A Biologically Plausible Spiking Neuron Model of Fear Conditioning

Carter Kolbeck (ckolbeck@uwaterloo.ca)
Trevor Bekolay (tbekolay@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada, N2L 3G1

Abstract

Reinforcement learning based on rewarding or aversive stimuli is critical to understanding the adaptation of cognitive systems. One of the most basic and well-studied forms of reinforcement learning in mammals is found in fear conditioning. We present a biologically plausible spiking neuron model of mammalian fear conditioning and show that the model is capable of reproducing the results of four well known fear conditioning experiments (conditioning, second-order conditioning, blocking, and context-dependent extinction and renewal). The model contains approximately 2000 spiking neurons which make up various populations of primarily the amygdala, periaqueductal gray, and hippocampus. The connectivity and organization of these populations follows what is known about the fear conditioning circuit in mammalian brains. Input to the model is made up of populations representing sensory stimuli, contextual information, and electric shock, while the output is a population representing an autonomic fear response: freezing. Using a novel learning rule for spiking neurons, associations are learned between cues, contexts, and the aversive shock, reproducing the behaviors seen in rats during fear conditioning experiments.

Keywords: Fear conditioning; learning; amygdala; neural engineering

Introduction

Fear conditioning is a widely studied paradigm in many areas of cognitive science. Thanks to its well-defined inputs and outputs, fear conditioning offers researchers one the most effective methods of examining learning, memory, and emotional processing in animal brains. The new model we propose here addresses four of the most well-known phenomena related to fear conditioning.

Conditioning: Associations are formed between a neutral stimulus (NS) and an aversive unconditioned stimulus (US) such that the NS comes to elicit a fear response.

Second-Order Conditioning: Associations are formed between a conditioned stimulus (CS) and an NS such that the NS comes to elicit a fear response.

Blocking: The presence of a strong CS prevents conditioning of an NS that is temporally paired with a US.

Context-Dependent Extinction and Renewal: A CS that is repeatedly presented in the absence of a US will lose its ability to elicit a fear response in the context in which it is extinguished.

Mathematical models reproducing these tasks have been previously demonstrated (e.g., Grossberg & Levine, 1987). However, models employing biologically plausible mechanisms are generally lacking.

The model presented here is a significant extension of a non-spiking neuron model previously developed by Krasne et al. (Krasne, Fanselow, & Zelikowsky, 2011). One difference between our model and the one developed by Krasne et al. is the level of biological plausibility in the mechanisms employed; our model is built using a spiking neuron modeling framework (the Neural Engineering Framework, or NEF) that models spike times, as well as post-synaptic behavior that reflects appropriate neurotransmitter dynamics (Eliasmith & Anderson, 2003). In addition, our model uses more sophisticated neural representations, reflecting multi-modal input through the same channel, and high-dimensional representations of context and stimuli. There are also critical structural differences between the models: most notably in the way context-dependent extinction is implemented. We also propose a novel, but biologically plausible hippocampal circuit that learns associations between conditioned stimuli and the context in which they occur.

The spike-timing dependent learning rule used in this model is also a novel contribution, combining previous work integrating supervised learning in the NEF (MacNeil & Eliasmith, 2011) with an unsupervised Hebbian term. This combined rule can be modulated by external error signals, such as an aversive US, and by the coactivation of inputs and outputs. The error-driven component and the Hebbian component can be combined at any ratio, allowing for flexible responses to bouts of high error or overall activity.

The model has three inputs. Multiple CS/NS representations, such as auditory, visual and tactile stimuli are simultaneously presented to the model as a single high-dimensional vector. Another high-dimensional input represents contextual information. The US input is a one dimensional input representing an electric shock. Depending on the magnitude of the inputs, temporal pairing, and previous learning, the model generates an output that initiates a fear response: freezing (a period of watchful immobility in rats).

Neural Engineering Framework

The Neural Engineering Framework developed by Eliasmith and Anderson (Eliasmith & Anderson, 2003) provides a method for representing and transforming information encoded in neurons. Using the NEF, complex algorithms can be encoded in neurons to generate models such as the one presented in this paper.

The NEF has three principles. The first is the representation principle, which details how a population of spiking neurons can represent high dimensional information (vectors). 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 aproximation to any function. In addition, these synaptic connection weights can be learned during a simulation to approximate any desired function. The third principle is the principle of dynamics, which we will not be considering in detail here.

The single neuron model used in the fear conditioning circuit presented here is the Leaky Integrate-and-Fire (LIF) neuron. While the NEF can support a wide variety of neural models, there are advantages to choosing LIF neurons: they capture a sufficient level of biological detail, and at the same time are computationally efficient to simulate. The properties of the LIF neurons used (for example their post-synaptic time constants and refractory periods) have been chosen to be consistent with what is known about the neurons in the brain regions being modeled.

The first NEF principle can be used to determine how a population of neurons can represent a high-dimensional vector. Each neuron in a population is taken to fire most strongly to one particular vector in that space: the neuron's preferred direction vector **e**. How strongly the neuron fires when representing any arbitrary vector is determined by the dot product of the vector being represented **x** and the neuron's preferred direction vector **e**. This value, multiplied by the neuron's gain α, plus the background current J_{bias}, determines the amount of current J that flows into the neuron (Eq. 1).

To represent an input vector **x** as neural activity, we use Eq. 1 to determine the somatic current, and use that to drive the single-neuron model to produce spiking. To perform the opposite operation - finding out what vector a population of spiking neurons is representing - we use Eq. 2. This equation solves for decoding weights **d** that when multiplied by the spiking activity of the neuron, filtered by a given post-synaptic current, gives back the optimal least-squares linear estimate of **x**.

We can perform a transformation **M** on a population of neurons representing **x** that will give us a second population that represents **Mx**. The synaptic connection weights between the two populations necessary to perform the transformation can be found using Eq. 3. The index i refers to a neuron in the population representing **x** and the index j refers to a neuron in the population representing **Mx**. Solving Eq. 3 for all neurons i in the first population and j in the second population gives the weight matrix needed to solve any linear operation defined by **M**. For non-linear operations, **d** values can be calculated as shown in Eq. 4.

$$J = \alpha \mathbf{e} \cdot \mathbf{x} + J_{bias} \tag{1}$$

$$\mathbf{d} = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \int a_i a_j \, dx \quad \Upsilon_j = \int a_j \mathbf{x} \, dx \tag{2}$$

$$\omega_{ij} = \alpha_j \mathbf{e}_j \mathbf{M} \mathbf{d}_i \tag{3}$$

$$\mathbf{d}^{f(x)} = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \int a_i a_j \, dx \quad \Upsilon_j = \int a_j f(\mathbf{x}) \, dx \tag{4}$$

These equations make it possible to translate a high-level algorithm that performs transformations on vectors into a spiking neuron model.

Learning

In many situations, including fear conditioning, the function to be computed by the connection between two populations cannot be determined a priori. In the case of fear conditioning, a particular stimulus cannot be said to evoke fear until it is associated with an aversive stimulus. The function also changes dynamically, as in the case of context-dependent extinction and renewal.

Learning has been previously explored in the NEF in (MacNeil & Eliasmith, 2011). They proposed a learning rule based on minimizing some provided error signal **E**.

$$\Delta\omega_{ij} = \kappa \alpha_j \mathbf{e}_j \mathbf{E} a_i$$

The error signal in the case of fear conditioning can be thought of as the aversive shock, which will be discussed in further sections.

It has been long thought that learning in the fear conditioning circuit is *Hebbian*; that is, the learning rule should be dependent on both input and output activity (a_i and a_j) (Sigurdsson, Doyère, Cain, & LeDoux, 2007). For this reason, we utilize a form of the well known BCM rule (Eq. 5), and combine it with the supervised rule from (MacNeil & Eliasmith, 2011) to yield Eq. 6.

$$\Delta\omega_{ij} = a_i a_j (a_j - \theta) \tag{5}$$

$$\Delta\omega_{ij} = \kappa \alpha_j a_i \left(S \mathbf{e}_j \mathbf{E} + (1-S) a_j (a_j - \theta)\right), \tag{6}$$

where θ is the average of the filtered output spiking activity (a_j) over some long time window ($\tau > 20s$), κ is the learning rate, and $0 \leq S \leq 1$ is how much the supervised component is weighted relative to the unsupervised component. The resulting rule utilizes spike-timing dependent plasticity (a well-established biological mechanism for learning (Markram, Gerstner, & Sjöström, 2011)), and includes both explicit error and self-organization terms. Such a rule has not been used in past models of fear conditioning.

The Model

Amygdala

The amygdala is a brain region that has long been associated with emotions - particularly fear - in mammals. Extensive connections with cortical areas that process sensory information provide the amygdala with the inputs needed to form complex associations between multiple sensory cues. There are three main anatomical regions of the amygdala implicated in this processing: the lateral amygdala (LA), the lateral basal area (BL), and the medial central nucleus (CEm), all of which are included in this model.

LA has been found to be a site of convergence of neutral and aversive stimuli (Sigurdsson et al., 2007). This convergence suggest that LA is where associations between an NS

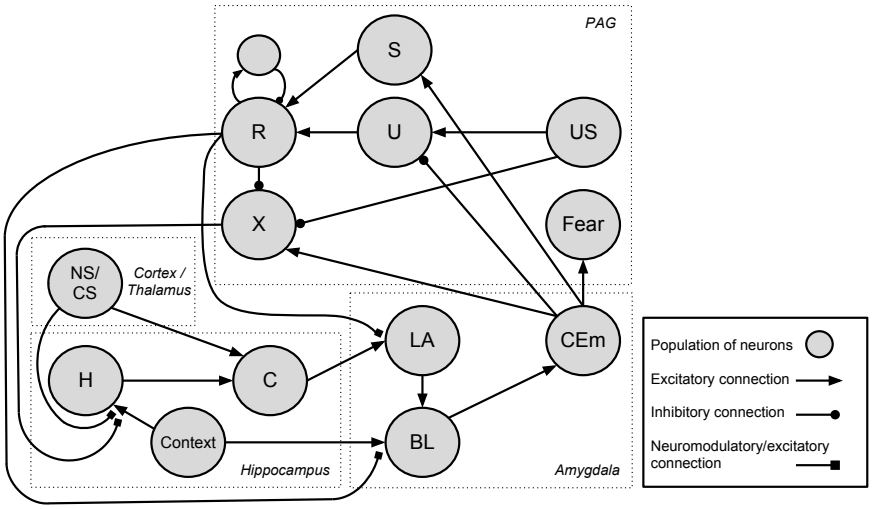

Figure 1: The proposed model circuitry. See text for definition of anatomical areas, and functional details.

and a US are formed. This learning is thought to rely on neuromodulatory signals as well as Hebbian processes in which post synaptic facilitation is necessary for learning (Krasne et al., 2011). The learning rule used here takes into consideration both of these requirements.

BL receives projections from the hippocampus, and is thought to be involved in processing contextual information. The learning between contextual information and a US is assumed in this model to be similar to the learning that occurs in LA. In addition to being a centre of convergence for contextual stimuli and USs, BL also serves as a gateway through which LA activity can reach CEm.

CEm receives input from BL. Activity of CEm travels to the periaqueiductal gray and drives freezing and feedback circuits that effect learning in LA and BL.

Periaqueductal Gray

Although there are many possible routes for US information to reach the amygdala, one possible route is through the periaqueductal gray (PAG) (Krasne et al., 2011). Because of its potential access to US information, in this model PAG is responsible for the reinforcement signals that facilitate learning in the amygdala. The control of these signals - also proposed to occur in PAG (Krasne et al., 2011) - enables more complex phenomena 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 fear and anxiety (Behbehani, 1995). In the context of the fear conditioning circuit, PAG is thought to be responsible for initiating fear responses such as freezing. The output of CEm triggers these processes through PAG.

Hippocampus

The hippocampus, a part of the cerebral cortex, has long been implicated in memory and spatial navigation. Contextual information - information about an animal's surroundings - plays an important role in the extinction of a CS that repeatedly occurs in the absence of a US. In this model, we propose a circuit based in the hippocampus that forms associations between contexts and sensory cues, and affects processing of CSs in the amygdala. There is prior evidence that the hippocampus may be involved in this kind of association (Moita, Rosis, Zhou, LeDoux, & Blair, 2003). In addition to its role in extinction, the hippocampus is also the source of contextual information that can come to elicit fear responses via learning in BL.

Implementation

As shown in Figure 1, the model consists of 14 neural populations that correspond to regions of the PAG, amygdala, thalamus, cortex, and hippocampus. Each population is made up of between 100-300 LIF neurons with parameters consistent with those found in the relevant areas of the mammalian brain. While in rats and other mammals each region of the brain being modeled here contains many more neurons, 100-300 neurons per population adequately represents the information needed for the model to run in its current configuration. Adding more neurons to the NS/CS and context populations would allow for higher dimensional representations and hence more sophisticated contexts and stimuli to be processed. However, for demonstration purposes, the inputs have been limited to three stimuli or contexts per population.

Experiments Modeled

To test the model, we chose to replicate four well-known fear conditioning experiments. The results of the simulations of these experiments are shown in Figures 2-5; they show the

decoded activity (the numeric values represented by a population of neurons) of the relevant regions of the model, with different colours representing different cues or contexts. To generate the figures in this paper, learning parameters (learning rate, strength of neuromodulators) have been adjusted so that conditioning requires only one pairing between an NS and a US. While freezing can be conditioned in a single trial (Bevins, McPhee, Rauhut, & Ayres, 1997), parameters of the learning rule can be adjusted so that conditioning occurs only after multiple pairings between an NS and a CS, as well (results not shown here). The simulations used for this paper were all run with the same structural parameters; the model was not adjusted between the simulated experiments. Single neuron model parameters were randomly generated within a biologically realistic range at the start of each experiment.

Conditioning

Simple fear conditioning experiments demonstrate classical Pavlovian conditioning. These experiments begin with a neutral, usually auditory or visual stimulus and a US such as an electric shock to an animal's foot. Initially, presentation of the NS has no behavioural effect on the animal, but after being temporally paired with the US, the NS becomes a CS that can elicit fear responses in the absence of the US.

Neurons in the amygdala have been shown to fire only for a short time at the onset of a US (Johansen, Tarpley, LeDoux, & Blair, 2010). 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 which 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 post-synaptic time constants - the R population will be inhibited, and the reinforcement signal will no longer be sent to the amygdala.

Recall the discussion of learning in the NEF. In the case of conditioning, the connection weights being adjusted are those between the CS/NS population and LA. The neuromodulatory error signal is the value represented by the R population of neurons driven by the US (neuromodulators serving as error signals in learning is thought to occur in other brain regions as well (Schultz & Dickinson, 2000)). The R population is also connected directly to LA to facilitate the unsupervised Hebbian portion of the learning rule. Initially, the synaptic weights between the neurons representing the NS and LA are such that any activity of an NS will induce no, or very little, activity in LA. However, at the onset of a US, the synaptic weights between the neurons representing the NS and LA will change according to Eq. 6 such that the value represented by LA will be increased (if a positive error signal is present for a sufficient duration, the value represented at LA will saturate at its maximum). It should be noted that the error signal used here only takes on values between 0 and 1. Because of this, it can only be used to drive activity in LA higher. Extinction is handled by a different mechanism in the brain, which we discuss later.

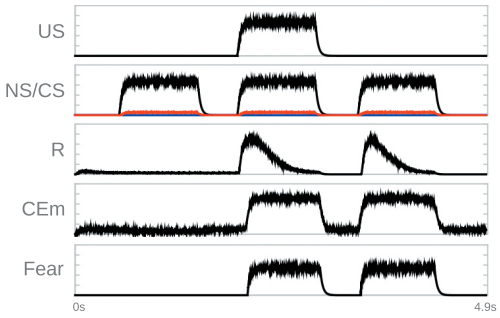

Figure 2: Simple conditioning. Initially, a stimulus is presented that evokes no response. The stimulus is then conditioned via pairing with a US, resulting in a fear response. Finally, the stimulus is shown to elicit a fear response in the absence of the US. All plots in this and following figures are representations decoded from spike trains in the relevant populations using the NEF.

This learning is specific to the stimulus present during training; the synaptic weights are adjusted such that only the CS that was present during the reinforcing signal will be able to elicit the learned response in LA in the future. Activity of LA travels to BL and then to CEm which then has the potential to activate the population of neurons that initiate freezing. Figure 2 shows the results of a simulation of conditioning. Note that, although not shown, the model allows for conditioning of a context in the same way an NS can be conditioned.

Second-Order Conditioning

Second-Order conditioning experiments begin with a CS (previously conditioned with a US) and an NS. The NS is temporally paired with the CS; after this training, the former NS will become a CS able to elicit a fear response.

The learning in this case is similar to that in the conditioning discussed above. However, in this case, the US is not present to activate the R neurons and provide the learning signals needed for conditioning. Instead the reinforcing signal arrives at the R population from CEm through the S population. The activity of a strong CS is enough to activate LA, CEm, the S population, and the R population, which facilitates strengthening of the synapses between neurons representing an NS and LA. Figure 3 shows the values represented by neural populations during a simulated second-order conditioning experiment.

Blocking

Blocking occurs when a strong CS, which has been trained to elicit a fear response, is present while an NS is paired with a US. Experiments show that after training under these conditions, the NS does not come to elicit a fear response on its own. The CS is already a strong predictor of the US, so when paired with an NS, the NS is overshadowed, or blocked, by the CS.

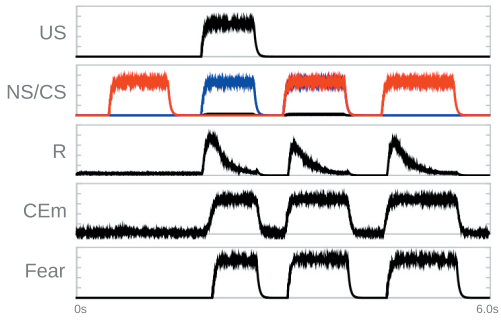

Figure 3: Second-order conditioning. Initially, the red stimulus is presented, evoking no response. Next, the blue stimulus is conditioned via pairing with a US, resulting in a fear response. The red stimulus is then conditioned via pairing with the blue stimulus, resulting in a fear response. Finally, the red stimulus is shown to elicit a fear response in the absence of the US or another CS.

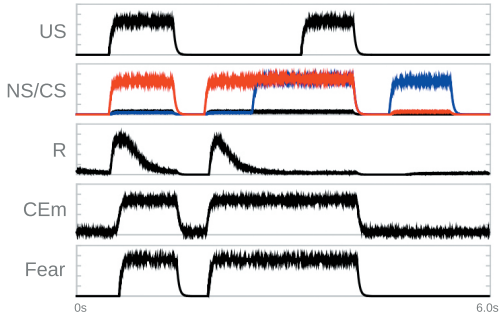

Figure 4: Blocking. Initially, the red stimulus is conditioned via pairing with a US, resulting in a fear response. Next, the red stimulus is shown to evoke a fear response. The blue stimulus is then presented. Subsequently, a US is presented. The presence of the red CS blocks conditioning of the blue NS (note the lack of activity in the R population at the second presentation of the US). Finally, the blue NS is presented and fails to evoke a fear response.

In order to account for this phenomenon in the model, an inhibitory population blocks the reinforcing effects of a US when a strong CS is already present. CEm inhibits the U population through which the signal from the US reaches the reinforcing population R. In the presence of sufficient CEm activity supplied by a strong CS, the US will be unable to activate the R population of neurons. The results of a simulation shown in Figure 4 demonstrate this effect.

Extinction and Renewal

After training, a CS can come to elicit fear responses in the absence of a US. However, if the CS is repeatedly presented in the absence of the US, it will eventually lose its ability to elicit a fear response. It is thought that this effect is not simply an unlearning process on the synapses between the neurons representing the CS and the neurons of LA (Rescorla, 1993). If a CS is extinguished in one particular context, it is capable of eliciting a fear response in a different context. This finding suggests that inhibition of a CS after extinction is itself a learned state that is context dependent.

We can think of this as each context being able to learn the state of the CS in that context (Bouton, 2004). That is, each context tells us whether any CS is inhibited or uninhibited in that context. This information is passed along to a population of neurons that calculates the effective strength of the CS in that context and relays that information to LA.

This extinction association between a CS and a context should only be learned in the case that a CS is exciting LA in the absence of a US. The reinforcing signal for this circuit is generated by the X population of neurons. As shown in Figure 1, if a CS is responsible for exciting neurons in CEm, but is not accompanied by a US, the X neurons will be excited - notice that the US inhibits the X population. The error signal in this case is not represented by the X population; rather, the X population enables learning on a more complex, multi-dimensional error signal.

The transformation being learned is between the populations of neurons representing context, and the H population. The error signal is supplied by the NS/CS population. In the presence of the X neuromodulator, the synaptic weights between the context population and the H population will be adjusted such that the current context evokes the state of the NS/CS population in the H population. The value represented in the H population is passed to the C population where its representation is subtracted from the representation of the NS/CS population. This results in suppression of the CS, or CSs, associated with that context. The learning of the synaptic weights here is also governed by the rule in Eq. 6.

To demonstrate this, an ABC extinction and renewal experiment is simulated using the model. In ABC renewal, conditioning is performed in context A, extinction is performed in context B, and the CS is able to elicit a fear response when tested in a new context, C. Note that contextual information is being conditioned along with the CS; however, it is being extinguished as well (see Discussion). Figure 5 shows the values being represented by the populations of the model during this processes.

Discussion

We have shown that a biologically plausible model is capable of replicating the results of several well-known fear conditioning experiments. The model demonstrates the connection between the neural circuitry of fear conditioning and the behavioural responses that rely on it. We can see, through dynamic simulations, how populations of spiking neurons give rise to emotional processing through the association of intrinsically aversive stimuli and high-dimensional information in the cortex and hippocampus.

The hippocampal circuit we have introduced demonstrates how emotionally significant information can be linked to contexts. It allows us to explore the neural mechanisms govern-

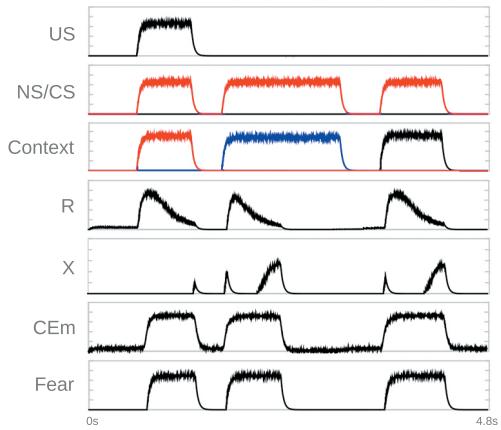

Figure 5: Extinction and renewal. Initially, the stimulus is conditioned to elicit a fear response via pairing with the US in the red context. Next, when tested in the blue context, the stimulus elicits a fear response temporarily. However, in the absence of a US, activity in the X population is uninhibited, initiating extinction (accelerated for demonstration) of the stimulus. Notice that fear responding ceases during the second presentation of the stimulus. Finally, despite extinction in the blue context, the stimulus evokes a fear response in the black context.

ing conditioning and extinction of fear memories in relation to context: something that could be critical in, for example, the study of post traumatic stress disorder (Shin, Rauch, & Pitman, 2006). Additionally, this circuit has significance as a general model of how salient information can be used as an error signal in order to learn associations between high-dimensional populations. This type of circuit can naturally be integrated with what is currently the world's largest functional brain model (Eliasmith et al., 2012), and which accounts for a variety of perceptual, motor, and cognitive phenomena. The proposed circuit is critical for extending such large-scale models to include more emotional and reinforcement based processing.

The learning rule we have presented here is shown to successfully characterize the synaptic plasticity required for the adaptability of the fear conditioning circuit. Critically, the rule explains both learning mechanisms implicated in fear conditioning: error modulated, and Hebbian learning. The rule demonstrates how neuromodulators modify the fear conditioning circuit: a processes that is critical to understanding adaptability in emotional processing. It also explains how Hebbian processes contribute to the development of emotionally significant associations, and more generally, the formation of associations between salient temporally paired stimuli.

While the model offers significant contributions to the biological understanding of fear conditioning, some properties of the model need to be further defined. For example, how to handle extinction of a conditioned contextual stimulus is still unclear. The model currently handles this by performing unlearning on the synapses between the context population and BL when the X population is active, but other possibilities may exist. However, a major contribution of the model is that it provides a biologically plausible platform through which to explore this and other more complex details of fear conditioning.

References

Behbehani, M. M. (1995). Functional characteristics of the midbrain periaqueductal gray. *Progress in Neurobiology*, *46*, 575– 60.

Bevins, R. A., McPhee, J. E., Rauhut, A. S., & Ayres, J. J. (1997). Converging evidence for one-trial context fear conditioning with an immediate shock: importance of shock potency. *Journal of experimental psychology. Animal behavior processes*, *23*(3), 312–24.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learn. Mem.*, *11*, 485–494.

Eliasmith, C., & Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., et al. (2012, November). A Large-Scale Model of the Functioning Brain. *Science*, *338*(6111), 1202–1205.

Grossberg, S., & Levine, D. S. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: blocking, interstimulus interval, and secondary reinforcement. *Applied optics*, *26*(23), 5015–30.

Johansen, J. P., Tarpley, J. W., LeDoux, J. E., & Blair, H. T. (2010). Neural substrates for expectation-modulated fear learning in the amygdala and periaqueductal gray. *Nature Neuroscience*, *13*(8), 979–86.

Krasne, F. B., Fanselow, M. S., & Zelikowsky, M. (2011). Design of a neurally plausible model of fear learning. *Frontiers in Behavioral Neuroscience*, *5*, 41.

MacNeil, D., & Eliasmith, C. (2011). Fine-tuning and the stability of recurrent neural networks. *PloS One*, *6*(9), 22885.

Markram, H., Gerstner, W., & Sjöström, P. J. (2011). A history of spike-timing-dependent plasticity. *Frontiers in synaptic neuroscience*, *3*, 4.

Moita, M. A., Rosis, S., Zhou, Y., LeDoux, J. E., & Blair, H. T. (2003). Hippocampal place cells acquire location-specific responses to the conditioned stimulus during auditory fear conditioning. *Neuron*, *37*(3), 485–97.

Rescorla, R. A. (1993). Inhibitory associations between S and R in extinction. *Animal Learn. & Behav.*, *21*, 327-336.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci.*, *23*, 473–500.

Shin, L. M., Rauch, S. L., & Pitman, R. K. (2006). Amygdala, medial prefrontal cortex, and hippocampal function in PTSD. *Annals of the New York Acad. of Sci.*, *1071*, 67–79.

Sigurdsson, T., Doyère, V., Cain, C. K., & LeDoux, J. E. (2007). Long-term potentiation in the amygdala: a cellular mechanism of fear learning and memory. *Neuropharmacology*, *52*(1), 215–27.