population of neurons by measuring the firing rate of every neuron. The task here is explained by the following equation.

$$\hat{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{d}_i a_i \tag{6.4}$$

Multiplying a vector of the same dimensionality as $\mathbf{x}$, $\mathbf{d}$, by the activity of a neuron, a, and summing all of those multiplications for every one of the n neurons in the population, gives back an estimate of the information encoded in that neural population ($\hat{\mathbf{x}}$). This variable $\mathbf{d}$ is called the neuron's decoder value, and is constant - it stays the same regardless of what the input to the neuron is. Each neuron in a population has an associated decoder.

We want to solve equation 6.4 for $\mathbf{d}$ such that the difference between $\mathbf{x}$ (the actual value of the input) and $\hat{\mathbf{x}}$ is minimized. The following matrix algebra solves for $\mathbf{d}$ in this way.

$$\mathbf{X} = [-x, -x + dx, ..., x - dx, x]$$

$$\mathbf{A} = \begin{bmatrix} a_1(\mathbf{X}) \\ \vdots \\ a_n(\mathbf{X}) \end{bmatrix}$$

$$\mathbf{d} = \mathbf{\Gamma}^{-1}\mathbf{\Upsilon}, \text{where } \mathbf{\Gamma} = \mathbf{A}\mathbf{A}^{\mathbf{T}} \text{ and } \mathbf{\Upsilon} = \mathbf{A}\mathbf{X}^{\mathbf{T}} \tag{6.5}$$

In equation 6.5, $\mathbf{d}$ is an m by n matrix containing every neuron's decoder, where m is the number of dimensions in the input. The activity matrix A is an $\frac{xrange}{dx}$ by n matrix that contains the activity of every neuron for every value of input to be considered.

The $\mathbf{d}$ solved for in equation 6.5 allows us to go from spike rates of a population of neurons back to the original signal. However, the activity of neurons cannot always be characterized by a spike rate. If the activity of a population is changing over time, then we need a way to determine the original signal only from the sequence of spikes and not from a spike rate.

The outputs of neurons are not simply discrete spikes; they are spikes filtered by a postsynaptic current (PSC). A biologically determined function that represents this filter is $PSC(t)$ (equation 6.6). The function depends on the synaptic time constant, τ_{PSC}.

$$PSC(t) = \frac{1}{\tau_{PSC}} e^{-t/\tau_{PSC}} \tag{6.6}$$

Filtering a spike train with this filter results in an output that looks like the plot in figure 6.3.

Given this output of our neurons, which we will call $a(t)$, to determine the estimate of the original signal, $\hat{\mathbf{x}}$, we must find the appropriate decoders, $\mathbf{d}$. Having solved for the decoder of every neuron in the population in the above equation so as to minimize the difference between $\mathbf{x}$ (the actual value of the input) and $\hat{\mathbf{x}}$, the original signal can be reconstructed by multiplying the activity, $a(t)$, of each neuron by its corresponding $\mathbf{d}$, and summing the result for all neurons in the population:

$$\begin{aligned} \hat{\mathbf{x}}(t) &= \sum_{i=1}^{n} \mathbf{d}_i a_i(t) \\ &= \sum_{i=1}^{n} \mathbf{d}_i \sum_{s_i=1}^{nspikes_i} PSC(t - t_{s_i}), \end{aligned} \tag{6.7}$$

where s_i indexes each of the $nspikes_i$ spikes produced by a neuron.

In the next section we consider how we can use the encoding and decoding processes discussed here to compute functions between populations of neurons.

6.3 Transformation

In the example from the previous section, the state represented by the neurons was eye position (more specifically, the activity needed to drive the muscles that control an eyeball). However, most of the inputs to neurons in the brain are the outputs of other neurons. One population of neurons can be directly connected to another population of neurons as is shown in figure 6.4.

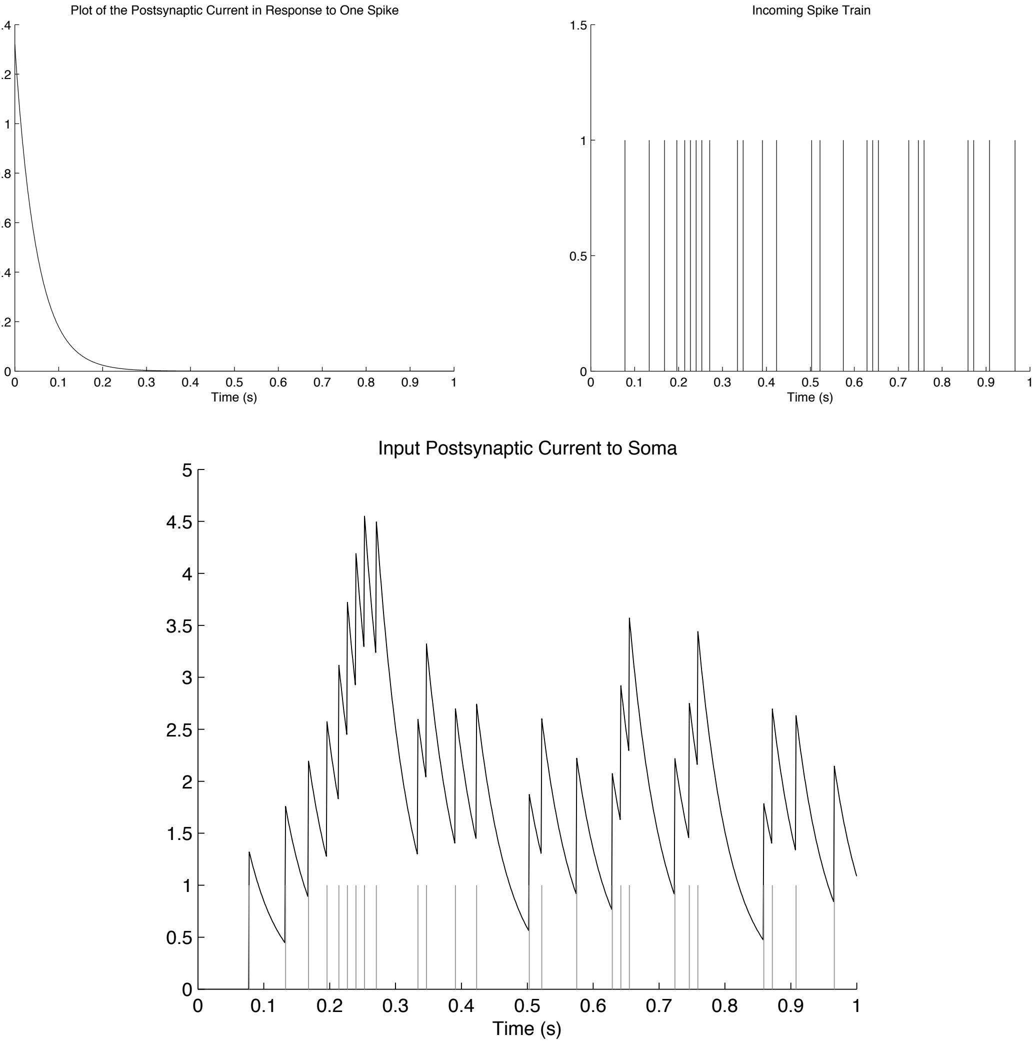

Figure 6.3: Plot of postsynaptic current (top left). Plot of spikes (top right). Plot of spikes filtered by postsynaptic current (bottom). From [16]. These plots show how a time varying signal can be approximated by spiking neurons.

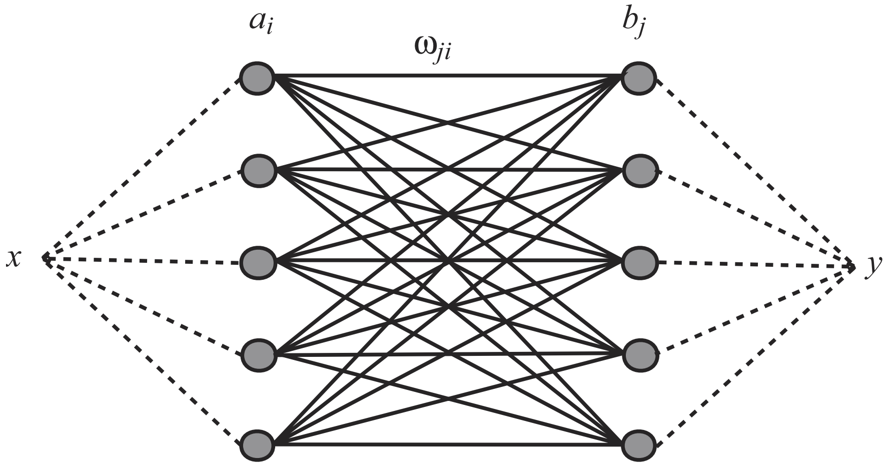

Figure 6.4: A communication channel between two populations of neurons, from [21]. A vector, x, is represented in a population, a, with neurons indexed by i. A second population, b, represents a vector, y, and has neurons indexed by j. For a communication channel, the weights, ω_{ij}, are chosen so that $x = y$.

In the simplest case, the neurons in population a are connected to the neurons in population b such that population b represents the same information as population a; such a connection between two populations is called a ‘communication channel’.

When we wanted to determine the amount of current required for a neuron to represent an external input, we used equation 6.2, shown again here.

$$J = \alpha \mathbf{e} \cdot \mathbf{x} + J_{bias}$$

In the communication channel described here, we want $\mathbf{x}$ in equation 6.2 to be replaced with $\hat{\mathbf{x}}$: the value that the a population is representing. From the previous section (equation 6.4) we know that

$$\hat{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{d}_i a_i.$$

Substituting the right hand side of equation 6.4 into equation 6.2 for $\mathbf{x}$ gives the amount of current to be injected into a neuron in population b (indexed by j) in terms of the activity of the neurons in a neural population a (indexed by i) in order to implement a communication channel.

$$\begin{aligned} J_j &= \alpha_j \mathbf{e}_j \cdot \hat{\mathbf{x}} + J_{bias_j} \\ J_j &= \alpha_j \mathbf{e}_j \cdot \sum_{i=1}^{n} \mathbf{d}_i a_i + J_{bias_j} \end{aligned} \tag{6.8}$$

Having found the current injected into each neuron of the b population, we can apply the neuron model function $G(J_j)$ to determine the activity of a neuron in the b population (b_j) when it is representing $\hat{\mathbf{x}}$.

$$b_j = G(\alpha_j \mathbf{e}_j \cdot \sum_{i=1}^{n} \mathbf{d}_i a_i + J_{bias_j}) \tag{6.9}$$

As discussed earlier, a weight is the term used to describe how well activity from the axon of a presynaptic neuron induces current in a dendrite of a postsynaptic neuron. We can collect α_j, and the dot product between $\mathbf{e}_j$ and $\mathbf{d}_i$ from equation 6.9 into one weight term.

$$\omega_{ij} = \alpha_j \mathbf{e}_j \cdot \mathbf{d}_i \tag{6.10}$$

Substituting into 6.9 gives

$$b_j = G(\sum_{i=1}^{n} \omega_{ij} a_i + J_{bias_j}). \tag{6.11}$$

The weights determined by equation 6.10 are the weights needed to create a communication channel. But if we want population b to represent something other than what population a is representing, we need only replace $\hat{\mathbf{x}}$ in equation 6.8 by $\mathbf{C}\hat{\mathbf{x}}$ where $\mathbf{C}$ is any matrix that can be multiplied by $\hat{\mathbf{x}}$. This allows population b to represent any linear transformation of the value represented by population a. Weights can now be represented as

$$\omega_{ij} = \alpha_j \mathbf{e}_j \mathbf{C} \mathbf{d}_i. \tag{6.12}$$

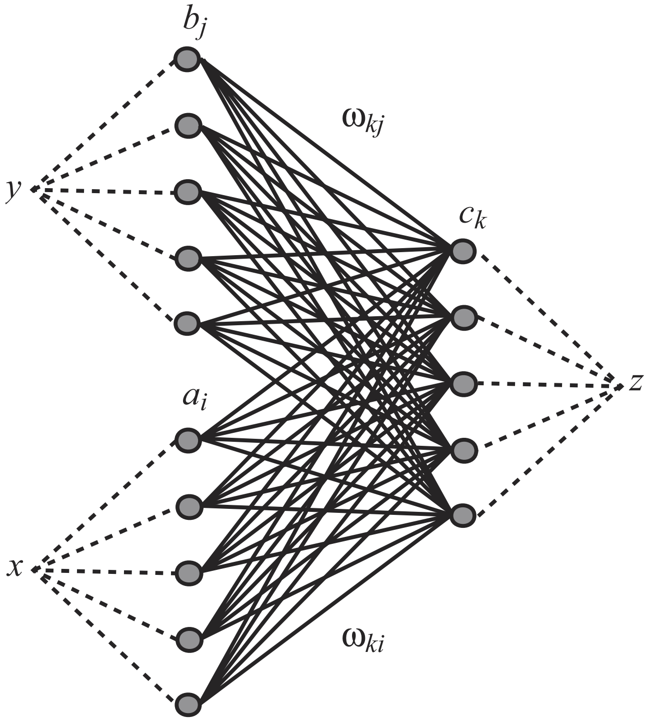

Figure 6.5: A neural population, c, with inputs from neural populations a and b, from [21]. The value represented in population c (z) is a linear combination of the values represented in b (y) and a (x).

Population b can also receive more than one input. Let's introduce another population, c (whose neurons are indexed by k), into this example, and connect the populations as shown in figure 6.5. We can easily modify equation 6.8 for this situation.

$$\begin{aligned} J_k &= \alpha_k \mathbf{e}_k \cdot (\hat{\mathbf{x}} + \hat{\mathbf{y}}) + J_{bias_k} \\ \text{or with transformations, } J_k &= \alpha_k \mathbf{e}_k \cdot (\mathbf{C}_1 \hat{\mathbf{x}} + \mathbf{C}_2 \hat{\mathbf{y}}) + J_{bias_k} \end{aligned} \tag{6.13}$$

Now the activity at population c will be

$$\begin{aligned} c_k &= G(\alpha_k \mathbf{e}_k \cdot (\mathbf{C}_1 \sum_i \mathbf{d}_i a_i + \mathbf{C}_2 \sum_j \mathbf{d}_j b_j) + J_{bias_k}) \\ c_k &= G(\sum_i \omega_{ik} a_i + \sum_j \omega_{jk} b_j + J_{bias_k}) \ . \end{aligned} \tag{6.14}$$

This method can be applied to any number of input populations. A population can therefore represent any combination of linear transformations applied to its inputs. The same derivation holds for nonlinear functions of the input as well, although the decoders optimized for will change (see [21]).

6.4 Learning

In section 5.2, the commonly held belief that learning occurs by adjusting weights between neurons was introduced. Equations 6.11 and 6.12 demonstrate how changing weights between neurons leads to a different transformation being calculated between two populations.

Indeed, in the NEF, the assumption is that learning in brains occurs by adjusting the weights, and therefore the transformations computed, between populations of neurons. The following sections discuss two different ways in which the weights of NEF neurons can be adjusted online, rather than by solving the least-squares minimization in equation 6.5.

6.4.1 Error minimization

The error minimization technique of learning involves comparing the value represented by a population of neurons to a value that we want the population to represent, and using that difference to adjust the input weights to that population in order to achieve the desired representation (see figure 6.6).

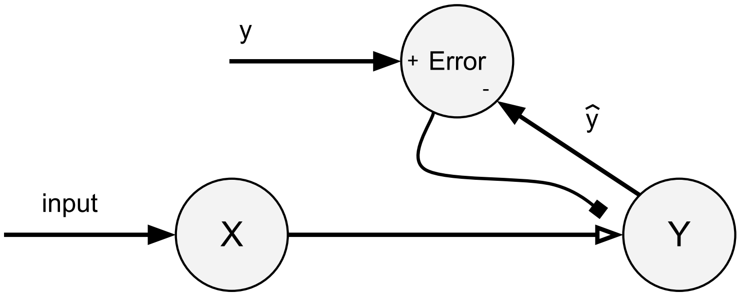

Figure 6.6: A general error-modulated learning circuit. The difference between the desired value represented by population Y, $\mathbf{y}$, and the actual value represented by population Y, $\hat{\mathbf{y}}$, is calculated by the *Error* population, and used to modify the weights between X and Y such that Y comes to represent the desired value $\mathbf{y}$ for a given input.

MacNeil and Eliasmith demonstrated a least squares error method locally implemented in a spiking network that accomplishes this [42]. The least squares technique aims to reduce the squared difference between the actual representation of the output population and the desired representation. The error, $\mathbf{E}$, is represented as

$$\mathbf{E} = \mathbf{y} - \hat{\mathbf{y}}. \tag{6.15}$$

We will represent the squared error as follows.

$$SE = \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^2 \tag{6.16}$$

Next we use equation 6.4 to replace $\hat{\mathbf{y}}$ in the above equation and take the derivative with respect to $\mathbf{d}_i$.

$$\begin{aligned} SE &= \frac{1}{2}(\mathbf{y} - \sum_{i=1}^{n} \mathbf{d}_i a_i)^2 \\ \frac{dSE}{d\mathbf{d}_i} &= (\mathbf{y} - \sum_j \mathbf{d}_j a_j) a_i \end{aligned}$$

In the above equation, i indexes only the neuron whose connection is being optimized, while j indexes all neurons in the population. Notice that the term in brackets is equivalent to the error we defined in equation 6.15.

$$\frac{dSE}{d\mathbf{d}_i} = (\mathbf{E})a_i \tag{6.17}$$

If we rewrite this using delta rule form and add a learning rate parameter κ, we get

$$\Delta\mathbf{d}_i = \kappa\mathbf{E}a_i. \tag{6.18}$$

To get this expression in terms of weights, we multiply both sides by the encoding vector $\mathbf{e}_j$, and the neuron gain α_j.

$$\begin{aligned} \Delta\mathbf{d}_i \cdot \mathbf{e}_j\alpha_j &= \kappa\alpha_j\mathbf{e}_j \cdot \mathbf{E}a_i \\ \Delta\omega_{ij} &= \kappa\alpha_j\mathbf{e}_j \cdot \mathbf{E}a_i \end{aligned} \tag{6.19}$$

Equation 6.19 describes how to adjust weights of a specific neural connection in order to minimize the representation error of a population.

It is important that such a proposed learning rule could potentially be implemented in a biologically plausible way. One way to check for biological plausibility is to ensure 'locality' of the learning rule; i.e., that the parameters on the right hand side of equation 6.19 are available to individual synapses. α and $\mathbf{e}$ are neural parameters, and a is a measurement of presynaptic cell activity which is directly available to the synapse. The learning rate parameter, κ, is a constant that accounts for how quickly synaptic weights can change; this process relies on biochemical factors such as changes to neurotransmitter release and uptake. It is possible for the error, $\mathbf{E}$, to be available to every synapse as well, under specific assumptions. One assumption is that errors, and therefore desired values, are represented in neural populations in the brain. The other assumption is that error populations are connected to the synapses between the input and output neural populations as shown in figure 6.6.

6.4.2 Hebbian learning

In addition to the error modulated learning rule, a Hebbian learning rule has also been implemented in the NEF [5]. As discussed earlier, Hebbian learning is the modification of synaptic weights governed by the timing of input and output activity of a neuron. A simplified characterization of this relationship was given in equation 5.3 and is repeated here.

$$\Delta\omega_{ij} = a_i a_j$$

This equation can be translated directly into terms used by the NEF. We can include the scaling factors for the learning rate (κ) and the neural gain (α) from equation 6.19 to represent the Hebbian learning rule as

$$\Delta\omega_{ij} = \kappa\alpha_j a_i a_j. \tag{6.20}$$

As mentioned earlier, purely Hebbian learning would cause runaway synaptic weight changes at all neurons. To combat this, the Hebbian learning rule is modified using the BCM rule. As shown in equation 5.4, the BCM rule can be described as

$$\Delta\omega_{ij} = a_i a_j (a_j - \theta).$$

The BCM rule in the NEF is formed by substituting the Hebbian component from equation 6.20 into the BCM rule.

$$\Delta\omega_{ij} = \kappa\alpha_j a_i a_j (a_j - \theta) \tag{6.21}$$

Because the NEF is a spiking-neuron architecture, the θ parameter is taken to be the average of the filtered spike train of the postsynaptic neuron (see [34] for a biological justification of such a parameter). This average is computed over a long time frame (greater than 20ms) prior to the current time step. As shown in [6], this learning rule seems to have the effect of creating sparse representations in neural populations.

6.4.3 Combined learning rule

There is evidence to suggest that learning in some parts of the brain (including the brain region in the model described in this thesis) utilizes a combination of the learning methods described above (this will be discussed later in the thesis) [69]. To account for this phenomenon, a rule combining the error modulated learning rule and the BCM learning rule for the NEF was developed called the hPES (homeostatic Prescribed Error Sensitivity) rule [5] [6].

$$\Delta\omega_{ij} = \kappa\alpha_j a_i (S\mathbf{e}_j \cdot \mathbf{E} + (1 - S) a_j (a_j - \theta)) \tag{6.22}$$

This rule is simply a combination of the rules in equation 6.19 and equation 6.21 with the addition of a scaling factor S (where $0 <= S <= 1$) that determines what proportion of each rule contributes to the combined rule. The combined rule has been shown to minimize error as well as the error minimization rule alone, and has shown to be beneficial when learning nonlinear transformations [6].

Chapter 7

A new model of fear conditioning

Figure 7.1 shows the new NEF model of the fear conditioning circuit and is shown here to serve as a reference while reading the following sections. The structure, function, and implementation of the model will be explained in detail in this chapter.

7.1 Implementation

The model is organized as a collection of interacting NEF neuron populations as shown in figure 7.1. These populations correspond to regions of the PAG, amygdala, thalamus/cortex, and hippocampus.

The design of the model began by determining what computations were required to generate the behaviour seen in fear conditioning experiments. Much of this work had already been done in the development of the Krasne et al. model [35]; however, additional computations were required to maintain biological plausibility and to expand the capabilities of the model. The populations shown in figure 7.1 are the main functional populations in the model; however, there are some populations that contain sub-populations (not shown) that are either necessary to support the main function, or make the implementation of the model with the NEF easier. These include sub-populations for gating to ensure learning only occurs at the desired times, and sub-populations specifically for plastic connections in populations involved in learning.

Some of the populations in the model have known direct mappings to actual neural populations in the rat brain. These populations are the lateral amygdala (LA), the lateral basal area (BL), and the medial central nucleus (CEm). The remainder of the populations

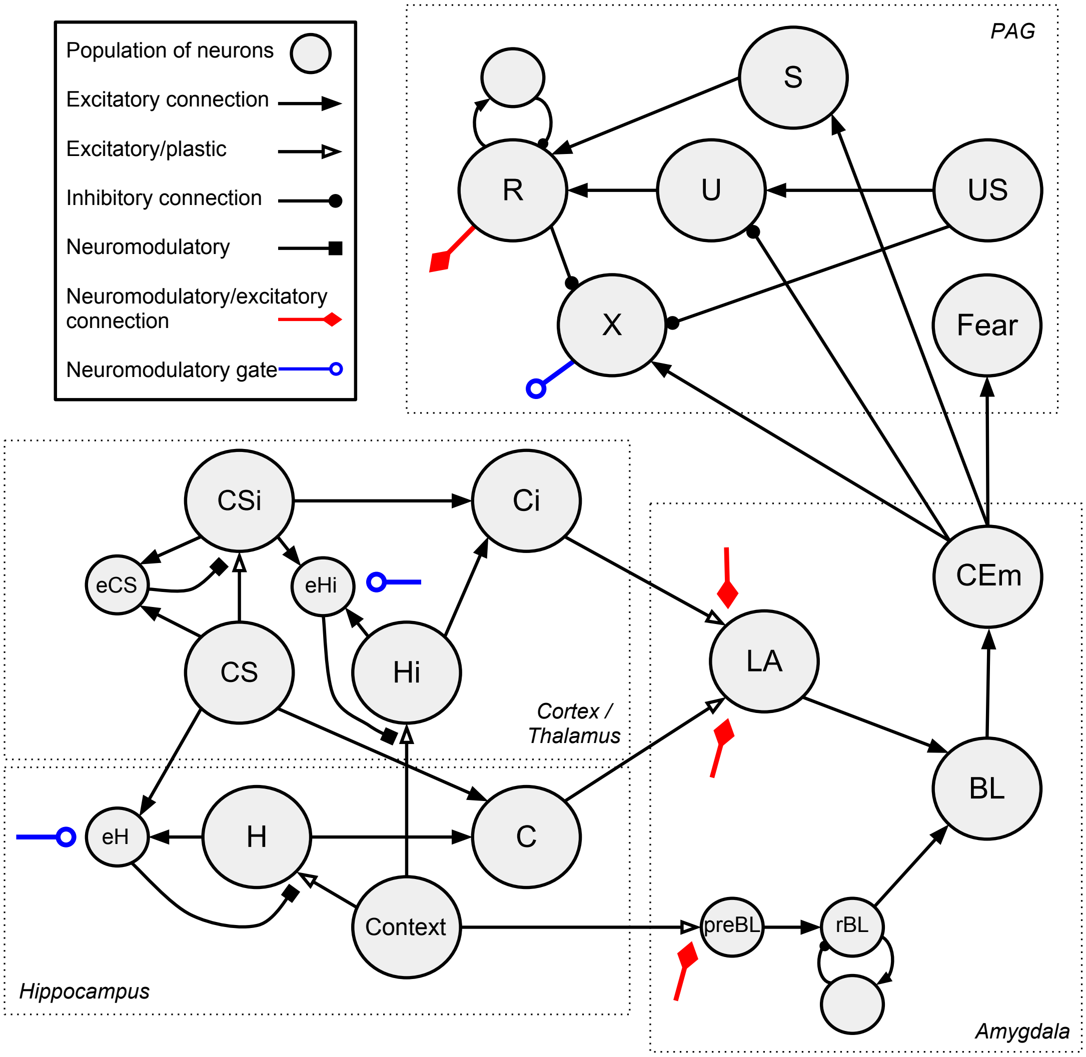

Figure 7.1: The proposed model. The connections between the anatomical regions, as outlined in 3.1, are shown in detail here along with the connections between sub-populations of these regions. Some of the population labels contain abbreviations: CS stands for conditioned stimulus, US for unconditioned stimulus, LA for lateral amygdala, BL for the lateral basal area, CEm for the medial central nucleus, i for image, r for recurrent, and e for error. The function of every neural population is explained in sections 7.2.1 through 7.2.6. Neural populations are explained in the sections corresponding to behaviours for which they are required.

do not correspond directly to an anatomically identified sub-population of neurons, and have been given arbitrary labels. However, even these populations implement functions that are thought to be performed in their respective anatomical region. These functions are all supported by biological data, and their implementations in this model should be thought of as specific computational proposals for how the brain could be performing the required computations. The S, R, U, X, and Fear populations share a similar function to their correlates in the Krasne et al. model.

The number of neurons assigned to each population depends on the value that the population is required to represent. For example, populations that represent a scalar value were assigned 100 neurons (to achieve a root-mean-square precision of around 1%), while populations representing three dimensional vectors were assigned 300 neurons (for the same precision in higher dimensions). These numbers do not reflect the actual number of neurons in these brain regions; they are only what is required to create a functioning model.

While all neurons making up the populations in the model are LIF neurons, their encoders, gains, and current biases (see chapter 6) are randomly generated within biologically plausible ranges at the start of every simulation. This means that neural populations are not identical from simulation to simulation. However, this does not have a significant impact on the function of the model, demonstrating the robustness of the design to specific neuron parameter values.

After the main anatomical areas had been divided into functional populations, the transformations between populations were specified. It is these transformations (computations on the values represented by populations of neurons) that govern the behaviour of the model. Some of these transformations are changed as the model learns through the course of a simulation; however, it is important to point out that the initial transformations between neural populations are exactly the same at the start of every simulated experiment performed. This means that no functional parameters of the model were changed to suit specific experiments. The function of the model is identical at the start of every simulation; it is the combination of inputs that it receives during the simulation that changes its function.

The model receives three different inputs representing CSs, contexts, and USs. The CS and context inputs are each three-dimensional vectors; each dimension can be thought of as representing one CS or one context. While three dimensions was chosen for these simulations, any number of dimensions could be used. It is likely that the brain can form associations between a wide variety of sights, sounds, textures, and environments; in order to account for the wide variety of possible inputs, the input dimensions would have to be increased substantially, and so would the number of neurons in downstream populations.

However, in order to demonstrate the chosen fear conditioning experiments, only three input dimensions were needed.

The model was generated and run using a software package called Nengo, which can be downloaded from http://nengo.ca, that allows users to create and simulate models through a user interface or through a scripting language. For this model, Nengo reads inputs from a file once the model has been generated. The times at which these inputs change values during the simulation are also read from the file. The input values and the times at which they change are chosen to correspond to the fear conditioning experiment being performed.

The data collected from the model simulations (shown in the plots in the following sections) includes the changing decoded values of various populations including the primary output population, which indicates the presence or absence of a fear response. These decoded values are filtered spike trains of the neurons in an NEF population. Different coloured lines in the same population represent the decoded values of different dimensions.

The x-axis of the results plots is time; however, the model has not been tested on precise timing results from actual behavioural experiments. Consequently, the x-axis is in seconds, but it can be thought of as a unit-less value used only to show the progression of time and the order of events.

7.2 Model description and simulations

In this section, the model will be explained in the context of the experiments it has been designed to model. Explanations of how the model is involved in each individual experiment are built upon explanations in previous sections, so it is recommended that the following sections are read in order. Simulation results will also be provided and discussed here. Detailed descriptions of the experiments modelled have been given in chapter 2.

7.2.1 First-order conditioning

As discussed in earlier sections, the inputs for first-order conditioning experiments are a CS and a US, and the output is a behavioural fear response (freezing). To explain how the model replicates the results of conditioning experiments, let us start by looking at the US population in figure 7.1. The US population represents a scalar input signal which is high in the presence of a footshock, and zero in the absence of a footshock. The value represented by the US population is projected through the U population to the R population. For now, ignore the purpose of the U population; for first-order conditioning it just relays the output of the US population on to the R population (we will also defer explanation of the projection from the US population to the X population until a later section). The R population represents an error signal. When its represented value is high, learning is enabled between CSs and the LA population.

Neurons in the amygdala have been shown to fire for only a short time at the onset of a US [30]. In order to replicate these findings, the R population of neurons is only excited by increases of its input. This is accomplished by a recurrent inhibition circuit that consists of the R population connected to an inhibitory population, which in turn inhibits the R population. At the onset of a US, this circuit is excited. After a short time - determined by the neurons' postsynaptic time constants - the R population will be inhibited, and the reinforcement signal will no longer be sent to the amygdala.

The CS population represents a high-dimensional (in these simulations, three-dimensional) vector that represents different input stimuli. The value represented by the CS population is projected to the LA population through the C population (for first-order fear conditioning, the C population simply relays the value represented by the CS population to the LA population). However, the connection between C and LA is not a direct communication channel. The weights of this connection undergo modification and account for learning in first-order conditioning experiments. Three populations play a role in this learning: the R, C, and LA populations. This learning circuit is shown in figure 7.2. It should be noted that LA represents a scalar value, not a high-dimensional value like C. In fear conditioning experiments, the activity in the lateral amygdala has been shown to increase after receiving paired US and CS inputs [59]. To account for this increase in activity, LA is modelled as representing a scalar activity level.

The R population affects learning in two ways: one is by providing the error signal referred to in section 6.4.1. Notice that there is no feedback between the output population, LA, and the error population, R. This means that if R is representing a high value, the weights of the connection between C and LA will be modified such that the value in LA is increased. The error signal does not decrease as the value in LA increases, so the value in LA will eventually be driven to saturation if R remains high for a long enough period of time. Also, the value represented by R can only be positive; this means that R will never be responsible for a decrease in the value represented by LA. Inhibition of LA activity is handled by another mechanism which will be discussed later.

After this learning process, the value represented in C (a CS) while the value represented in R was high will elicit an increase in the value represented in LA. It is important to

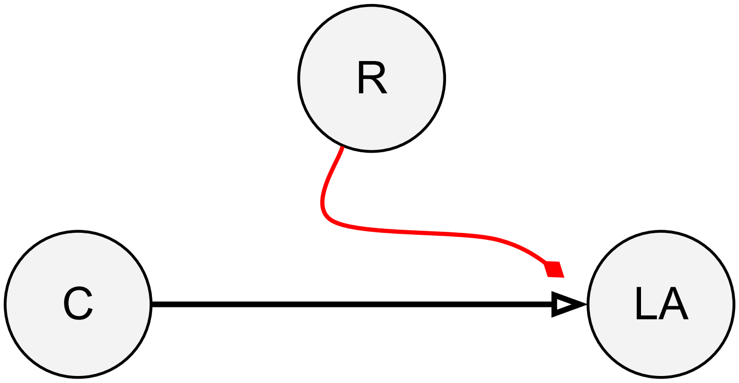

Figure 7.2: Learning sub-circuit from figure 7.1 for LA. The circuit allows for representations of CSs in C to activate the LA (lateral amygdala) in the presence of a reinforcing signal from R.

note that learning between the C population and the LA population only occurs if the C population is representing a vector with at least one non-zero value. If there is no CS present, the neurons in C will not be active, and a high value at R will not cause an increase in the value represented in LA via the error modulation mechanism.

The other way that the R population affects learning is through the Hebbian mechanism described in section 6.4.2. When the value represented by R increases, the value represented by LA also increases (albeit only slightly) because of the excitatory connection between the two populations. This activity changes the weights between R, C, and LA as described by equation 5.3.

The result of this R/C/LA learning circuit is that stimuli (represented in C) that are paired with the US (which causes the value represented by R to go high) gain the ability to increase the value represented in the LA. This learning is specific to the stimulus present during training; the synaptic weights are adjusted such that only the CS or CSs that were present during the reinforcing signal will be able to elicit the learned response in LA in the future. As seen in figure 7.1, LA projects to the BL population (which represents the lateral part of the basal amygdala), which projects to the CEm. The BL population represents a scalar value and, in the case of first-order conditioning, acts as a relay between the LA and the CEm.

Increases in CEm activity have been found to be closely correlated with fear responses [20]. The CEm is modelled as a neural population representing a scalar value. It projects to the Fear population, which also represents a scalar value that corresponds to freezing behaviour. The Fear population has a threshold; when the CEm reaches that threshold the Fear population goes high, at which point the animal freezes.

The above description of the pathways from US to LA, from CS to LA, and from LA to Fear accounts for first-order conditioning behaviour. Figure 7.3 shows the results of a first-order fear conditioning experiment simulated with this model.

7.2.2 Second-order conditioning

Second-order conditioning involves many of the same model populations as first-order conditioning. The main difference is that in second-order conditioning, a well conditioned CS is able to increase the value represented by the R population, and in turn, impact weight changes between the C and LA populations.

The CEm population has a projection to the S population, which has a projection to the R population. After training with a US and a CS (CS1), when CS1 is present in the C population, LA will go high even in the absence of a US. If the value in LA is large enough, CEm will go high and R will go high via the S population. This means that if there is another stimulus (CS2) represented at C (remember C represents high-dimensional information, so it can represent multiple stimuli at once), the weights between C and LA will be modified such that CS2 is able to elicit a fear response as well.

The inclusion of the S population was to ensure that second-order conditioning only occurs when a well-conditioned CS1 is present; the S population does not respond until its input has reached a certain threshold. A direct connection from the CEm to the R population would not have provided this functionality.

Figure 7.4 shows the results of a second-order fear conditioning experiment simulated with this model.

7.2.3 Sensory preconditioning

Sensory preconditioning requires that stimuli can be associated with each other before any pairing with a US occurs. To enable this, the CS population is connected to the CSi (i for ‘image’) population (representing values of the same dimensionality as the CS population) via a plastic connection (figure 7.5). The weights between these two populations of neurons are modified so that the value represented in CSi is the same as the value represented in CS. This is achieved using the error minimization learning rule for the NEF and having the

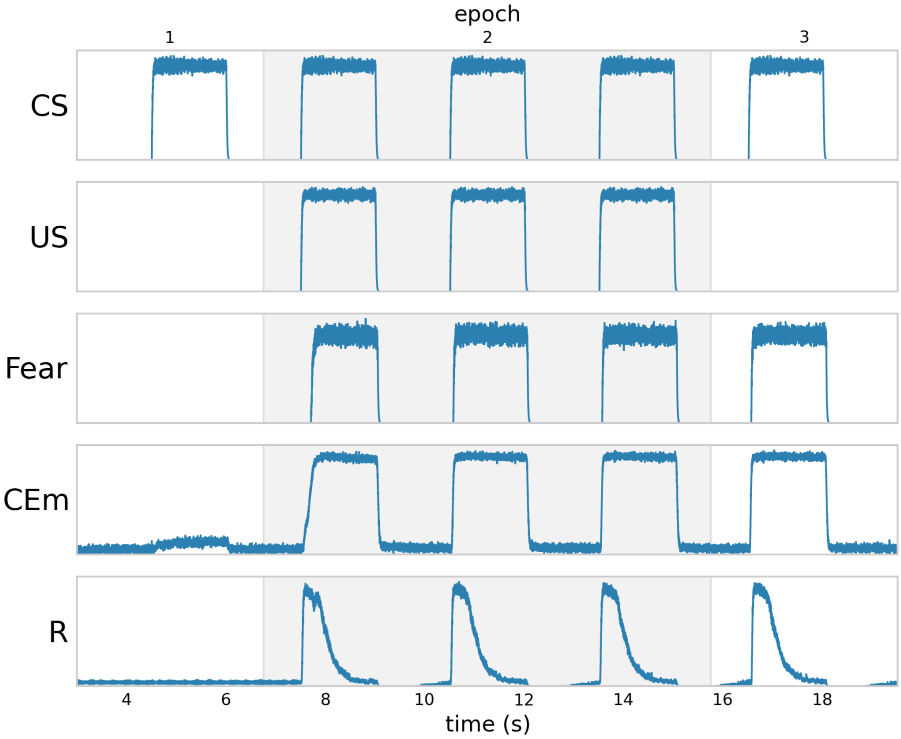

Figure 7.3: Simulation results for a first-order fear conditioning experiment. In epoch 1, a CS is presented to show that there is no fear response to the stimulus yet. In epoch 2, the CS is paired with the US a number of times. In epoch 3, when the CS is presented in the absence of the US, it elicits a fear response. Learning here occurred in one trial; the first pairing of the CS and US in epoch 2 elicited a fear response. The R population (the reinforcement signal) responds with each presentation of the US or a CS that has been sufficiently paired with a US.

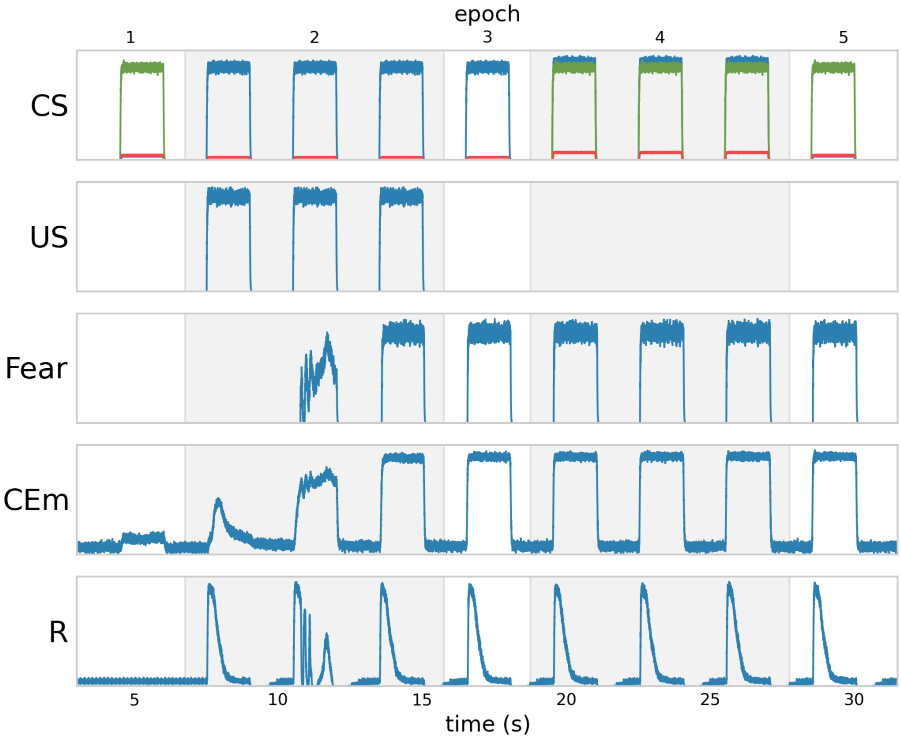

Figure 7.4: Simulation results for a second-order fear conditioning experiment. In epoch 1, a CS (green) is presented to demonstrate that it does not elicit a fear response. In epoch 2, a different CS (blue) is paired with the US multiple times. Epoch 3 shows that the blue CS is able to elicit a fear response after the training. In epoch 4, the blue CS and the green CS are paired together multiple times. In epoch 5, the green CS is now able to elicit a fear response in the absence of the blue CS or US even though it was never paired directly with the US.

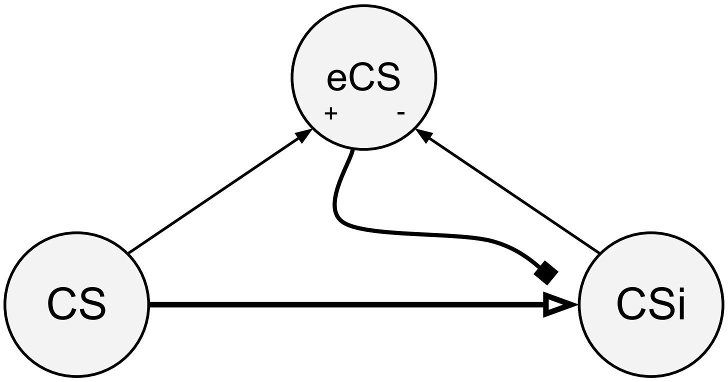

Figure 7.5: Learning sub-circuit from figure 7.1 for CS and CSi populations. The circuit allows for representations in CS to be mirrored in CSi.

error population (eCS: e for 'error') set to be the difference between the values represented in CSi and the values represented in CS (CS-CSi). When an input is first presented to CS, there will be a large discrepancy between the values represented by CS and CSi. This will cause the error signal to be large and the weights will be adjusted in order to minimize that discrepancy. Critically this error signal is of the same dimensionality as the values represented in CS and CSi, so weights between CS and CSi will be adjusted so that the value of each dimension of CSi will be the same as the value of the corresponding dimension of CS.

This connection allows for the association of stimuli. If the representation in CS indicates the presence of two stimuli, A and B (i.e., the values of two of the dimensions are high), then the weights between CS and CSi will be changed such that CSi has two dimensions high as well. In this case, the weights between the neurons activated by representation A in CS and the neurons in CSi will be adjusted such that the presence of A in CS causes the presence of A and B in CSi. Similarly, the weights between the neurons activated by representation B in CS and the neurons in CSi will be adjusted such that the presence of B in CS causes the presence of A and B in CSi. This is crucial to the process. Because of this type of weight modification, after pairing between A and B for a long enough time such that both CS and CSi represent high values of A and B, subsequent presentations of either A or B at CS will bring about high values of both A and B in CSi. Thus, after association, A at CS can bring about a representation of B (along with A) at CSi, and B at CS can bring about a representation of A (along with B) at CSi.

It is important to note that the ability for stimulus A or B to bring about a representation of the other stimulus will not last forever. If A is present at CS and A and B are present at CSi, there is a discrepancy between the values represented at the populations and the error minimization rule will begin to work towards making the representations at CS and CSi the same once again. However, before that process is complete, the stimuli will have the opportunity to affect fear conditioning in other parts of the model.

The CSi population is connected to the LA in the same way as the CS population; it goes through an intermediary Ci population. This Ci population is connected to the LA population in the same way as the C population is connected to LA, and the modification of the weights between Ci and LA is subject to the same learning rules as those between C and LA.

Here is an example of how this allows for sensory preconditioning. After two stimuli, A and B, have been present for some time, they become associated such that the presence of A or B at CS elicits a representation of both A and B at CSi (as explained above). If A is subsequently paired with a US, learning occurs at two places: between C and LA (a representation of A will come to cause a high value to be represented at LA), and between Ci and LA (a representation of A will come to cause a high value to be represented at LA, and a representation of B will come to cause a high value to be represented at LA). Subsequently, if B is present at CS, it will not cause LA to go high via its representation at C; however, it does elicit representations of both A and B at CSi and Ci, which in turn causes LA to go high.

Figure 7.6 shows the results of a sensory preconditioning experiment simulated with this model.

7.2.4 Blocking

The critical population in this model that allows blocking to occur is the U population. As discussed earlier, U acts as an intermediary between the US population and the R population. U also receives an inhibitory input from the CEm population. When the CEm reaches a certain level, it inhibits U, bringing the value that it represents down to zero regardless of what input it receives from US.

If a CS has been sufficiently paired with a US, it will cause the value represented by CEm to go high in the absence of the US. Because the R population only responds to changes in its input, it will go high for a period of time after the CS (CS1) is presented because of the connection from S, which receives input from CEm. If CS1 remains present, it will maintain a representation of a high value at CEm and cause inhibition of the U

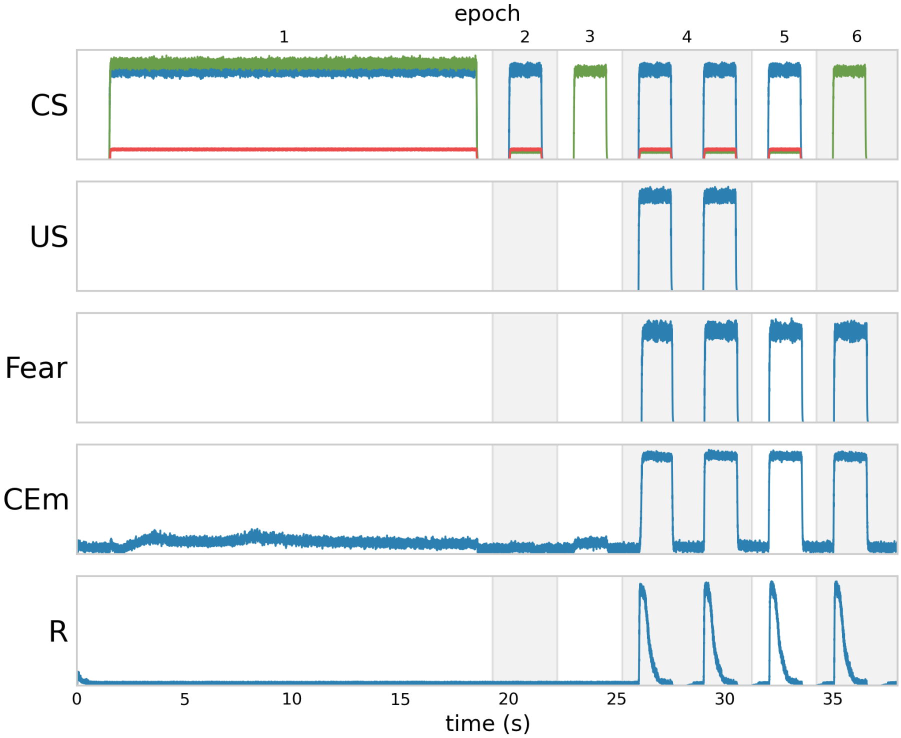

Figure 7.6: Simulation results for a sensory preconditioning fear conditioning experiment. In epoch 1, a green CS and a blue CS are paired for considerable time. Epoch 2 shows that the blue CS does not yet elicit a fear response, and epoch 3 shows that the green CS does not yet elicit a fear response. In epoch 4, the blue CS is paired with the US. Epoch 5 shows that the blue CS has gained the ability to elicit a fear response after the training. Epoch 6 shows that the green CS can elicit the fear response as well.

population. During the presence of the CS1, if a second CS (CS2) was paired with a US, no learning would occur between CS2 and the LA because the US will be unable to supply the activation of the R population that is needed for learning at LA. Hence, CS1 blocks CS2 from obtaining affective significance.

It should be mentioned that the traditional blocking experiments first performed by Kamin [31] included simultaneous presentations of a combined representation of CS1 and CS2 along with the US after CS1 had been conditioned. In this model, that experimental setup would induce second-order conditioning with the reinforcing signal being supplied by CS1 (see figure 7.4). However, blocking of US induced reinforcement signals would occur as well (see figure 7.7). In actual animal tests the balance between blocking and second-order reinforcement is more complex than this model can account for, and is likely affected by the timing of stimulus presentations and attentional effects [52]. The simulation demonstrating blocking was set up to eliminate second-order conditioning and demonstrate blocking of US reinforcing signals only.

Figure 7.7 shows the results of blocking in a fear conditioning experiment simulated with this model.

7.2.5 Context conditioning

Context conditioning with the model shares some similarities with first-order conditioning, but here the learning occurs between the Context population and a proposed preBL population. The Context population, like C, represents high dimensional information; preBL, like LA, represents a scalar value; and the value represented in R has the same effect on weight modification between the two populations as it does for the connection between C and LA.

The value represented in preBL passes through a population, rBL, with a recurrent inhibition connection before it reaches BL. This is done to replicate the finding that an animal does not maintain freezing behaviour for the entire time that it is in a context in which it received an aversive stimulus [14]. If there is a learned association between a context and a US via the Context/preBL connection weights, then the rBL population will subsequently only go high at the onset of the context. The recurrent inhibition connection here is essentially the same as the recurrent inhibition connection on the R population; however, the inhibition here is slower.

Figure 7.8 shows the results of a context fear conditioning experiment simulated with this model.

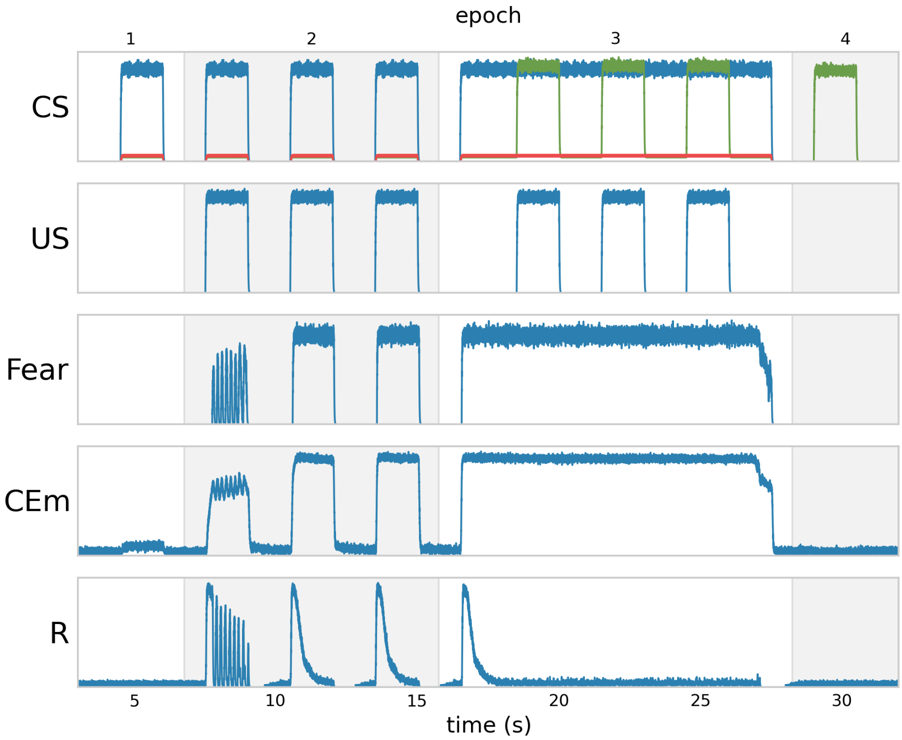

Figure 7.7: Simulation results for a blocking fear conditioning experiment. Epoch 1 demonstrates that the blue CS does not elicit a fear response. In epoch 2, the blue CS is paired with the US. At the start of epoch 3, the blue CS is introduced causing a fear response. While the blue CS is present, the green CS is paired with the US multiple times. Epoch 4 shows that despite the pairing in epoch 3, the green CS does not elicit a fear response.

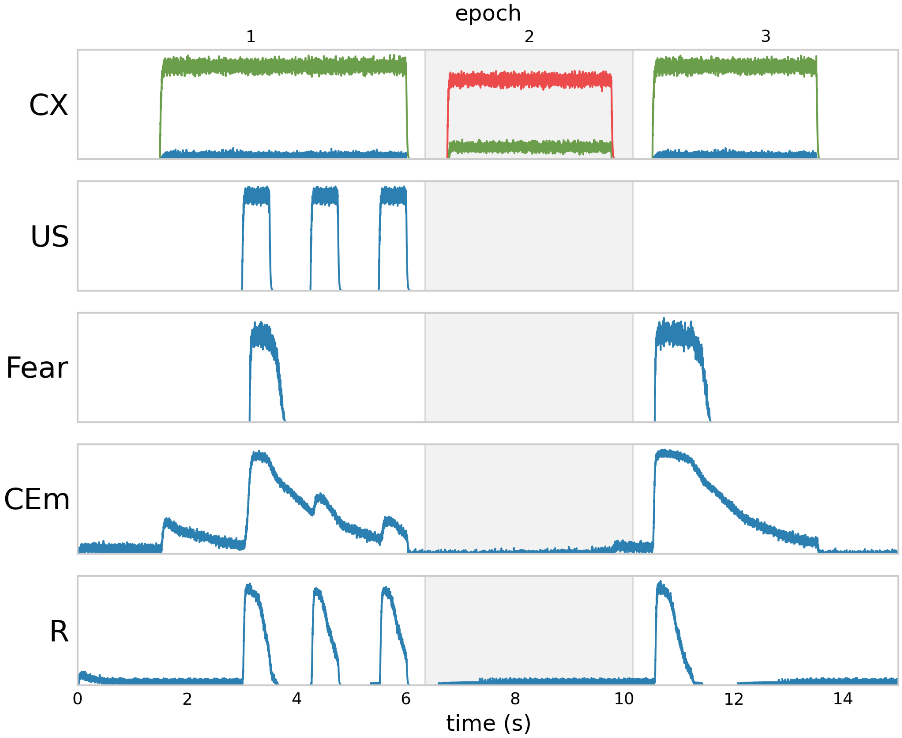

Figure 7.8: Simulation results for a context fear conditioning experiment. Before pairing with the US in epoch 1, the green context does not elicit a fear response. When paired with the US, a fear response is activated. CEm activity increases quickly, but decreases over time; in this model, contexts do not maintain a prolonged fear response. Epoch 2 shows that a red context does not elicit a fear response. At the reintroduction of the green context in epoch 3, a fear response is activated.

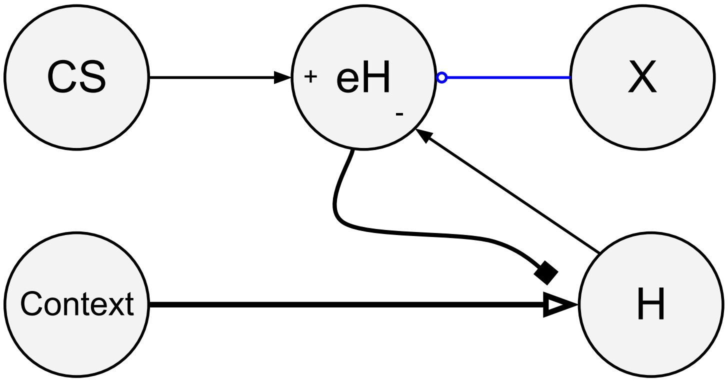

Figure 7.9: Learning sub-circuit from 7.1 for context-dependent extinction. The circuit allows contexts to evoke representations in H of CSs that they were previously paired with in the presence of a signal from X.

7.2.6 Extinction and renewal

Extinction and renewal with first-order conditioning

Context dependent extinction and renewal in this model is made possible by associating each context with the extinction state (whether the stimulus has been extinguished in that context or not) of every stimulus. This extinction state is saved if a stimulus that excites the CEm is present for a length of time in the absence of the US. Because the stimulus present is no longer predicting the occurrence of the US, it should lose its ability to activate the CEm.

The two additional populations used to allow for extinction and renewal are the H population and the eH (the e is for 'error') population. The Context population projects to the H population via plastic synaptic connections. The error population receives inputs from the CS population and the H population and calculates the difference between them (CS-H). The error population represents the same number of dimensions as CS and is therefore capable of adjusting the weights between Context and H such that each context brings about in H the representation present in CS at the time of learning. Figure 7.9 shows the learning circuit involved in this process.

Critically, the error population is gated by the X population. The X population receives excitatory input from CEm and inhibitory input from US and R. The result of these connections is that X is only high once R activity has decreased after a stimulus causes significant activity in CEm in the absence of the US. As discussed earlier, a CS that has gained the ability to excite the CEm will also excite the R population through the S population. The value represented by the R population only remains high for a short period of time after the presentation of the CS. During this time the X population is inhibited so that second-order conditioning is possible before extinction of the stimuli occurs.

When the value represented by the X population is high, it allows the eH population to modify the weights between Context and H. This is the time that the context learns what stimuli should be extinguished. After learning, the H population represents the stimuli to be extinguished, and projects that representation to the C population where it is subtracted from the projection from the CS population. If a CS (CS1) is represented in population H, the projection from H to C will inhibit the representation of CS1 coming from the CS population in population C, and prevent CS1 from increasing the value represented in LA.

It is important to remember that each different context will bring about a different extinction state in H depending on what extinction has occurred in that context in the past. This means that a particular CS could be extinguished in one context, but not in another. Figures 7.10, 7.11, and 7.12 show the results of model simulations using the different extinction and renewal schedules discussed earlier in section 2.5.1.

The extinction method described here is also applicable to representations in the CSi population. The Hi, and eHi populations have exactly the same function as the H and eH populations. Hi receives input from Context just like H. eHi receives input from CSi and Hi in the same way that eH receives input from CS and H. Hi projects to Ci just like H projects to C, and the representations from Hi and H both have the same impact on representations from Ci and C respectively. The purpose of extinction in the CSi circuit is explained in the next section.

Extinction and renewal with higher-order conditioning

As discussed in section 2.5, the results of extinction of a first-order conditioned CS on higher-order conditioned CSs differs depending on the method of higher-order conditioning. The way that higher-order conditioning is performed in the model accounts for this observation.

In second-order conditioning, a CS2 is capable of eliciting a fear response because it was associated with the fear response brought about by a CS1 that was previously associated with a US. The ability for CS2 to elicit a fear response is the result of the modification of weights between C and LA. When CS1 and CS2 were both present at CS, CS1 caused

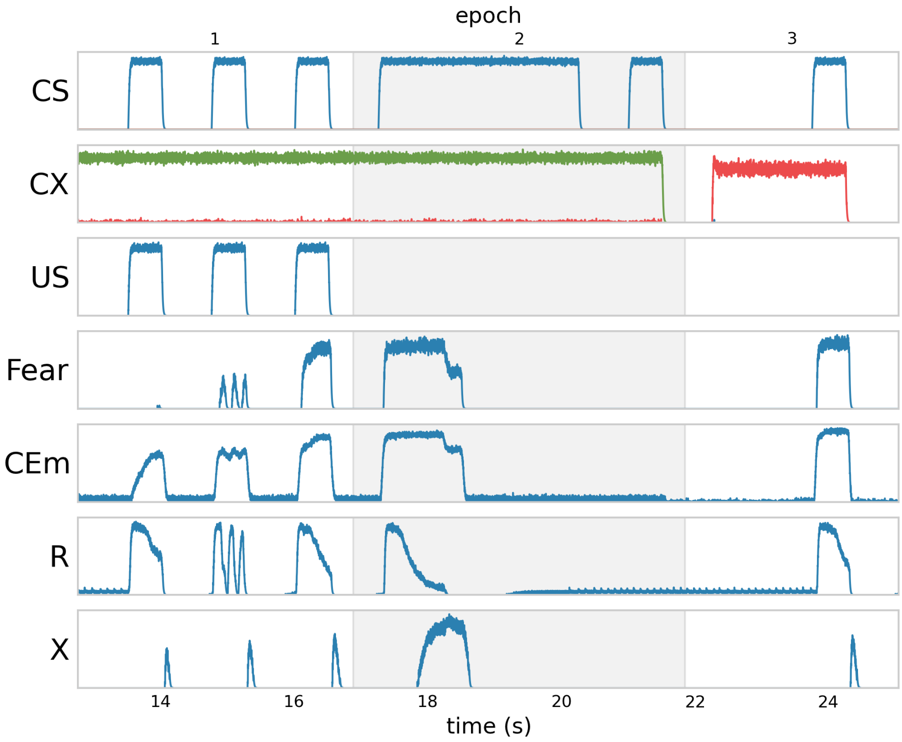

Figure 7.10: Simulation results for an AAB extinction and renewal experiment. In epoch 1, the blue CS is paired with the US while in the green context. At the start of epoch 2, the blue CS is able to elicit a fear response in the green context. After some time being present in the absence of the US, the CS is extinguished in the green context. The extinction process begins at the onset of activity in the X population. The blue CS is presented again at the end of epoch 2 to demonstrate that it has been extinguished in the green context. In epoch 3, we move to the red context. The same blue CS is presented in the red context, but is renewed and elicits a fear response.

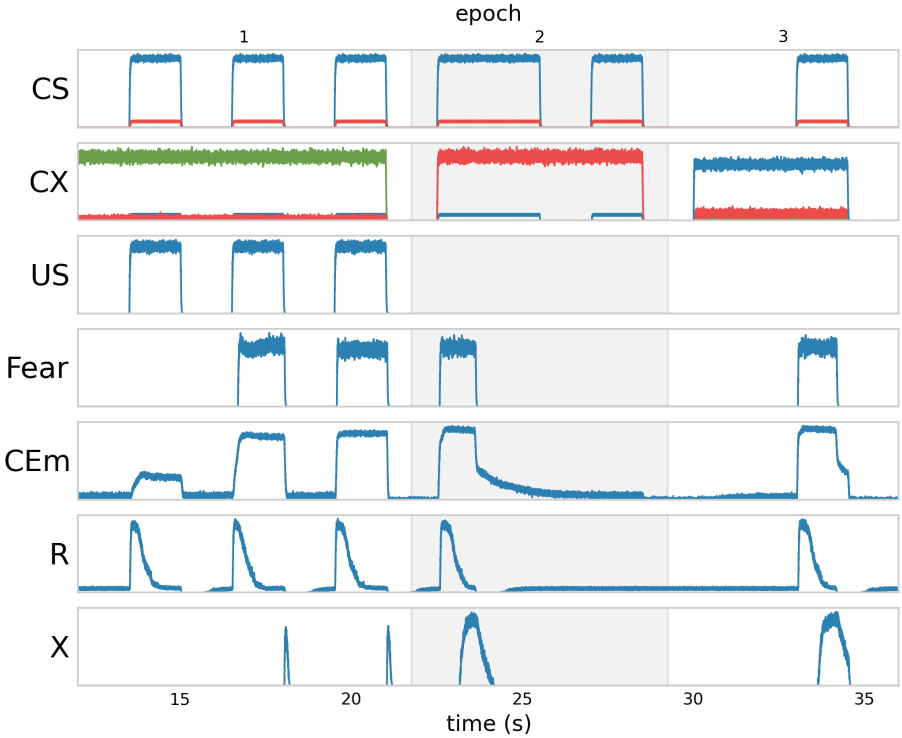

Figure 7.11: Simulation results for an ABC extinction and renewal experiment. The blue CS is paired with the US in the green context in epoch 1. In epoch 2, the blue CS is able to elicit a fear response in the red context, but is eventually extinguished. It is presented at the end of epoch 2 to demonstrate that it is no longer able to elicit a fear response in the red context. In epoch 3, we move to the blue context, in which the blue CS is capable of eliciting a fear response.

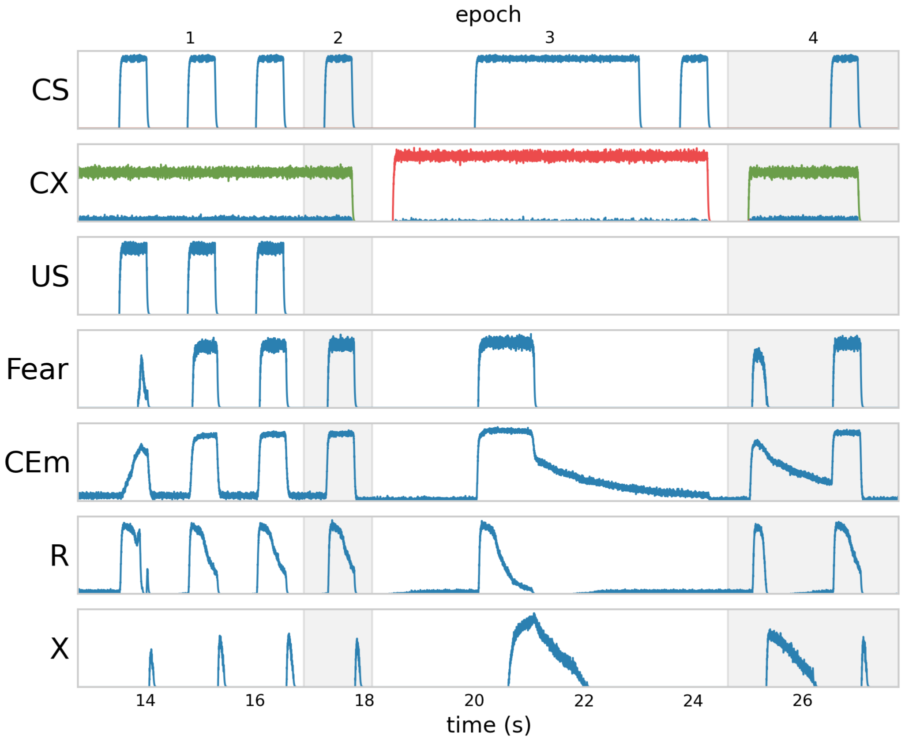

Figure 7.12: Simulation results for an ABA extinction and renewal experiment. In epoch 1, the blue CS is paired with the US in the green context. Epoch 2 shows that the blue CS is now capable of eliciting a fear response. In epoch 3, we change to the red context. The blue CS is presented, and after a while the X population initiates extinction. A presentation of the blue CS shows that it has been extinguished in the red context. In epoch 4, we move to the green context. The green context itself is able to elicit a fear response. Furthermore, the blue CS that was extinguished in the red context elicits a fear response in the green context.

activity in R to go high, which then facilitated the weight modification between C and LA that allowed a representation of CS2 at C to bring about increased activity in LA.

If CS1 is subsequently presented and extinguished in a particular context, there will be an inhibiting representation at C from H in that context that prevents CS1 from activating LA. However, a presentation of CS2 at C will not be inhibited (given that CS2 was not present at the same time that CS1 was extinguished). The result is that extinguishing CS1, the first-order conditioned CS, does not affect the ability of CS2, the higher-order conditioned CS, to elicit a fear response. Figure 7.13 shows the results of a simulation demonstrating this kind of extinction.

In sensory preconditioning, CS2 is capable of eliciting a fear response because of its association with CS1 before CS1 was associated with the US. In this case, a representation of CS2 at C does not increase activity in LA. It is the representation of CS2 at Ci that increases activity in LA. If CS1 is presented, it brings about a representation of both CS1 and CS2 at the CSi population. If CS1 is present for a long enough time without being accompanied by a US, the extinction state represented in Hi will include both CS1 and CS2. Subsequent presentations of CS2 will project to Ci, but will be inhibited by the extinction state from Hi. The result is that extinguishing CS1, the first-order conditioned CS, does affect the ability of CS2, the higher-order conditioned CS, to elicit a fear response. Figure 7.14 shows the results of a simulation demonstrating this kind of extinction.

The descriptions in this chapter have shown how the model is capable of reproducing the results of fear conditioning experiments demonstrating first and second-order conditioning, sensory preconditioning, blocking, and extinction and renewal - both first-order and higher-order. Moreover, the model achieves this using spiking neuron populations corresponding to known anatomical regions. This combination of behavioural and biological detail is unique in fear conditioning models and provides advantages over other approaches. These advantages will be discussed in the next chapter.

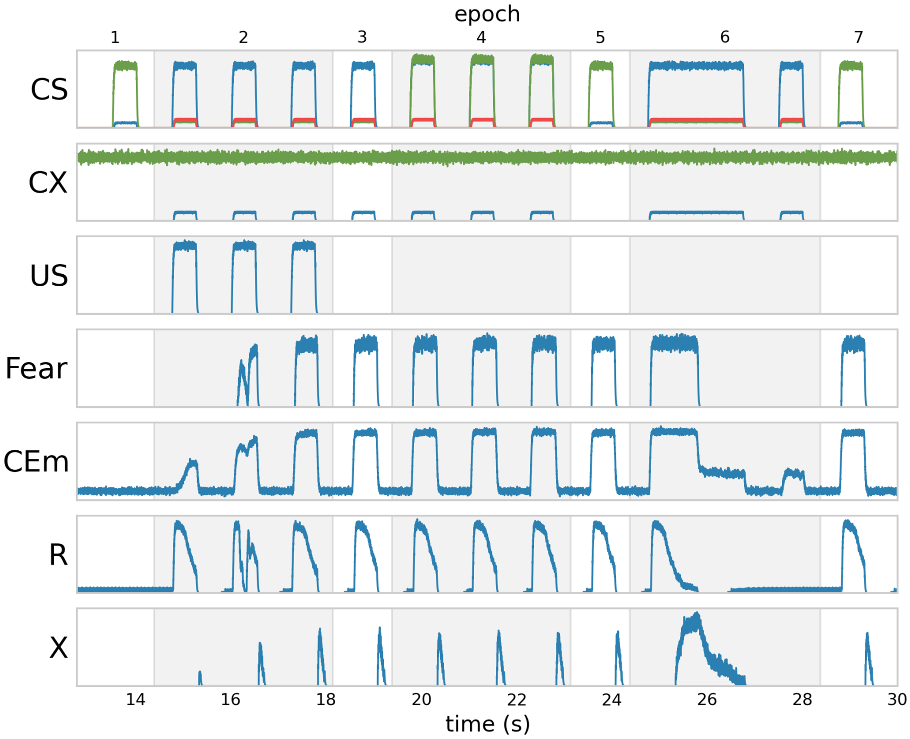

Figure 7.13: Simulation results for a second-order conditioning extinction and renewal experiment. Epoch 1 shows that the green CS is not able to elicit a fear response. In epoch 2, the blue CS is paired with the US. Epoch 3 shows that the training has allowed the blue CS to elicit a fear response. In epoch 4, the blue CS is paired with the green CS multiple times. Epoch 5 shows that the green CS is now capable of eliciting a fear response. In epoch 6, the blue CS is extinguished. Despite the extinction of the blue CS, the green CS elicits a fear response in epoch 7.

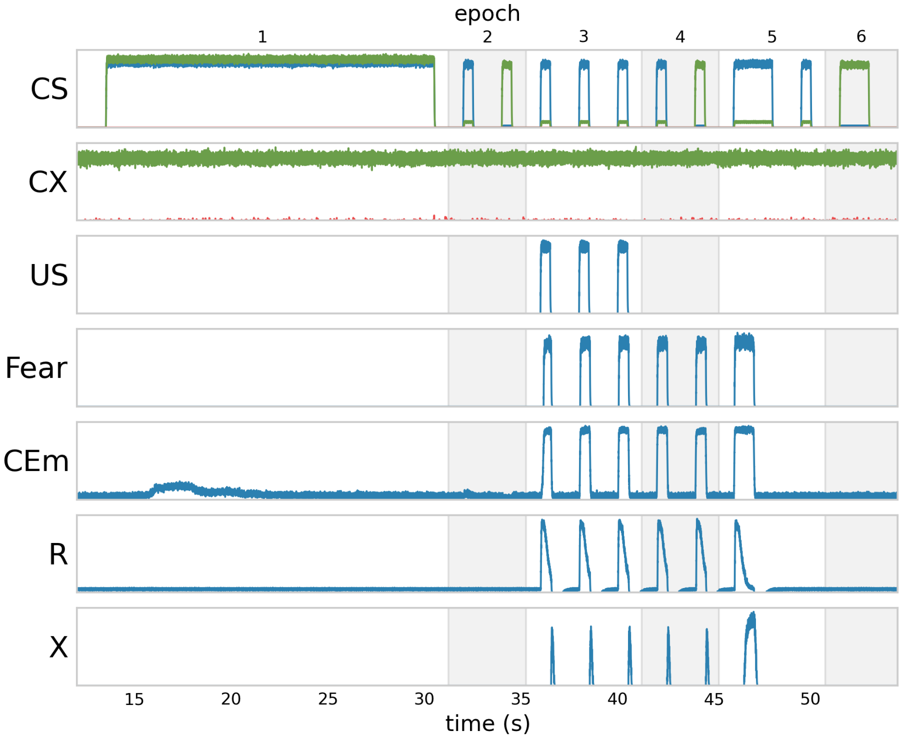

Figure 7.14: Simulation results for a sensory preconditioning extinction and renewal experiment. In epoch 1, the blue and green CSs are paired for considerable time. Epoch 2 shows that neither the blue or green CS is capable of eliciting a fear response. In epoch 3 the blue CS is paired with the US. Epoch 4 shows that both of the CSs are capable of eliciting a fear response. In epoch 5, the blue CS is extinguished. Because of the extinction of the blue CS in epoch 5, the green CS cannot elicit a fear response in epoch 6.

Chapter 8

Discussion and conclusions

The primary contribution of this thesis is demonstrating the ability to construct an integrative model of fear conditioning. To my knowledge, the model presented here is the first to reproduce high-level behavioural data from such a wide range of fear conditioning experiments with a spiking neuron model adhering to mammalian anatomy and known neuroscientific data. Some of the differences between this model and the other fear conditioning models were pointed out in section 4.6, but it is useful to now draw attention to the main differences between it (an NEF-based model) and the Krasne et al. model on which much of its structure was based.

One primary conceptual and implementational difference involves how high-dimensional data is represented in the models. In the Krasne et al. model, each dimension (a cue or a context) is assigned to one idealized neuron. In the NEF, a population of neurons can represent many dimensions, and each dimension need not be assigned to particular neurons; rather, the NEF employs distributed representations so that each neuron in a population can represent a part of the population's high-dimensional input. This is the more biologically plausible solution [65], and has practical implications for the development of the model. For example, when adding a new context or cue, the connections (including the plastic connections) in the model can be left as they are. This allows us to avoid the trouble of reorganizing a model just because we want it to be able to handle more complex representations. Furthermore, with this approach the representational capacity scales exponentially with the number of dimensions represented [21]. The result is that models developed using the NEF can easily be scaled up into more complex models.

Another significant difference between the two models is how they are constrained by biology. By restricting the model to spiking neurons, we are forced to implement functions in biologically plausible ways that may not be the easiest or most intuitive. For example, the model requires that the activities of the R and rBL populations increase only at the onset of activity from their inputs. There are different ways to model this behaviour mathematically, but by restricting ourselves to spiking neurons, some implementations seem more likely than others. For instance, we know that recurrent inhibitory connections exist in the brain (e.g. [19]), so this seems like a reasonable way to implement the function. It may not be the correct implementation, but it provides a more reasonable starting place than specification of an arbitrary mathematical function that does not have the implementational constraints of spiking neurons.

The biological constraint of spiking neurons also affects how learning is done in this model. A learning rule that works with spiking neurons is employed, and again, because we are forced to use a method that works with more realistic neurons, we expect that the proposed mechanism is more likely to be correct.

The performance of the NEF model can also be compared to a wider range of experimental data than the Krasne et al. model. Because many mathematical functions are straight forward to implement using the NEF, high-level behavioural data can be matched by the model (as was shown in this thesis). And because these functions are implemented in spiking neurons, we can also compare the results of the model to electrophysiological data, or other low-level neuroscientific data. The model presented here has not yet been analyzed at that level, but the framework used will allow for that kind of comparison as the model is further developed.

One last notable advantage of the model is that it allows us to make more detailed predictions. Although informed by anatomy, the model takes many informed guesses as to how a neural system could implement the functions needed to explain results from fear conditioning experiments. These guesses have been implemented in sufficient detail so as to allow for validation in neuroscientific experiments. For example, the organization of populations in the hippocampus is, to my knowledge, novel. Experiments could be done to test if there is a population of neurons like the H population that has significant activity after extinction events, and that inhibits activity in another population of neurons, which could be a correlate of the C population.

The design of such experiments may be difficult; it is likely that a population of neurons that performs a specific function may be dispersed throughout a brain region. It may also be the case that the brain performs the functions required for fear conditioning in completely different ways. For example, the complex neurons of the brain may be able to perform with one neuron what was in this model proposed with several populations. That being said, testing the theories proposed by this model may be a good place to start.

The limitations of this model also need to be mentioned here. One of the most significant limitations is that the model as implemented here does not provide a detailed account of timing. For example, in the model, the number of times a CS is paired with a US affects how well conditioning will proceed; however, this number is not matched to actual biological data. Additionally, in experiments, the length of time before or after the occurrence of a US that a CS is presented affects conditioning, but this effect was not modelled here.

Another limitation of the model is that it essentially uses only one type of neuron. The majority of LIF neurons used here can be thought of as excitatory pyramidal neurons. However, the brain contains many different types of neurons with different characteristics related to plasticity, neuromodulators, etc., which may have a subtle, or significant impact on the information processing in the circuit. This model may be useful for investigating where specific types of neurons might be involved in the fear conditioning circuit; the functions that this model performs could be found to be better suited to certain classes of neurons.

One last limitation of the model should be noted; it constrains itself to only certain anatomical regions of the brain. There are other brain regions, not modelled here, that likely play a role in the fear conditioning circuit: most notably, the PFC [50] [47]. However, developing a highly functional model which utilizes only a few anatomical areas is a good place to start exploring the potential role of other brain regions. Neuroscientific evidence may later suggest that certain functions of the circuit are better mapped to other brain regions, or that more advanced operation of the fear conditioning circuit requires the use of other brain regions. Changes to this model can be made in order to test these theories.

Future work on this project should include testing the theories proposed by the model, as well as addressing its limitations. More detail can be added to the model to match the current experimental data on the fear condoning circuit, and in the future the model can be expanded to account for new experimental data that will surely be generated. Inevitably, as the model is grown to account for more phenomena, other brain regions and functional circuits will need to be recruited. Continuing to develop such a model will hopefully give us a better look into the effects of fear on the brain and help us understand the role of emotional processing in general.

References

[1] Alison Abbott. Laboratory animals: The Renaissance rat. *Nature*, 428(6982):464–466, 2004.

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

[3] Bardia F Behabadi, Alon Polsky, Monika Jadi, Jackie Schiller, and Bartlett W Mel. Location-dependent excitatory synaptic interactions in pyramidal neuron dendrites. *PLoS computational biology*, 8(7):e1002599, 2012.

[4] Michael M Behbehani. Functional characteristics of the midbrain periaqueductal gray. *Progress in neurobiology*, 46(6):575–605, 1995.

[5] Trevor Bekolay. Learning in large-scale spiking neural networks. Master's thesis, University of Waterloo, 2011.

[6] Trevor Bekolay, Carter Kolbeck, and Chris Eliasmith. Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 169–174, 2013.

[7] R.L. Berkowitz, J.D. Coplan, D.P. Reddy, and J.M. Gorman. The human dimension: how the prefrontal cortex modulates the subcortical fear response. *Rev Neurosci*, 18(3-4):191–207, 2007.

[8] Elie L Bienenstock, Leon N Cooper, and Paul W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, 2(1):32–48, 1982.

[9] Robert J Blanchard and D Caroline Blanchard. Crouching as an index of fear. *Journal of comparative and physiological psychology*, 67(3):370, 1969.

[10] M. Bouton and R. Bolles. Contextual control of the extinction of conditioned fear. *Learning and Motivation*, 10:445–466, 1979.

[11] Judson S. Brown, Harry I. Kalish, and I. E. Farber. Conditioned fear as revealed by magnitude of startle response to an auditory stimulus. *Journal of Experimental Psychology*, 41:317–328, 1951.

[12] P Brown. Physiology of startle phenomena. *Adv Neurol.*, 67:273–287, 1995.

[13] Byron A Campbell and Julian Jaynes. Reinstatement. *Psychological Review*, 73(5):478, 1966.

[14] Pascal Carrive. Conditioned fear to environmental context: cardiovascular and behavioral components in the rat. *Brain research*, 858(2):440–445, 2000.

[15] Mircea I Chelaru and Valentin Dragoi. Efficient coding in heterogeneous neuronal populations. *Proceedings of the National Academy of Sciences*, 105(42):16344–16349, 2008.

[16] Feng-Xuan Choo. The ordinal serial encoding model: Serial memory in spiking neurons. Master's thesis, University of Waterloo, 2010.

[17] J Debiec and JE LeDoux. Disruption of reconsolidation but not consolidation of auditory fear conditioning by noradrenergic blockade in the amygdala. *Neuroscience*, 129(2):267–272, 2004.

[18] IT Diamond and WC Hall. Evolution of neocortex. *Science (New York, NY)*, 164(3877):251, 1969.

[19] Valentin Dragoi and Mriganka Sur. Dynamic properties of recurrent inhibition in primary visual cortex: contrast and orientation dependence of contextual effects. *Journal of Neurophysiology*, 83(2):1019–1030, 2000.

[20] Sevil Duvarci, Daniela Popa, and Denis Paré. Central amygdala activity during fear conditioning. *The Journal of Neuroscience*, 31(1):289–294, 2011.

[21] C. Eliasmith and C. Anderson. *Neural Engineering: Computation, representation, and dynamics in neurobiological systems.* MIT Press, Cambridge, 2003.

[22] Michael S. Fanselow and Robert C. Bolles. Naloxone and shock-elicited freezing in the rat. *Journal of Comparative and Physiological Psychology*, 93:736–744, 1979.

[23] JM Farook, Q Wang, SM Moochhala, ZY Zhu, L Lee, and PT-H Wong. Distinct regions of periaqueductal gray (pag) are involved in freezing behavior in hooded pvg rats on the cat-freezing test apparatus. *Neuroscience letters*, 354(2):139–142, 2004.

[24] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[25] J. C. Gewirtz. Using Pavlovian Higher-Order Conditioning Paradigms to Investigate the Neural Substrates of Emotional Learning and Memory. *Learning & Memory*, 7:257–266, 2000.

[26] Stephen Grossberg and Daniel S Levine. Neural dynamics of attentionally modulated pavlovian conditioning: blocking, interstimulus interval, and secondary reinforcement. *Applied optics*, 26(23):5015–5030, 1987.

[27] Benjamin Harris. Whatever Happened to Little Albert? *American Psychologist*, 34:151–160, 1979.

[28] Donald Olding Hebb. *The organization of behavior: A neuropsychological approach*. John Wiley & Sons, 1949.

[29] Erich D Jarvis. Evolution of the pallium in birds and reptiles. In *Encyclopedia of Neuroscience*, pages 1390–1400. Springer, 2009.

[30] Joshua P Johansen, Jason W Tarpley, Joseph E LeDoux, and Hugh T Blair. Neural substrates for expectation-modulated fear learning in the amygdala and periaqueductal gray. *Nature Neuroscience*, 13(8):979–86, 2010.

[31] Leon Kamin. Predictability, surprise, attention, and conditioning. In B A Campbell and R M Church, editors, *Punishment and aversive behavior*, pages 279–296. Appleton-Century-Crofts, New York, 1969.

[32] Eric R Kandel. The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, 294(5544):1030–1038, 2001.

[33] Eric R Kandel, James H Schwartz, Thomas M Jessell, et al. *Principles of neural science*, volume 4. McGraw-Hill New York, 2000.

[34] Alfredo Kirkwood, Marc G Rioult, and Mark F Bear. Experience-dependent modification of synaptic plasticity in visual cortex. *Nature*, 381(6582):526–528, 1996.

[35] Franklin B Krasne, Michael S Fanselow, and Moriel Zelikowsky. Design of a neurally plausible model of fear learning. *Frontiers in behavioral neuroscience*, 5, 2011.

[36] Joseph E LeDoux. Emotion: Clues from the brain. *Annual review of psychology*, 46(1):209–235, 1995.

[37] Joseph E LeDoux. *The emotional brain: The mysterious underpinnings of emotional life.* Simon & Schuster, 1996.

[38] Joseph E LeDoux, Piera Cicchetti, Andrew Xagoraris, and Lizabeth M Romanski. The lateral amygdaloid nucleus: sensory interface of the amygdala in fear conditioning. *The Journal of neuroscience*, 10(4):1062–1069, 1990.

[39] Joseph E LeDoux, Claudia R Farb, and Lizabeth M Romanski. Overlapping projections to the amygdala and striatum from auditory processing areas of the thalamus and cortex. *Neuroscience letters*, 134(1):139–144, 1991.

[40] Joseph E LeDoux, J Iwata, PRDJ Cicchetti, and DJ Reis. Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *The Journal of Neuroscience*, 8(7):2517–2529, 1988.

[41] 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.

[42] David MacNeil and Chris Eliasmith. Fine-tuning and the stability of recurrent neural networks. *PloS One*, 6(9), 2011.

[43] Stephen Maren. Neurobiology of pavlovian fear conditioning. *Annual review of neuroscience*, 24(1):897–931, 2001.

[44] Stephen Maren and Michael S Fanselow. Synaptic plasticity in the basolateral amygdala induced by hippocampal formation stimulation in vivo. *The Journal of neuroscience*, 15(11):7548–7564, 1995.

[45] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[46] Frances K McSweeney and Calvin Bierley. Recent developments in classical conditioning. *Journal of Consumer Research*, pages 619–631, 1984.

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

[48] Shigeru Miyagawa, Robert C Berwick, and Kazuo Okanoya. The emergence of hierarchical structure in human language. *Frontiers in psychology*, 4, 2013.

[49] Marta AP Moita, Svetlana Rosis, Yu Zhou, Joseph E LeDoux, and Hugh T Blair. Hippocampal place cells acquire location-specific responses to the conditioned stimulus during auditory fear conditioning. *Neuron*, 37(3):485–497, 2003.

[50] Maria A Morgan and Joseph E LeDoux. Contribution of ventrolateral prefrontal cortex to the acquisition and extinction of conditioned fear in rats. *Neurobiology of learning and memory*, 72(3):244–251, 1999.

[51] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*, volume 3. Clarendon Press Oxford, 1978.

[52] Diana B Padlubnaya, Nirav H Parekh, and Thomas H Brown. Neurophysiological theory of kamin blocking in fear conditioning. *Behavioral neuroscience*, 120(2):337, 2006.

[53] Jaak Panksepp and Lucy Biven. *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. WW Norton & Company, 2012.

[54] Shauna L Parkes and R Frederick Westbrook. The basolateral amygdala is critical for the acquisition and extinction of associations between a neutral stimulus and a learned danger signal but not between two neutral stimuli. *The Journal of Neuroscience*, 30(38):12608–12618, 2010.

[55] I.P. Pavlov and G.V. Anrep. *Conditioned reflexes*. Dover Publications, Incorporated, 1927.

[56] Verena Pawlak, Jeffery R Wickens, Alfredo Kirkwood, and Jason ND Kerr. Timing is not everything: neuromodulation opens the stdp gate. *Frontiers in synaptic neuroscience*, 2, 2010.

[57] Marie A Pezze and Joram Feldon. Mesolimbic dopaminergic pathways in fear conditioning. *Progress in neurobiology*, 74(5):301–320, 2004.

[58] R. G. Phillips and J. E. LeDoux. Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience*, 106:274–285, 1992.

[59] Gregory J Quirk, J Christopher Repa, and Joseph 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.

[60] Pedro Rada, NM Avena, and BG Hoebel. Daily bingeing on sugar repeatedly releases dopamine in the accumbens shell. *Neuroscience*, 134(3):737–744, 2005.

[61] R. A. Rescorla. *Experimental extinction*, pages 119–154. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2001.

[62] R. A. Rescorla and A. W. Wagner. *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*, chapter 3, pages 64–99. Appleton-Century-Crofts, New York, 1972.

[63] Ross C Rizley and Robert A Rescorla. Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*, 81(1):1, 1972.

[64] Trevor W Robbins and Barry J Everitt. Neurobehavioural mechanisms of reward and motivation. *Current opinion in neurobiology*, 6(2):228–236, 1996.

[65] Yoshio Sakurai. Population coding by cell assemblies - what it really is in the brain. *Neuroscience research*, 26(1):1–16, 1996.

[66] Harold Schlosberg. Conditioned responses in the white rat: II. conditioned responses based upon shock to the foreleg. *The Pedagogical Seminary and Journal of Genetic Psychology*, 49(1):107–138, 1936.

[67] Wolfram Schultz and Anthony Dickinson. Neuronal coding of prediction errors. *Annual review of neuroscience*, 23(1):473–500, 2000.

[68] NRW Selden, BJ Everitt, LE Jarrard, and TW Robbins. Complementary roles for the amygdala and hippocampus in aversive conditioning to explicit and contextual cues. *Neuroscience*, 42(2):335–350, 1991.

[69] Torfi Sigurdsson, Valérie Doyère, Christopher K Cain, and Joseph E LeDoux. Long-term potentiation in the amygdala: a cellular mechanism of fear learning and memory. *Neuropharmacology*, 52(1):215–227, 2007.

[70] Marius C Smith et al. Cs-us interval and us intensity in classical conditioning of the rabbits nictitating membrane response. *Journal of Comparative and Physiological Psychology*, 66(3):679–687, 1968.

[71] Brian L Thomas, Niccole Larsen, and John JB Ayres. Role of context similarity in aba, abc, and aab renewal paradigms: Implications for theories of renewal and for treating human phobias. *Learning and Motivation*, 34(4):410–436, 2003.

[72] 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), 2011.

[73] John B. Watson and Rosalie Rayner. Conditioned emotional reactions. *Journal of Experimental Psychology*, 3(1):1–14, 1920.