Lateral Inhibition Explains Savings in Conditioning and Extinction

Ashish Gupta & David C. Noelle
({ASHISH.GUPTA, DAVID.NOELLE}@VANDERBILT.EDU)
Department of Electrical Engineering and Computer Science
Vanderbilt University; Box 1679 Station B
Nashville, TN 37235, USA

Abstract

Animal conditioning experiments have shown that acquired behaviors that are subsequently extinguished are reacquired at a faster rate than during their initial acquisition. Residual synaptic plasticity in the relevant neural circuits has been the prevalent explanation for this form of savings. According to this theory, extinction training does not completely revert synaptic changes induced during initial acquisition, resulting in faster reacquisition. This account cannot explain more recent findings, however, that show that subsequent extinctions are also faster than the first. Rescorla has proposed an alternative to the residual synaptic plasticity account, in which acquisition and extinction involve the formation of two separate kinds of associations. We have explored this dual-pathway account using a neurocomputational model of conditioning. In our model, associations related to acquisition and extinction spontaneously become segregated as a result of the interaction between general neural learning processes and the presence of lateral inhibition between neurons. This model exhibits appropriate savings in both acquisition and extinction, and it captures the experimental results that prompted Rescorla to hypothesize separate acquisition and extinction pathways.

Introduction

The relationship between the *learning* of an association and the *unlearning* of that same association is commonly thought to involve a singular representation of the strength of association, with that strength rising during learning and falling during unlearning. In animal conditioning, this view suggests that the extinction of a behavior involves reversing the synaptic modifications made during the initial acquisition of that behavior. During acquisition training, the association between the conditioned stimulus (CS) and the unconditioned stimulus (US) is encoded by changing the strength of the synaptic interconnections between certain neurons in the brain. During extinction training, the changes made to these connections are reversed, causing the animal to stop producing the conditioned response (CR) (Kehoe, 1988; Klopf, 1988; Medina et al., 2001). While this theory is simple and elegant, it is not consistent with a growing body of behavioral findings.

Evidence from numerous behavioral studies points to the possibility that extinction isn't a mere reversal of the associations formed during acquisition. Phenomena like *savings*, *spontaneous recovery*, *renewal*, and *reinstatement* (Rescorla, 2001; Bouton, 2004) suggest that extinction training involves the superimposition of some separate decremental process that works to inhibit previously learned responses, leaving most of the originally acquired CS-US association intact. These phenomena involve a restoration of responding to a CS that was first associated with the US through acquisition training and then disassociated through extinction training. The phenomenon of *savings* involves the relatively small amount of reacquisition training needed to restore responding after extinction in comparison to the amount of initial acquisition training. In *spontaneous recovery*, responding to the CS is restored simply by the passage of time, after extinction training. *Renewal* is said to occur when a shift in environmental context away from that in which extinction training took place results in renewed responding. In *reinstatement*, the response is restored through the presentation of US, alone. All of these phenomena suggest that learned associations are not completely lost during extinction training.

Recognition of this retained association knowledge, even after responding has been extinguished, has led to theories involving *residual synaptic plasticity* and *subthreshold responding*. These theories hold that extinction training does not completely reverse synaptic changes made during initial acquisition, but only reverses these changes enough to effectively inhibit responding. When presented with the CS after extinction, the neural system involved in producing a response continues to become somewhat active, but not sufficiently active to produce an actual response. Thus, only small changes in association strength are needed to return this system to a state in which responding to the CS is robust.

Theories based on residual synaptic plasticity cannot account for some important additional observations, however. In particular, there is evidence that, just as extinction does not remove associations built up during previous acquisition training, subsequent reacquisition training does not remove the inhibitory force built up during previous extinction training. For example, animals continue to show spontaneous recovery — a phenomenon that only arises after extinction training — even if they experience a subsequent period of reacquisition that removes the behavioral impact of the previous extinction process (Rescorla, 2001). Also, just as reacquisition after extinction is faster than initial acquisition, subsequent extinctions are also faster than the first extinction (Reynolds, 1975; Rescorla, 2001). Ob-

servations of this kind have led Rescorla to hypothesize that the effects of acquisition and those of extinction are maintained within dual pathways, with the competition between these separated pathways determining the magnitude of response to the CS (Rescorla, 2001). His experiments also have led him to conjecture that these pathways interact, with the strengthening of one pathway making the other more sensitive to further training (Rescorla, 2002; Rescorla, 2003). (Some of these experiments are described in the next section.)

In this paper, we show that fundamental principles of neural computation, embodied in the Leabra modeling framework (O'Reilly and Munakata, 2000), spontaneously capture these phenomena of conditioning and extinction without requiring the incorporation of separate modules for acquisition and extinction. In particular, we demonstrate how synaptic plasticity, bidirectional excitation between cortical regions, and lateral inhibition within cortical regions interact so as to spontaneously segregate neural pathways associated with acquisition from those associated with extinction, allowing the effects of previous acquisition and extinction sessions to be retained. Of particular importance are processes of lateral inhibition, which introduce competition between neurons involved in the encoding of stimuli. Along with the mechanisms of synaptic learning, this competition separates the neurons that associate the stimulus with responding, called *acquisition neurons*, from those that associate the stimulus with non-responding, called *extinction neurons*. During extinction training, for example, synaptic strengths change so as to encourage the activation of extinction neurons and discourage the activation of acquisition neurons. Importantly, the weakening of excitatory synapses on acquisition neurons only continues until these neurons begin to lose their competition with extinction neurons, brought about by lateral inhibition, at which point the activation levels of the acquisition neurons drop dramatically, causing the synaptic modification process to effectively cease. Thus, much of the associational knowledge embedded in the synapses of the acquisition neurons is retained even after extinction. Similarly, many of the changes in extinction neuron synapses wrought during extinction training are retained after reacquisition training. Through this retention of synaptic strengths, our model exhibits a speeding of both subsequent reacquisitions and subsequent re-extinctions, demonstrating the savings seen in animals. It also captures the patterns of performance observed by Rescorla, without requiring an explicit mechanism for modulating the speed of learning within one pathway or the other.

Background

Behavioral Results

Rescorla identified two different mechanisms that might be responsible for faster reacquisition of responding after extinction. First, it is possible that the association with the CS is not completely removed by extinction training — that residual synaptic plasticity retains some associational connection. Second, it might be the case that extinction training triggers faster subsequent learning — that a CS undergoing retraining is particularly quick to acquire new associative connections with the US due to its prior history. To investigate these two alternatives, Rescorla conducted the following experiments.

In one experiment, two stimuli, A and C, were initially trained and then extinguished. Two other stimuli, B and D, were presented without reinforcement. Once A and C were extinguished, A and B then each received the same number of conditioning trials, encouraging responding to these stimuli. At the end of this training sequence, A elicited stronger responding than B. This is a demonstration of savings, since A was previously acquired and extinguished and B was not. This observation does not distinguish between Rescorla's two alternatives, however. The A stimulus could have begun reacquisition training with some residual synaptic plasticity or the reaquisition process could have operated at a faster rate for A. In order to separate these hypotheses, Rescorla tested responding to the compound stimuli AD and BC. Any residual synaptic plasticity in A should also be present in C, so responding to these two compounds should be roughly equivalent if both A and B grow equally in associational strength during reacquisition training. If, however, an association to A is learned faster because of its previous extinction, then greater responding should be seen to the AD compound. Surprisingly, neither of these outcomes were observed. Responding to BC was stronger than responding to AD. Rescorla concluded the A's dominance over B was the result of residual synaptic plasticity, and he explained the dominance of the BC compound in terms of blocking-like effect. If associative change is governed by an error-correction learning mechanism, and if stimulus A begins reacquisition training with a "head start" over stimulus B, there will be less error when stimulus A is presented, so the associational strength for A will grow more slowly than that for B. Since A's residual synaptic plasticity is shared by C, and since B's associational strength grows faster than that of A during reaquisition training, the BC compound dominates over AD (Rescorla, 2002).

This explanation gave rise to a question: Would A or B show greater associative change if the error signal during reacquisition training was equilibrated between them? In another experiment, Rescorla addressed this question by presenting the AB compound stimulus, rather than A and B separately, during reacquisition training (Rescorla, 2003). When this was done, greater responding was generated by the AD compound than by the BC compound. Hence, Rescorla concluded that, in addition to leaving residual associative strength, extinction also causes the stimulus to gain new associative strength at a faster rate when it is, once again, reinforced. Through a similar set of experiments, he concluded that a stimulus that was previously extinguished and reacquired is more sensitive to subsequent non-reinforcement.

Leabra Modeling Framework

In this paper, we show that these results arise naturally from the mechanisms of neural computation em-

bodied in the Leabra modeling framework. The Leabra cognitive modeling framework (O'Reilly and Munakata, 2000) offers a collection of integrated formalisms that are grounded in known properties of cortical circuits but are sufficiently abstract to support the simulation of behavior. The framework has been used to model a broad range of cognitive processes, including aspects of perception, attention, language, learning, and memory. Leabra includes dendritic integration using a point-neuron approximation, a firing rate model of neural coding, bidirectional excitation between cortical regions, fast feedforward and feedback inhibition, and synaptic plasticity that incorporates both error-driven and Hebbian learning. Of particular relevance to our model is Leabra's lateral inhibition formalism.

The effects of inhibitory interneurons tend to be strong and fast in cortex. This allows inhibition to act in a regulatory role, mediating the positive feedback of bidirectional excitatory connections between brain regions. Simulation studies have shown that a combination of fast feedforward and feedback inhibition can produce a kind of "set-point dynamics", where the mean firing rate of cells in a given region remains relatively constant in the face of moderate changes to the mean strength of inputs. As inputs become stronger, they drive inhibitory interneurons as well as excitatory pyramidal cells, producing a dynamic balance between excitation and inhibition. Leabra implements this dynamic using a *k-Winners-Take-All (kWTA)* inhibition function that quickly modulates the amount of pooled inhibition presented to a layer of simulated cortical neural units, based on the layer's level of input activity. This results in a roughly constant number of units surpassing their firing threshold. The amount of lateral inhibition within a layer can be parameterized in a number of ways, with the most common being either the absolute number or the percentage of the units in the layer that are expected, on average, to surpass threshold. A layer of neural units with a small value of this k parameter (e.g., 10-25% of the number of units in a layer) will produce sparse representations, with few units being active at once.

In our model, acquisition-related and extinction-related learning occurs in two distinct sets of neurons that compete with each other via this lateral inhibition mechanism. Indeed, it is lateral inhibition, in conjunction with Leabra's synaptic learning mechanism, that gives rise to a segregation between acquisition neurons and extinction neurons.

The Model

The learning performance of a simple multi-layer Leabra network, as shown in Figure 1, was examined. For simplicity, each stimulus (CS) was encoded as a single input unit. The stimulus was recoded over the firing rates of 40 units grouped into a hidden layer. This hidden layer incorporated strong lateral inhibition, using a kWTA parameter of $k = 3$, encouraging only 3 of the 40 units to be active at any one time. The hidden layer had a bidirectional excitatory projection to the output layer. The output layer contained 7 units, with $k = 5$. The

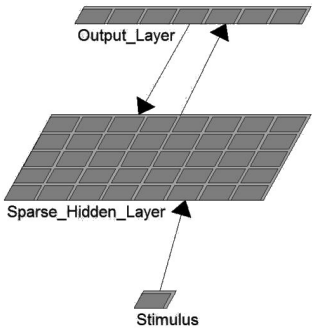

Figure 1: The Leabra network. Each gray box corresponds to a neural processing unit. Each arrow represents complete interconnectivity between the units in two layers.

first 5 units were interpreted as encouraging a positive response in the face of the stimulus, the aggregate activation over these units determining the strength of the response. The remaining 2 units in the output layer coded for a null response, and they offered a means to suppress activity in the first 5 units via lateral inhibition.

When simulating more than one CS, each was encoded over a separate input unit and a separate layer of 40 hidden units. All of the hidden layers participated in bidirectional excitatory connections with a single shared output layer, identical to the one previously described. Thus, different stimuli could not be represented using shared neural resources. This amounts to an assumption that the stimuli are all highly dissimilar, with each activating different neurons in the brain. This simplifying assumption is not a critical feature of this model.

Leabra's default parameters were used in these simulations, with only a few exceptions. To accommodate the relatively small size of this network, the range of initial random synaptic weights was reduced ($[0.0, 0.1]$ rather than the default range of $[0.25, 0.75]$) and learning rate for synaptic modification was set to a smaller value (0.005, half of the default of 0.01). Also, individual neuron bias weights were removed. Modifications of these kinds are common in smaller Leabra networks.

During acquisition training, the network was expected to activate the first 5 units in the output layer. During extinction training, it was expected to activate the last 2 units. Each training session was terminated when the sum squared error (SSE) between the network's output and these expected output patterns fell below a criterion value of 1. All simulation experiments were repeated 20 times, and mean results across these runs are reported.

Experiments

Experiment 1

Our first simulation experiment was designed to uncover the degree to which our model exhibits savings. Recall that animals are faster to reacquire an extinguished behavior, as compared to initial acquisition, and they are faster to extinguish a reacquired behavior, as com-

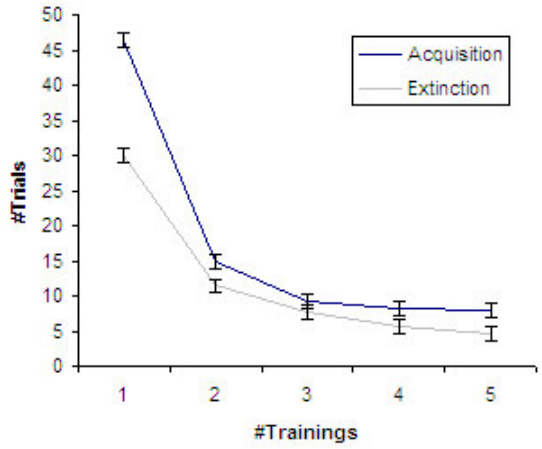

Figure 2: The number of training trials required to reach criterion (Y axis) decreases as number of prior acquisition and extinction training sessions (X axis) increases. Error bars report the standard error of the mean.

pared to initial extinction. A randomly initialized network was trained to respond upon the presentation of the CS (A+). Once this training reached criterion, the network was trained to not-respond upon the presentation of the CS (A-). This pattern was repeated 5 times. Figure 2 shows the number of trials required for successive acquisition and extinction trainings. Note that the required time quickly decreases. The model predicts that the required number of trials will asymptote to a small value after just a few acquisition-extinction iterations.

Why does this model exhibit savings? The network starts with small initial synaptic weights. Hence, a large change in weights is required for success during the first acquisition training session, and these weights are slowly built up on those acquisition neurons in the hidden layer that happen to win the competition imposed by lateral inhibition. During the first extinction training session, feedback from the output layer to the hidden layer encourages a different set of hidden units to become active, and these units take on the role of extinction neurons. The weights to the acquisition neurons start decreasing and the weights to the extinction neurons start increasing. As soon as the extinction neurons win the inhibitory competition, the acquisition neurons tend to fall below their firing threshold. At this stage, the weights to the acquisition neurons stop decreasing, as these neurons are no longer contributing to erroneous outputs. The weights to the extinction neurons continue to increase until the training criterion is met. During the second acquisition training, the weights to the acquisition neurons begin increasing again and the weights to the extinction neurons start to decrease. Once again, as soon as the extinction neurons lose the inhibitory competition, their activity falls essentially to zero, and their weights do not decrease further. Over successive acquisition and extinction trainings, the amount of change in weights keeps decreasing. Thus, acquisition and extinction associations are eventually maintained side by side in the network, allowing for the rapid switching between them based on recent conditioning feedback.

Table 1: The three training sessions, and single testing session, used in Experiment 2. Letters correspond to different stimuli. A plus indicates acquisition training, and a minus indicates extinction training.

Acquisition	Extinction	Reacquisition	Test
A+	A-		
B-	B-	A+	AD
C+	C-	B+	BC
D-	D-		

Table 2: The three training sessions, and single testing session, used in Experiment 3. Letters correspond to different stimuli. A plus indicates acquisition training, and a minus indicates extinction training. Note that "AB+" indicates that both A and B were presented together, as a compound, and this compound was reinforced.

Acquisition	Extinction	Reacquisition	Test
A+	A-		
B-	B-	AB+	AD
C+	C-		BC
D-	D-		

Experiment 2

The design of our second experiment is shown in Table 1. As previously discussed, Rescorla designed this experiment to assess whether the rapidity of reacquisition was a result of residual synaptic plasticity or of an increase in acquisition speed after extinction (Rescorla, 2002). A randomly initialized network was first trained on two CSs (A+ and C+) while two other stimuli were non-reinforced (B- and D-). Once the network reached criterion, it was then trained to extinguish A and C (A- and C-). During this session, B and D were presented in a non-reinforced manner as well (B- and D-). This was followed by training on A and B (A+ and B+) for 20 trials.[1] At the end of these training sessions, the response to A was much stronger than the response to B ($t(38) = 26.1$, $p < 0.001$), as shown in Figure 3. This is in accordance with Rescorla's observations. Finally, the network was tested on the compounds: AD and BC. As observed behaviorally, the network showed greater responding for BC than for AD ($t(38) = 2.9$, $p < 0.006$). See Figure 4.

In addition to capturing these general empirical re-

[1]This number of trials was chosen to make these results comparable to those from Experiment 3. In Experiment 3, it was found that 20 trials were needed, on average, to train the AB compound to criterion.

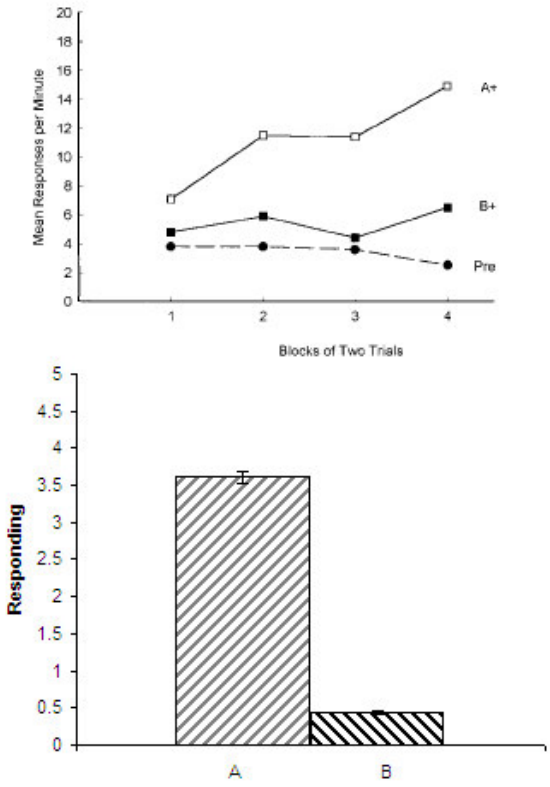

Figure 3: Experiment 2. Top: Results from Rescorla's experiment — mean responding for A, B, and with no stimulus present (Pre) during the reacquisition phase. Bottom: Simulation result — response magnitude for A and B at the end of the reacquisition phase, with error bars showing standard errors of the mean.

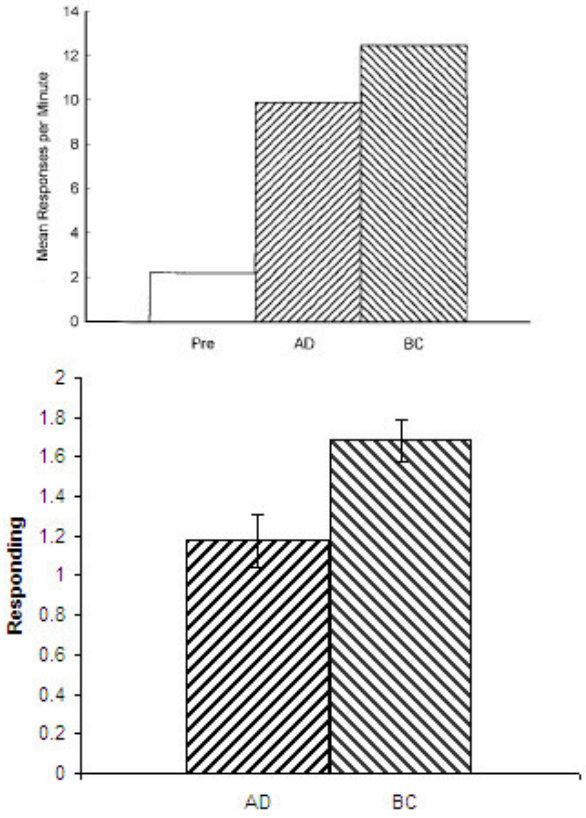

Figure 4: Experiment 2. Top: Results from Rescorla's experiment — mean responding for AD, BC, and with no stimulus present (Pre). Bottom: Simulation result — response magnitude for AD and BC compounds, with error bars showing standard errors of the mean. Note that BC produced a stronger response than AD.

sults, our model also matches more subtle nuances in Rescorla's data. First, the compound BC produced a much stronger response in animals than either B or C alone. Second, the compound AD was found to produce significantly weaker responding than stimulus A, alone. Our model captures both of these results. How can these results be explained? First, weights to B's acquisition neurons were strong, due to acquisition training, but it maintained only weak extinction neuron weights, since B was never extinguished. After extinction, C was left with strong weights for both acquisition neurons and extinction neurons. Interacting through bidirectional excitation with the output layer, the acquisition neurons for both B and C were able to mutually support each other, producing a strong overall response. In comparison, the weakly extinction-biased weights of D, when combined with the strong but balanced weights of A, were enough to start to tip the inhibitory competition in the direction of a null response when A and D were combined.

Experiment 3

Table 2 shows the design of our third experiment. As previously discussed, Rescorla designed this experiment to identify speeded learning after extinction by equalizing for the amount of error experienced by both A and B during reacquisition training. A randomly initialized network was trained as before, with the only difference being the use of a compound stimulus (AB+) during reacquisition training. Reaching criterion during reacquisition required 20 trials, on average. As observed in animals, the network produced stronger responding for AD than for BC ($t(38) = 5.0$, $p < 0.001$). See Figure 5.

Our model contains no mechanism for increasing the rate of learning for A after its extinction. So, how did our model capture this pattern of performance? This was not due to a speeding of learning with regard to the A stimulus, but due to a blocking of learning with regard to the B stimulus. When we measured the overall increase in the weights from the input layer to the acquisition neurons for the B stimulus, we found that this increase was only 0.083 for this experiment, while an increase of 0.407 was found for Experiment 2. At the beginning of the reacquisition phase, the AB compound produced very little responding. Hence, the error signal driving synaptic changes was strong. However, within

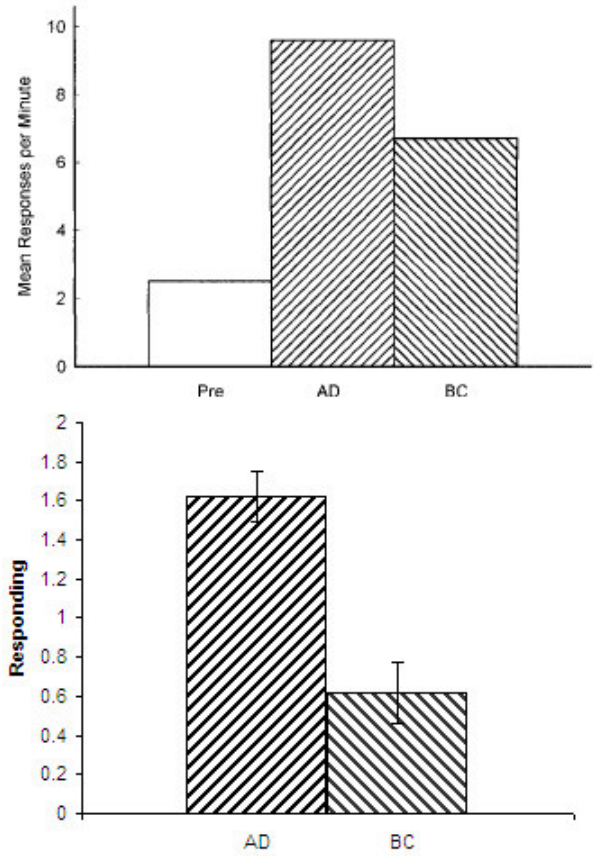

Figure 5: Experiment 3. Top: Results from Rescorla's experiment — mean responding for AD, BC, and with no stimulus present (Pre). Bottom: Simulation result — response magnitude for AD and BC compounds, with error bars showing standard errors of the mean. Note that AD produced a stronger response than BC.

only a few trials the network started producing a strong response, due to A's prior history of acquisition. At this point the magnitude of error signal dropped substantially, and B's weights stopped growing. In contrast, Experiment 2 included reaquisition trials involving B, alone, which produced a small response for the 20 trials in this phase, driving strong weight changes throughout.

Discussion

We have proposed a neurocomputational model for savings in conditioning and extinction. This model rejects the notion that extinction involves only a reversal in previously acquired synaptic associations, positing, instead, the existence of a separate pathway for extinction effects. This separate pathway is not an isolated architectural component of the model, however. Segregated acquisition and extinction pathways arise spontaneously through the interaction of foundational neural processes, including error-driven synaptic plasticity, bidirectional excitation, and strong lateral inhibition. We have shown that our model captures the relevant patterns of performance exhibited by animals.

In this paper, we have conceptualized the output of the network as encoding the propensity to produce the conditioned response. Alternatively, the network output could be interpreted as encoding an expectation of reward. In this case, separate neural circuitry would be responsible for converting this reward expectation into a response. Note that both of these interpretations are consistent with the results that we have presented here.

It is important to note that the learning mechanisms of our model are very similar to those used in other simple associative models of conditioning. Thus, this model can easily capture common conditioning results like blocking, summation, and overshadowing. Our model also seems well suited to explore other extinction-related behavioral results, such as conditioned inhibition, counter conditioning, latent inhibition, reinstatement, renewal, and spontaneous recovery. We are in the process of modeling all of these phenomena.

This work is part of a broader effort to explain the full range of conditioning phenomena in terms of the fundamental properties of neural circuits. Notice that our model depended not at all on the specific properties of particular brain areas. If successful, this effort will help us understand why vastly different brains produce similar patterns of learning.

Acknowledgments

The authors extend their thanks to the members of the Vanderbilt Computational Cognitive Neuroscience Laboratory (CCNL) and to three anonymous reviewers.

References

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory*, 11:485–494.

Kehoe, E. J. (1988). A layered network model of associative learning: Learning to learn and configuration. *Psychological Review*, 95(4):411–433.

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology*, 16(2):85–125.

Medina, J. F., Garcia, K. S., and Mauk, M. D. (2001). A mechanism for savings in the cerebellum. *The Journal of Neuroscience*, 21(11):4081–4089.

O'Reilly, R. C. and Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press.

Rescorla, R. A. (2001). Retraining of extinguished pavlovian stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, 27(2):115–124.

Rescorla, R. A. (2002). Savings tests: Separating differences in rate of learning from differences in initial levels. *Journal of Experimental Psychology: Animal Behavior Processes*, 28(4):369–377.

Rescorla, R. A. (2003). More rapid associative change with retraining than with initial training. *Journal of Experimental Psychology: Animal Behavior Processes*, 29(4):251–260.

Reynolds, G. S. (1975). *A Primer of Operant Conditioning*. Scott, Foresman and Company.