Learn Behav (2014) 42:22–38
DOI 10.3758/s13420-013-0112-z

An elemental model of retrospective revaluation without within-compound associations

Patrick C. Connor · Vincent M. Lolordo · Thomas P. Trappenberg

Published online: 29 June 2013

Abstract When retrospective revaluation phenomena (e.g., unovershadowing: AB+, then A−, then test B) were discovered, simple elemental models were at a disadvantage because they could not explain such phenomena. Extensions of these models and novel models appealed to within-compound associations to accommodate these new data. Here, we present an elemental, neural network model of conditioning that explains retrospective revaluation apart from within-compound associations. In the model, previously paired stimuli (say, A and B, after AB+) come to activate similar ensembles of neurons, so that revaluation of one stimulus (A−) has the opposite effect on the other stimulus (B) through changes (decreases) in the strength of the inhibitory connections between neurons activated by B. The ventral striatum is discussed as a possible home for the structure and function of the present model.

Keywords Associative learning · Classical conditioning · Cue competition · Generalization · Retrieval

Introduction

Early reports of retrospective revaluation phenomena identified two different effects—namely, recovery from overshadowing (Kaufman & Bolles, 1981; Matzel, Schachtman, & Miller, 1985) and backward blocking (Shanks, 1985). In the first phase of recovery from overshadowing, a compound stimulus is conditioned (AB+). This causes overshadowing, where the stimuli of the compound each receive less associative strength than they would have had they been conditioned alone. In the second phase, only one stimulus from the original compound is presented, and it is not reinforced (A−), extinguishing that stimulus's associative strength. Recovery from overshadowing is the finding that responding to B, the absent stimulus, is increased relative to a control group (e.g., that receiving C− in the second phase). In backward blocking, the first phase also involves conditioning a compound stimulus (AB+). In the second phase, however, one element is presented and reinforced (A+). Backward blocking is the finding that responding to stimulus B decreases below the level in a relevant control group. The phenomenon gets its name because it is the reverse of the standard blocking procedure (phase 1, A+; phase 2, AB+). Research in the years that followed these reports turned up several other forms of retrospective revaluation—namely, backward conditioned inhibition (Chapman, 1991; Urcelay, Perelmuter, & Miller, 2008), recovery from forward blocking (Blaisdell, Gunther, & Miller, 1999), recovery from conditioned inhibition (Lysle & Fowler, 1985), and others. Retrospective revaluation phenomena have not always been found when tested for, however (e.g., Dopson, Pearce, & Haselgrove, 2009; Shevill, & Hall, 2004).

The associative models in the early days of retrospective revaluation findings were unable to account for these phenomena. For example, the Rescorla–Wagner model (Rescorla & Wagner, 1972) predicts that no change to the absent stimulus will occur during the second phase. The Rescorla–Wagner model is defined as

$$\Delta V_i = \alpha_i \beta(\lambda - \Sigma V), \tag{1}$$

where V_i is the associative strength of stimulus i and parameters α_i and β (both range between 0 and 1) are learning rates related to the salience of the conditioned stimulus (CS) and unconditioned stimulus (US), respectively. The

P. C. Connor (✉) · T. P. Trappenberg
Faculty of Computer Science, Dalhousie University,
Halifax, Nova Scotia, Canada
e-mail: patrick.connor@dal.ca

V. M. Lolordo
Department of Psychology, Dalhousie University,
Halifax, Nova Scotia, Canada

parameter λ represents the total associative strength supportable by the US (a positive value when present and 0 when absent), and $\Sigma V = \Sigma_j V_j$ is the sum of the associative strengths of all stimuli present on a trial. The associative strength of a CS is updated after each trial in proportion to the surprisingness of the US ($\lambda - \Sigma V$) and the learning rate parameters associated with the CS and US. In recovery from overshadowing, if A and B have equal salience, an AB+ phase leaves each stimulus with half of the total associative strength. In phase 2, an A− treatment extinguishes stimulus A's associative strength but does not affect the associative strength of stimulus B, because its absence gives it a salience, α_B, of zero. Likewise, Wagner's standard operating procedures (SOP) model (Wagner, 1981) could not account for the phenomena. In SOP, associative elements representing each stimulus can be either in the inactive state or in one of two active states (A1 or A2, corresponding to being in focal attention or peripheral attention, respectively). Whenever a stimulus is presented, some of its elements are moved from the inactive state to A1. Over time, these elements' activity decays, moving from the A1 state into A2 and, eventually, into the inactive state. When a stimulus is presented, elements of its past associates enter A2, and again decay back to the inactive state. Associations between stimuli occur depending on the states in which their elements reside. If the elements of two stimuli are in the A1 state, the strength of their association is increased. If the elements of one stimulus are in the A1 state and the elements of another stimulus are in the A2 state, an inhibitory association from the first stimulus to the second stimulus is formed, but not vice versa. In the first phase of recovery from overshadowing, the elements of stimulus A and B are associated with those of the US and of one another. In the second phase, stimulus A is presented alone, causing its elements to enter A1 and activating elements of its associates B and the US into A2, thereby reducing its association with them. Stimulus B's association with the US remains unchanged, since no associations are formed between two stimuli concurrently present in the A2 state.

In response to retrospective revaluation phenomena, the Rescorla–Wagner model and SOP were retrofitted to explain them (Aitken & Dickinson, 2005; Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994). Both models make use of within-compound associations presumably developed in the first phase to associatively retrieve the absent stimulus. Then each model uses a mechanism to revalue the absent stimulus. In the case of Dickinson and Burke's extension of SOP, two associative learning rules are added. The first is that when the elements of two stimuli are in the A2 state, excitatory connections form between them. Also, when one stimulus is in the A1 state and another is in the A2 state, inhibitory associations between the stimuli are formed in both directions (e.g., A→B and B→A), instead of only one, as before. Now, when A is presented without the US in the second phase (i.e., recovery from overshadowing), the representations of both B and the US are in A2, satisfying the condition for formation of an excitatory association between them. Thus, during test, responding to stimulus B increases. If, instead, A is presented and reinforced in the second phase (i.e., backward blocking), the representations of A and the US will enter into the A1 state, while the representation of stimulus B, their associate, will be called into the A2 state. B in A2 and the US in A1 result in the formation of an inhibitory association between B and the US, thus reducing its associative strength. So the associative strength of B and the response to it in a test phase are greater in the recovery from overshadowing design than in backward blocking. In the case of the Van Hamme and Wasserman extension of the Rescorla–Wagner model, the absent stimulus is retrieved and given a negative alpha value. This leads to revaluation of the absent stimulus in the opposite direction as the presented stimulus. The Van Hamme and Wasserman model was recently elaborated upon by Witnauer and Miller (2011). Their model further develops the use of within-compound associations to enable it to account for second-order retrospective revaluation phenomena, which we will discuss later. Within-compound associations are also featured in other associative models of retrospective revaluation (e.g., Jamieson, Crump, & Hannah, 2012; Kasprow, Schachtman, & Miller, 1987; Kutlu & Schmajuk, 2012). For example, in the comparator hypothesis (Kasprow et al., 1987; Miller & Matzel, 1988; Stout & Miller, 2007), the within-compound associations become important during the test phase. The presence of a stimulus at test evokes previously paired stimuli through within-compound associations. The presented stimulus's response-evoking power becomes the associative strength of the presented stimulus B minus a fraction of the product of A's associative strength and the strength of the A→B within-compound association. In recovery from overshadowing, subsequent extinction of cue A makes the product of the B→A association and A→US association smaller than in a control group, thereby increasing the response-evoking power of cue B at test.

A few models have taken a different approach, explaining retrospective revaluation phenomena apart from within-compound associations. The APECS model (Le Pelley & McLaren, 2001; McLaren, 1993, 2011) explains these phenomena using configural units that represent memories of compound trials. A few elemental associative models (Dawson, 2008; Ghirlanda, 2005) also explain the phenomena apart from within-compound associations. Yet the models of Ghirlanda and Dawson bear a key flaw. As we will show, substantial revaluation in these models can occur without an associative history between the elements, an apparent failure to match the experimental data (e.g., Matzel et al., 1985) and the present general understanding of these phenomena. The present work pushes forward this enterprise by overcoming this particular issue. Since stimulus input excites activity in individual neurons of our model, the neurons begin to compete by inhibiting one another through lateral connections until only a subset (or ensemble) of neurons remains active. The activities of these neurons collectively express or represent the associative strength for the stimulus input. After a compound is conditioned, the constituent stimuli presented separately tend to activate a significantly similar ensemble of neurons. So, when one of the stimuli is subsequently conditioned/extinguished, learning in their common competitive lateral inhibitory connections is increased/decreased, which has the effect of oppositely revaluing the absent constituent. Without the compound conditioning phase, however, stimuli would activate largely separate ensembles of neurons (and lateral connections) and, thus, would not significantly affect one another during a revaluation.

Background

The models of Dawson (2008) and Ghirlanda (2005) represent simple elemental models of retrospective revaluation phenomena. Dawson offers a model similar to the Van Hamme and Wasserman (1994) extension but gives negative α values to all absent stimuli. Ghirlanda took a different approach, which we will now discuss in some depth. This model represents stimuli in a distributed format, instead of the usual one-to-one stimulus-input arrangement. Each stimulus in this model is described as a compound of many stimulus elements (ministimuli). For a feature such as color, we could model 100 stimulus elements spanning the visible color spectrum, each element representing a different wavelength. Here, Ghirlanda represents each punctate stimulus (e.g., stimulus A) as a Gaussian pattern of stimulus elements, as shown in Fig. 1. Formally, the input provided to Ghirlanda's model is

$$S_i = K + \sum_j \alpha_j e^{-\frac{\left(\frac{i}{N} - \mu_j\right)^2}{\sigma^2}}, \tag{2}$$

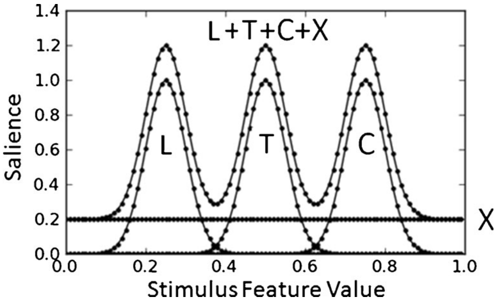

Fig. 1 Distributed stimuli used in a simulation of the Ghirlanda (2005) model. Gaussian-shaped stimuli L, T, and C represent conditioned stimuli, and the flat function X represents the context. The input to Ghirlanda's model is the sum of the present stimuli and the context, of which the example LTCX is given. There are 100 stimulus elements

where S_i represents the *i*th stimulus element's input salience, α_j is the salience of the *j*th stimulus (analogous to the Rescorla–Wagner model's α term), and N represents the number of stimulus elements. Each stimulus's Gaussian pattern of stimulus elements is centered about a specific feature value (e.g., wavelength) between 0 and 1 (μ_j) and has a certain width (σ). In simulations of Ghirlanda's model and our proposed model, we use 100 stimulus elements spanning between feature values of 0 and 1 and use $\sigma = \frac{1}{10\sqrt{2}}$ to fit several Gaussian-shaped stimuli into this range. The environmental context's representation does not have a Gaussian shape. Instead, it is represented as a flat function, such that all 100 stimulus elements have the same value, $K = 0.2$. In a simulated trial, the input provided to Ghirlanda's model (S_i) is the sum of the Gaussian patterns for each presented stimulus and the flat background context. The Gaussian shaped stimuli L, T, and C that we use in our simulations are shown in Fig. 1, as well as an example input LTC compound, which incorporates the context (X).

Learning proceeds after each trial as in the Rescorla–Wagner model, except that each stimulus element has an associative strength that is updated, instead of associative strengths for punctate CSs,

$$\Delta W_i = S_i \beta(\lambda - r_s) \tag{3}$$

$$r_s = \sum_i W_i S_i, \tag{4}$$

where W_i represents the *i*th distributed stimulus element's associative strength and r_s computes the total associative strength for all stimuli including the context on a given trial. For simple Pavlovian conditioning (i.e., AX+, X−, where X− represents the extinction of the context during the intertrial interval), the AX+ trials pull the W_i values upward toward $r_{AX} = 1$, while the X− trials pull W_i values downward toward $r_X = 0$. At the end of this tug-of-war, W_i values are found that satisfy both pulls, and thus asymptotes are reached. The resulting associative strengths can be pictured as a Gaussian curve shifted downward (negatively) by an amount similar to the context value, K.

One of the earliest investigations of recovery from overshadowing was reported by Matzel et al. (1985). In Experiment 3, they paired a light and a tone in the first phase, followed by reinforcement (TL+). They also reinforced separate presentations of a click stimulus (C+). In the second phase, they separated subjects into three groups: Group ET received nonreinforced presentations of the tone, Group EC received nonreinforced presentations of the click stimulus, and Group O was placed in conditioning chambers (as was done for the other groups), but no additional stimulus presentations were made. In the third phase, testing was performed. Results from their experiment and results from a simulation of this procedure using Ghirlanda's (2005) model are shown in Fig. 2. The first phase of simulation (TLX+, CX+, X−) leads to a set of W_i values that could be depicted as three negatively shifted Gaussians ($r_T = 0.50$, $r_L = 0.50$, $r_C = 1.0$). In the second phase, the extinction of the tone (r_T = 0.0) in Group ET inflates responding to the light (r_L = 0.61), which corresponds to the ordinal findings in Matzel et al. However, when we examine Group EC, where the separately conditioned click stimulus is extinguished, we find that responding to the light stimulus has also been inflated (r_L = 0.71), which is at variance with the experimental data in Fig. 2. The extinction of the tone stimulus in Group ET also inflated responding to the click stimulus above Group O, the control (Group ET, $r_C = 1.11$; Group O, $r_C = 1.0$) and extinction of the click stimulus in Group EC also inflated the tone above Group O (Group EC, $r_T = 0.71$; Group O, r_T = 0.50). These two revaluations also disagree with the experimental findings. In summary, these simulations show that retrospective revaluation in Ghirlanda's model does *not* require a history of compound conditioning. Instead, it predicts that the revaluing of a conditioned stimulus will substantially affect the associative strength of even separately conditioned stimuli. Dawson's (2008) model makes the same prediction, apparently employing a negative α for *all* absent stimuli. These simple associative models disagree not only with the findings of Matzel et al. and others (Cole, Barnet, & Miller, 1995; Miller, Barnet, & Grahame, 1992), but also with the current general understanding of these phenomena (but see Amundson, Escobar, & Miller, 2003; Escobar, Pineño, & Matute, 2002). In practical terms, if conditioning a stimulus could substantially alter the responses to unrelated stimuli, such interference could accumulate and confuse an organism about what each stimulus actually predicts.

Instead of simulating the experiment by Matzel et al. (1985), we could also have shown Ghirlanda's (2005) model in a within-subjects design retrospective revaluation paradigm, with AB+ and CD+ in the first phase and A+ and C− in the second phase. This combination of recovery from overshadowing and backward-blocking designs demonstrates retrospective revaluation when response to B in the test phase is smaller than the response to D (Chapman, 1991; Shanks, 1985). Since Ghirlanda's model revalues stimuli that have not been previously associated with one another, the conditioning of A affects stimuli C and D as much as it does stimulus B, and the extinction of C similarly affects A and B. Because the conditioning of A and extinction of C both pull on B and D, but in opposite directions, we will find a near zero change in the values of B and D relative to their pre-second-phase values, which is not the retrospective revaluation result commonly reported for this paradigm. Although interesting, the problematic mechanisms that bring about a

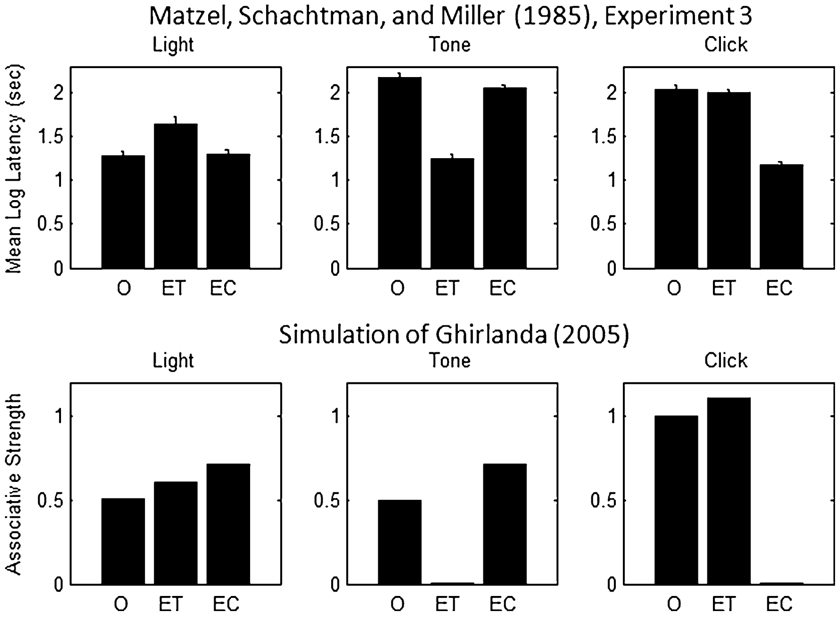

Fig. 2 Results of a lick suppression experiment (3) in Matzel, Schachtman, and Miller (1985) and its simulation using the model of Ghirlanda (2005). Responding shown in the upper panels is in terms of mean log latency (in seconds) to make 25 licks in the presence of the light stimulus. Longer latencies indicate greater suppression and greater associative strength. Corresponding simulations of associative strengths from Ghirlanda's model are provided in the lower panels. In the simulations, a procedure similar to that in the experiment was used ('X' is the context): phase 1**:** TLX+, X−, CX+, X−; phase **2:** Group O, X−, X−; Group ET, TX−, X−; Group EC, CX−, X−; phase 3**:** LX− TX−, CX− (all groups). Sufficient trials were used in each phase of simulation to ensure that responses to a stimulus reached asymptotic levels. In Ghirlanda's model, extinction of the tone in phase 2 (Group ET) inflated the light above the overshadowing control group (Group O), which corresponds to the findings of Matzel et al.. The extinction of the click (Group EC) in simulation, however, also strongly inflated the light, which is a failure to predict the associated experimental data. The extinction of the tone in the model also inflated the click and vice versa, but this also fails to occur in the data. Experimental data from Matzel el al. (1985), Experiment 3, used by permission

near zero change in this paradigm are readily understood from our simulation of recovery from overshadowing. Therefore, we do not pursue a within-subjects design in later comparisons.

As we will show, the present model overcomes the problems described above and yet, like the simple associative models, does not rely upon within-compound associations. In what follows, we will describe the present model and then look at the contributions of each of its mechanisms by enabling them one at a time while simulating several classical conditioning phenomena. Ultimately, we will arrive at an explanation for retrospective revaluation phenomena.

The present model

Here, we show how associative strength is computed and updated in our neural network model, as illustrated in Fig. 3. Where possible, we attempt to maintain biological plausibility and will briefly mention where this motivation influences certain design decisions. Possible neurobiological correspondences reflecting specific structural and learning rule details are discussed in a later section.

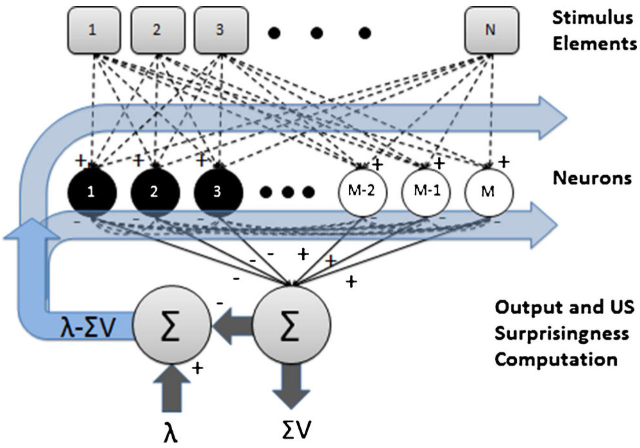

Fig. 3 The present model. The stimulus element inputs represented by the rounded boxes take exactly the same distributed input as that used in Ghirlanda's (2005) model, except that the context here is also modeled as a Gaussian pattern. Each dashed line in the model represents a connection that will (or will not) be established upon model initialization with some fixed probability. Neurons in the model, represented by circles, receive input and become excited. The connections between the neurons are inhibitory. These connections induce competition between the neurons, which reduces neuron activities and leads to a subset of neurons that dominates and suppresses all other neurons. The activities of the neurons are accumulated (bottom-center circle), where one half of these neurons add and the other half subtract from the sum. The total is appropriately scaled and represents the sum of associative strengths (ΣV) for the input stimuli. Conditioning is accomplished by changing the connection weights of model neurons. This is a function of the several factors including the US surprisingness (computed in the bottom-left circle), which is represented by the broad arrow leading back to the input and lateral connections. Importantly, the stimuli presented on a trial determine the ensemble of active neurons that develops through competition. Since it is the sum of activities of model neurons that gives the associative strength, the active neural ensembles come to represent the associative strengths of the stimuli that evoke them

We represent a stimulus in the same Gaussian form as in Ghirlanda's (2005) model. Unlike the Ghirlanda model, however, the present model does not use a flat function to represent the context. Instead, the context is expressed by its own Gaussian-shaped curve, to represent that an environmental context is really a collection of stimuli itself. This approach also allows the possibility of having different contexts, if so desired. Given a certain distributed CS as input, the model responds with activity in its neurons, which equals the excitatory input minus the lateral competition. Upon stimulus presentation, each neuron is allowed to settle into an internal activity (u_j) according to

$$\Delta u_j = \frac{1}{\tau}\left(u_j - \left(\frac{1}{N}\sum_{i=1}^{N} S_i w^I_{ij} - \frac{1}{M}\sum_{k=1}^{M} r(u_k) w^L_{kj}\right)\right) \qquad (5)$$

where $\tau = 10$, $N = 100$ is the number of stimulus elements, and $M = 2{,}500$ is the number of neurons in the model. We use a large number of model neurons because it improves the consistency of the simulation results. The distributed stimulus elements, S_i, are connected to each neuron with a certain connection probability ($P^I = .25$). Subsequent equations appear to invoke full connectivity. However, an absent connection is represented as an immutable connection weight of zero, which helps to simplify both the formal description and implementation of the model. This partial connectivity allows certain neurons to prefer activating in response to one stimulus or another and agrees with the reality that neural projections are not fully connected. The synaptic weights receiving stimulus input, $w^I{}_{ij}$, are initialized with a value of 20 for each connection made. Note that the indices *i* and *j* represent specific stimulus elements and specific model neurons respectively. So, instead of having a single weight per stimulus element, as in the Ghirlanda model, there is a single weight per stimulus element *for each neuron* in the model. Each neuron also has lateral synaptic weights, which receive inhibitory inputs from competing neurons. The lateral weight $w^L{}_{kj}$ is located on neuron *j* and receives input from competing neuron *k*. These weights are also initialized to 20 for connections made, and recurrent connections are permitted; the connection probability is $P^L = .25$. The term $r(u_k)$ is an activation function transforming the internal activation into a mean neuron firing rate,

$$r(u_k) = L(u_k)^2, \qquad (6)$$

where $L(u_k) = 0$ when $u_k < 0$ and otherwise $L(u_k) = u_k$p. $L(u_k)$ is a threshold-linear function (Usher & McClelland, 2001), which here reflects that neurons become silent when their internal activation goes below zero (analogous to real neurons). The outputs of these model neurons converge as a sum of the neuron firing rates, $r(u_j)$, with half of the neurons increasing the output and the other half decreasing it:

$$\Sigma V = \frac{\lambda_S}{M}\sum_{j=1}^{M} T(j) r(u_j), \tag{7}$$

$$T(j) = sgn\left(\frac{2j}{M} - 1\right), \tag{8}$$

where the sgn() function returns the sign of its argument and λ_S is a factor that translates the output from a sum of activities into units of associative strength (λ_S = 2,500). The function $T(j)$ is +1 for half of the neurons (hereafter called *positive* neurons) and −1 for the other half (*negative* neurons) based on their index, j. The final sum represents the combined associative strength of the input stimuli, or the expected associative strength (ΣV), the analogue to r_s in Equation 4 from Ghirlanda's model. This approach permits both positive and negative associative strengths by using a population of positively contributing and negatively contributing neurons. For a positive associative strength, the positive neurons are (on average) more active than the negative neurons. Negative associative strengths are expressed by the opposite difference of activity. The segregation of neurons into these two groups allows all input connection weights to be consistently positive, which is more plausible biologically speaking. There are few instances where a real neural connection can switch from having a positive to a negative influence. Also, as we will see, having two segregated groups of neurons contributes to generating a novel configural mechanism from elemental inputs.

Importantly, the ensemble of neurons that wins the competition and remains active is determined by the stimulus input provided. Because it is the activities of model neurons that ultimately combine to express associative strength, the active neural ensembles come to represent the associative strengths of the stimuli that evoke them. The learning rules cooperate by only updating the active ensemble neurons and only for nonzero stimulus elements (i.e., present stimuli). The learning rule used to update the input weights of each neuron is

$$\Delta w_{ij}^{I} = T(j) S_i \beta(\lambda - \Sigma V) L(u_j). \tag{9}$$

Note that updates to a neuron are proportional to its internal activation when it is above zero only (i.e., $L(u_j)$) and thus is part of the ensemble of active neurons. Also note that weights associated with distributed stimulus elements that have zero salience will also not change. This ensures that *input* weights are updated only for presented stimuli. The learning rule for the lateral weights, which receive inputs from other neurons, is

$$\Delta w_{kj}^{L} = T(j) H(u_k) \rho(\lambda - \Sigma V) L(u_j), \tag{10}$$

where $H()$ is the Heaviside or unit step function (1 when the argument is greater than zero and 0 otherwise), which means that learning will occur only if the sending neuron, indexed as k, is active. The parameter ρ in Equation 10 is the learning rate parameter for lateral weights (ρ = 0.5β). In this model, an individual neuron's weights, w_{ij}^{I} and w_{jk}^{L}, are always positive. This keeps the stimulus input influence excitatory and the lateral influence inhibitory in Equation 5. Weight changes must have opposite signs for the positive and negative neurons, so that these opposing pathways learn cooperatively. The function $T(j)$ defined in Equation 8 achieves this. Equation 9 is essentially a Hebbian learning rule (from its pre- and postsynaptic activity terms) modulated by the US surprisingness error term.

Classical conditioning simulations

The present model is readily integrated into trial-based simulations of classical conditioning experiments. A single trial consists of presenting stimuli, presenting the outcome (US or no US), computing the surprisingness, and adjusting synaptic weights according to the learning rules. Although we simulate phenomena that develop CS–US associations, the model does not explicitly exclude the notion of developing CS–CS associations, although these do not occur in the present simulations. The present model, like many others, does not define any CS–US timing, thereby excluding certain temporal phenomena (e.g., serial feature-positive discrimination) from its scope. Experimental findings and model predictions are ordinal in nature, so the usual assumption that associative strength is monotonically translated into conditioned responding is made here.

Salience levels play a role in our simulations. The US (β = .1) has a value of λ = 100 when the US is present and λ = 0 when it is absent. Conditioned stimuli have a salience of α = 1, while the context (X) has a salience of α = .2. Each intertrial interval is simulated like a single conditioning trial, where the context is presented but not reinforced (X−), just as in simulations of Ghirlanda's (2005) model. Associative strength accrued to the context is partly extinguished during these intertrial intervals. Parameters of the model used in simulations have been specified in the preceding section. A similar version of the present model was described in Connor and Trappenberg (2011) and shows how performance of the model varies within appropriate ranges for selected parameters.

An example of how conditioning proceeds in this network is shown in Fig. 4. In excitatory conditioning (CS→US)

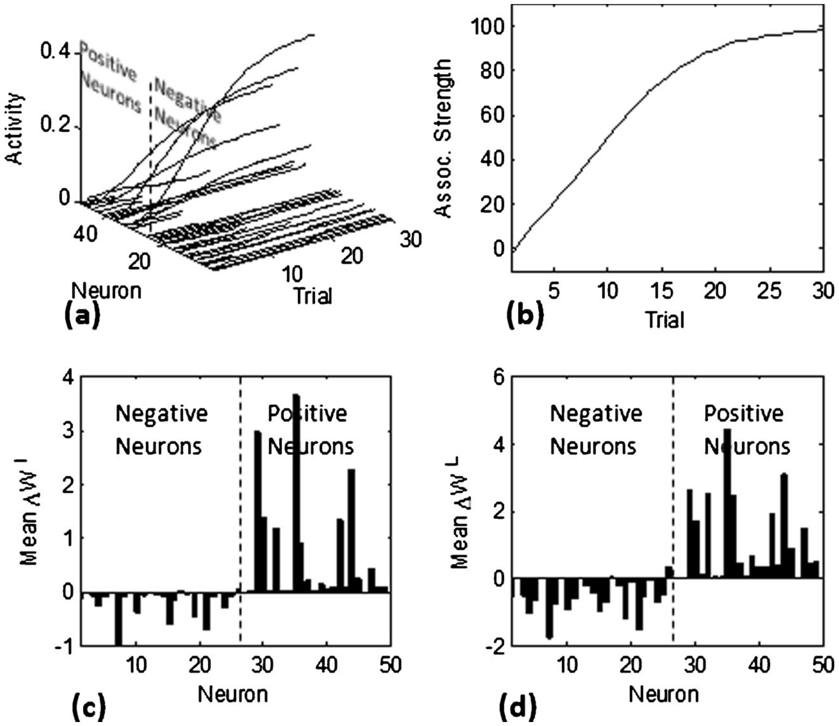

Fig. 4 The present model during excitatory conditioning, simulated using only 50 neurons for demonstration purposes. **a** Activity in some positive neurons (neurons 26–50) increases with the number of trials. Other neurons lose the competition and are silenced. Negative neurons (1–25) are either suppressed or very weakly active. **b** Overall associative strength increases, approaching asymptote within 30 trials. **c** The average change in input synaptic weights for each neuron between the first and last trials shows a substantial increase for positive neurons and a slight decrease for negative neurons. **d** Lateral synaptic weights also increase for positive neurons and decrease for negative neurons

trials, the synaptic weights of active positive neurons are increased, while the synaptic weights of active negative neurons are decreased. This results in increased activity in the positive neurons and decreased activity in the negative neurons for subsequent trials. The difference between the activity in these two pathways gives a final positive associative strength. As conditioning trials continue, the associative strength will grow until it matches that supportable by the US. More and more neurons are also silenced through lateral inhibition as an ensemble of neurons increasingly dominates. During extinction, the opposite process happens, cutting down positive neuron activity and restoring negative neuron activity. Also, the active ensemble will grow to include more neurons once again. In short, increases and decreases in associative strength track increases and decreases of synaptic weights in the positive neurons, and the opposite relationship exists between associative strength and the synaptic weights of negative neurons.

In the following sections, we use additional classical conditioning simulations to show how certain model mechanisms affect model behavior. Beyond the acquisition example in Fig. 4, we do not revisit demonstrations of the mechanisms borrowed from the Rescorla–Wagner model but, rather, focus on the unique mechanisms of the present model. Building upward, we first show how the combination of the activity-dependent learning term $L(u_j)$ and having dual pathways (i.e., positive and negative neurons) develops configural representations from individual stimuli. Then we demonstrate how adding lateral inhibition sculpts ensembles of active neurons and, finally, how learning in these lateral connections enables retrospective revaluation effects.

Activity-proportional learning and dual pathways perform configuration

Recall that there are positive and negative neurons in the model whose influences sum to provide the overall associative strength. Changes to the weights of these neurons are made in proportion to their internal activation, $L(u_j)$ (Equations 9 and 10). The combination of these two mechanisms leads to the development of configural cues. To demonstrate this, we simulate the negative patterning procedure, which in a single phase interleaves trials of AB− with A+ and B+ trials. The ordinal finding is that responding to the compound AB during a subsequent test is less than responding to either A or B alone (Delamater, Sosa, & Katz, 1999; Harris, Gharaei, & Moore, 2009; Redhead & Pearce, 1995; Woodbury, 1943). To demonstrate the combined efforts of the two mechanisms, Fig. 5 shows simulated negative patterning results for our model with and without each of them. In addition, results when lateral inhibition and lateral learning are enabled are also given to

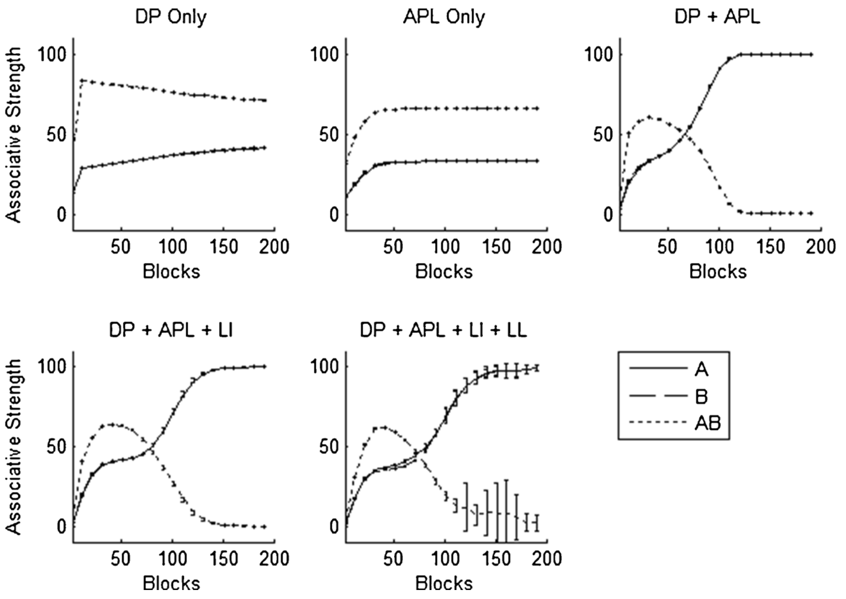

Fig. 5 Simulation of negative patterning using various configurations of the present model for 15 differently initialized models or stat rats (error bars are barely visible for all but one of the panels). Each block consists of three trials (A+, B+, AB−). Negative patterning requires that both the positive and negative neurons exist and that there is activity-proportional learning. The lateral inhibition and lateral learning mechanisms do not assist but also do not substantially interfere. DP, dual pathway; APL, activity-proportional learning; LI, lateral inhibition; LL, lateral learning

show that these additional mechanisms do not interfere. Note that to disable the dual-pathway nature of the model, we eliminate excitatory input to the negative neurons to silence them. To disable activity-proportional learning, we simulate without the last term in Equations 9 and 10. Disabling lateral inhibition is accomplished by fixing all lateral weights to zero and disabling lateral learning is done by setting $\rho = 0$.

As Fig. 5 shows, when either the dual-pathway or activity-proportional learning mechanisms work alone, negative patterning fails. When both mechanisms are engaged, however, the phenomenon emerges. The way in which the model accomplishes this can be seen from the input weights. Figure 6 shows that when activity-proportional learning and both pathways are enabled, positive neuron weights specialize for either stimulus A or B, while negative neuron weights grow similarly for each stimulus. When only a single stimulus (A or B) is present, the specializing positive neurons activate strongly, whereas the unspecialized negative neurons activate little. The sum of large positive neuron activations minus small negative neuron activations results in asymptotic (λ) conditioned responding. Now, recall that the activation function is $r(u_j) = L(u_j)^2$, which means that doubling the stimulus input (which doubles u_j) will quadruple a neuron's output. So, when both stimuli are present (AB), the negative neurons' activations are increased exponentially, which enables them to balance out the positive neuron activations, resulting in zero conditioned responding. The weights develop as follows. Initial stages of training show that for positive neurons that are activated more strongly for a certain stimulus (e.g., A), the input weights increase more on A trials than they decrease on AB− trials. Conversely, weights receiving connections from the stimulus for which a positive neuron's activation is weaker (e.g., B) will have a net decrease because their reinforced trial increases the weights less than does the decrease occurring from AB− trials. Negative neuron weights decrease in early stages because the reinforced trials have a larger positive error term (which decreases negative neuron weights) than the negative error term on AB− trials. As associative strength increases to the individual stimuli, however, this situation reverses, and negative neuron weights begin to grow more on the AB− trials and do so roughly evenly for A and B.

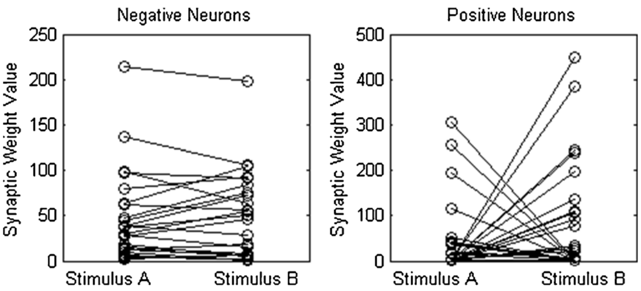

Fig. 6 Correlation between the weights in a random selection of model neurons and stimuli A ($\sum S_i^A w_{ij}^I$) and B ($\sum S_i^B w_{ij}^I$) when both pathways and activity-proportional learning are enabled (i.e., lateral inhibition and lateral learning are disabled). Negative neurons grow relatively evenly for both stimuli A and B, making them respond substantially more to the compound AB than to A or B alone. In contrast, positive neurons' weights tend to specialize (increase) for either stimulus A or B and decrease for the other stimulus

Although not the focus of the present article, this approach to developing configural cues is quite novel. Its elemental basis puts it in the same realm as Harris's elemental model (Harris, 2006) and the replaced elements theory (Brandon & Wagner, 1998; Wagner, 2003; Wagner & Brandon, 2001). In such models, each stimulus is represented by a set of elements. Elements that are active (or within an attentional buffer) during conditioning receive larger changes in associative strength than do the others. Certain elements for each stimulus are allowed to become activated depending on whether the stimulus is presented alone or in compound. Therefore, in the negative patterning procedure, some stimulus elements are primarily conditioned in the single-stimulus trials, but not the compound trials, and vice versa. This allows some elements to encode the single-stimulus associative strength and others to help represent an opposite compound associative strength. The present model departs radically from this idea, not needing to deactivate stimulus elements but, instead, deriving its configural ability from its dual pathways and activity-proportional learning in the context of a squared activation function. Additional work is needed to evaluate this approach by simulating other experimental paradigms and drawing thorough comparisons with the other configural models on the market. For the present work, however, we have focused on the model's ability to explain retrospective revaluation phenomena, which is supported by lateral inhibitory connections and the learning therein, to which we now turn.

Adding lateral inhibition

Figure 7 shows the neuron response following negative patterning for a model with only 200 neurons for illustration purposes, where both panels show models that have activity-proportional learning and both pathways are enabled. In the top panel, lateral inhibition is disabled, and in the bottom panel, it is enabled (without learning, $\rho = 0$). Although the difference in associative strength between the two conditions will be small (see, e.g., the comparison in Fig. 5), the neural activity takes a new form. Without lateral inhibition, all neurons are active for every input. When lateral inhibition is enabled, a unique ensemble of active neurons takes shape in response to the presentation of each stimulus or compound. On the surface, it may seem that this mechanism is very similar to the replaced elements mechanism of models, noted above, that perform configuration. Although lateral inhibition may technically be able to behave in this way, it is ancillary in our simulations. In fact, lateral inhibition does not appear to explain any additional phenomena in this context, besides taking part in helping lateral learning explain retrospective revaluation phenomena, which will be discussed later. A potential benefit, however, is that because it uses fewer neurons to represent the same information, the overall capacity of the system to learn further stimulus–outcome relationships should increase.

With lateral inhibition enabled, unique active ensembles emerge for specific input stimuli, such that when the input stimuli change, so will the active ensemble to some degree. Roughly speaking, the weights of the neurons in the ensemble match the distributed stimulus input profile more closely than do the weights of neurons that are silenced. Thus, the similarity between two ensembles depends on the similarity of the two stimuli and is reduced as stimuli become dissimilar. This is shown in Fig. 8, where the ensemble similarity between a previously conditioned stimulus having a specific

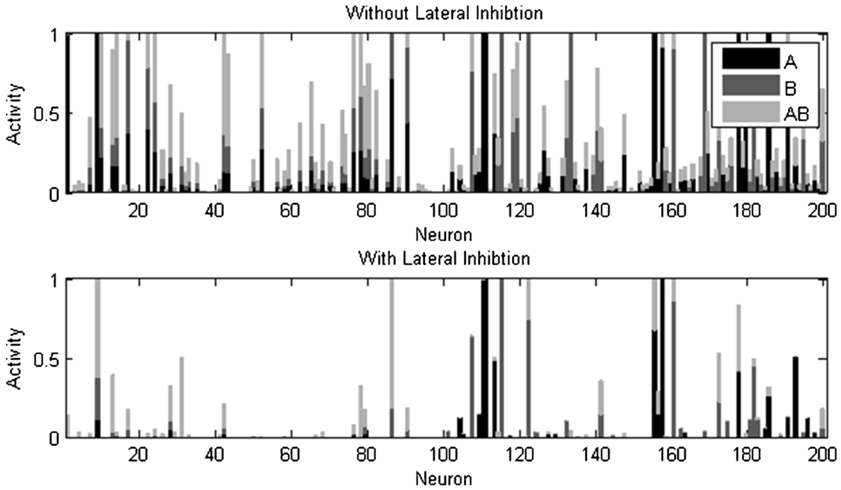

Fig. 7 Model neuron activity after negative patterning. Using only 200 neurons for demonstration purposes, the simulated activity for each stimulus or compound is computed and drawn as a stacked column in the bar graph, where each column represents one neuron. The length of each differently shaded bar in the stack is the amount of activity observed for the condition it represents. The left half of the neurons (1–100) are negative neurons, and the right half (101–200) are positive neurons. When lateral inhibition is disabled, all neurons respond to some degree for every stimulus and, thus, take part in representing every stimulus's associative strength. When lateral inhibition is enabled, however, only a fraction of the neurons are active for any given stimulus. This means that each neuron takes part in representing only certain stimuli's associative strengths

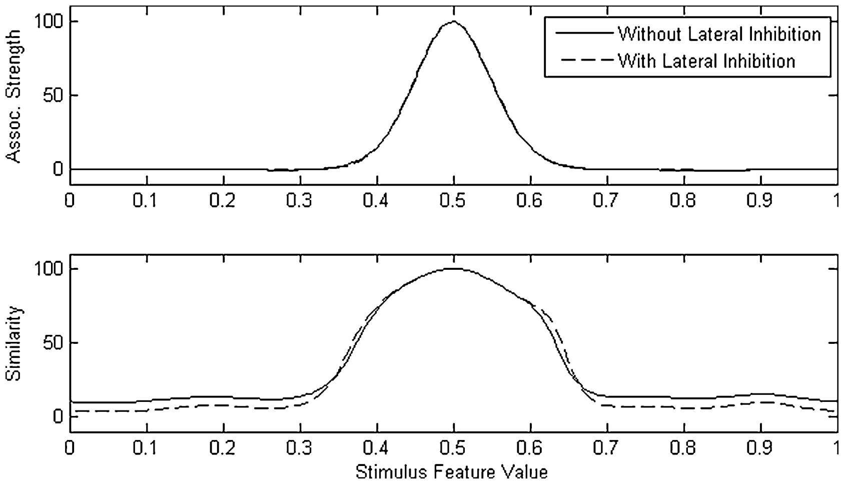

Fig. 8 Measures of associative strength and active ensemble similarity between a previously conditioned stimulus (feature value = .5) and all other feature values (0 to 1) with and without lateral inhibition. In both cases , we see that CSs with similar feature values evoke substantially similar ensembles and, thus, associative strengths. Adding lateral inhibition tends to lower the similarity between the ensembles activated by unrelated stimuli. Similarity is computed as the cosine of the angle (i.e., the normalized dot product) between the neural ensembles activated for the previously conditioned stimulus and the test stimulus

feature value (e.g., green light) and all feature values (i.e., green, red, blue, yellow, etc.) is computed, both for when lateral inhibition is disabled and for when it is enabled (but no lateral learning). Although there is no difference in associative strength with and without lateral inhibition (upper panel), the similarity between the neural activations of vastly different features (e.g., 0.5 vs. 0.2 or 0.5 vs. 0.9) is smaller with lateral inhibition (lower panel). Without lateral inhibition, there will be a greater similarity between the neural activations of different features, because every neuron is active for every feature.

Until now, we have discussed similarity between two distinct stimuli. Consider a related case in which one stimulus is joined by a second stimulus to make a compound. Because there is substantial similarity between the compound and its constituents, the activation of neurons by the compound will be more similar to the activation evoked by one of its constituents than to the activation evoked by an unrelated stimulus.

Adding lateral learning

Simulations of recovery from overshadowing using the present model are shown in Fig. 9, following Matzel et al. (1985) and the simulations of Ghirlanda's (2005) model described earlier. In particular, these simulations show (1) that recovery occurs only when lateral learning is enabled and (2) that revaluing a conditioned stimulus significantly affects only the associative strengths of stimuli with which it was previously paired, and not unrelated stimuli.

To understand how lateral learning accomplishes all of this, we will focus on two positive neurons and explain how recovery from overshadowing can occur, as shown in Fig. 10. Negative neurons do not play a major role in this phenomenon but may take a more significant role in other retrospective revaluation phenomena (e.g., recovery from conditioned inhibition by extinction of the excitor). Excitatory conditioning of a compound (phase 1) increases the input weights of its active neurons. Because $\rho > 0$, the lateral inhibitory connections between the active neurons grow as well. In phase 2, only one of the constituents (A) is presented. Because of its history of activating the ensemble associated with the compound (AB), there is a substantial degree of similarity between the ensembles activated by AB and A, and thus these same two positive neurons are activated again. When the presentation of A is followed by no reinforcement, these active neurons' A-specific input weights and their lateral weights are decreased. In phase 3, when the absent stimulus (B) is tested, we detect a change. Although the B-specific input weights did not change (because there was no input from stimulus B in phase 2), its active ensemble's lateral inhibitory weights are smaller. As a result, there is less inhibition, which increases these positive neurons' overall activities and, thereby, increases the associative strength of B. Intuitively speaking, excitatory conditioning of a compound ties its constituents together in terms of causing them to activate similar ensembles of neurons in future trials. Then interactions occur between these stimuli through lateral learning in their shared connections. The result is that extinguishing one increases the associative strength of the other (i.e., recovery from overshadowing), and increasing one's associative strength will decrease the other's (i.e., backward blocking). In this way, the shared lateral connections play a similar role as the within-compound associations found in other models but do not retrieve explicit stimuli per se. Without the compound conditioning step, as is the case for unrelated stimuli, there would be fewer shared neurons (and thus, lateral

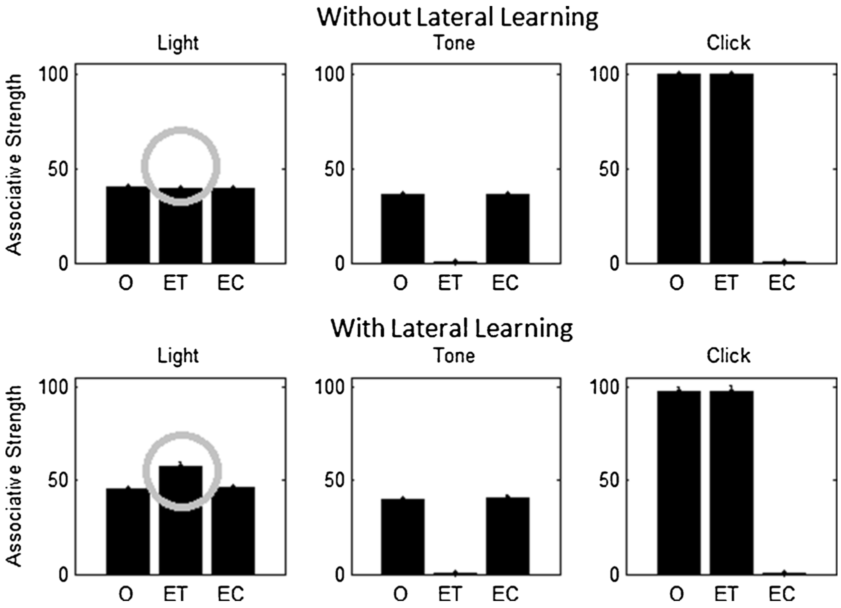

Fig. 9 Simulations of recovery from overshadowing (Matzel, Schachtman, & Miller, 1985, Experiment 3) using the present model when lateral learning is disabled (ρ = 0) and enabled (ρ > 0). Error bars represent the small deviation in results for 15 differently initialized models (stat rats). The simulation procedure matches that used for earlier simulations of the Ghirlanda (2005) model: phase 1 (50 trials)**:** TLX+, X−, CX+, X−; phase 2 (200 trials), Group O, X−, X−; Group ET, TX−, X−; Group EC, CX−, X−; phase 3 (1 trial), LX−, TX−, CX− (all groups). Circled in the results, we see that extinction of the tone in phase 2 of the simulation (Group ET) revalued (inflated) the light above the control group (Group O) when lateral learning is enabled, but not when it is disabled. Also in agreement with the experimental data, the simulations did not substantially revalue any other stimuli (regardless of whether or not lateral learning was enabled), in contrast to the simulations of Ghirlanda's model

connections) in the ensembles of the individual stimuli. As a result, there would be far less change in the lateral inhibition for an absent stimulus were an unrelated stimulus presented and revalued.

Second-order retrospective revaluation and relation to other models

In recent years, theorists have focused on the phenomena of second-order retrospective revaluation. The second-order retrospective revaluation procedure involves conditioning, in successive phases, two compounds that share a common element (i.e., phase 1, AX+; phase 2, XB+) and, in a third phase, revaluing one of the nonshared stimuli (A).

Houwer and Beckers (2002) ran three experiments, using a weapons/tanks procedure. In the first experiment they did the following: phase 1: CT1+, phase 2: T1T2+, and phase 3: either C+ or C− between groups. Group C+ had a much higher rating of C, a lower rating of T1, and a higher rating of T2. The next experiment looked at third-order retrospective revaluation: phase 1: CT1+, phase 2: T1T2+, phase 3: T2T3+, and phase 4: C+ or C− between groups. Group C+ had higher ratings for C and T2 than Group C− did, but lower ratings for T1 and T3. All the effects were substantial. The next experiment looked at second-order retrospective revaluation in a within-subjects design—essentially, phase 1: C1T1+, C2T3+, phase 2: T1T2+, T3T4+, and phase 3: C1+, C2−. Ratings of C1 were higher than those of C2, and T3 was rated higher than T1 (first-order retrospective revaluation). There was also a big second-order retrospective revaluation effect: T2 was higher than T4. Melchers, Lachnit, and Shanks (2004) obtained results similar to those of Houwer and Beckers in a within-subjects experiment within the foods/allergies setting. Their Experiment 3 looked at second-order retrospective revaluation (e.g., phase 1: AB+, BC+, phase 2: C+ vs. phase 1: DE+, EF+, phase 2: F−) and its direct analogue (e.g., where the element trials came before the compound trials). First-order retrospective revaluation occurred (e.g., B < E), and second-order retrospective revaluation was in the opposite direction (e.g., A > D). Denniston et al. (2001), using rats, employed a between-groups paradigm: phase 1: CA+, phase 2: BA+, then either C− or nothing. The conditioned response to B was lower in the C− group than in the controls. This finding is consistent with Houwer and Beckers and with Melchers, Lachnit, and Shanks (2004).

McLaren, Forrest, and McLaren (2012) reported an experiment on retrospective revaluation using the foods/allergies setting. First- and second-order retrospective revaluations were assessed in a within-groups design: phase 1: BC+, DE+, phase 2: AB+, EF+, and phase 3: A+, F−. Ratings of B and C both declined, relative to D and E, and the first-order effect was about as big as the second-order effect. Their

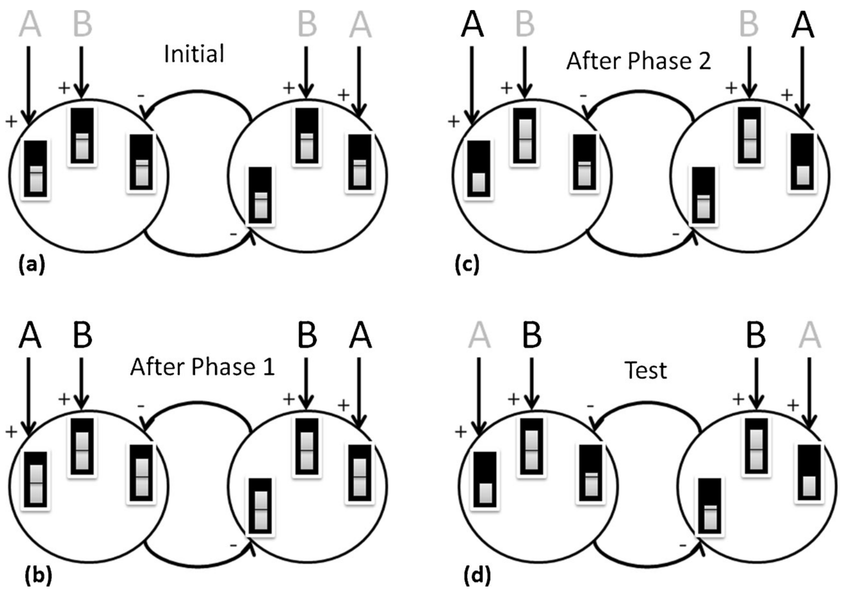

Fig. 10 Recovery from overshadowing as demonstrated in the present model. This diagram focuses on two positive neurons represented by circles that are active whenever A, B, or AB is presented. Each neuron receives excitatory inputs from stimuli A and B and an inhibitory connection from the other neuron. a The neurons' synaptic weights, which are represented thermometer style in the rectangles associated with each connection, are initialized to about half value. b After conditioning to compound AB (phase 1), input weights connecting A and B to the neurons are increased. Also increased are the lateral weights between these active neurons. c In the second phase, A is presented but not reinforced, which decreases its input weights and lateral weights. d Subsequent testing of B shows an increase of associative strength. Although B's input weights are unchanged, its lateral weights have decreased. Less inhibition means greater activity in these positive neurons, which translates into more associative strength (Equation 7)

second-order result is the opposite of the findings described above. McLaren et al. reported that if they instead provided all the data at once on handouts, which they interpreted as entailing a low memory load, then they got a different result; ratings of B and C moved in opposite directions after A+, and ratings of D and E moved in opposite directions after F−. They further suggested that the findings of Melchers et al., which were opposite to their own, were the result of a relatively low memory load. This hypothesis needs to be tested in an experiment that varies memory load.

So, we have a data conflict, but what do the models predict? McLaren et al. (2012) said that their participants who received all the data on handouts reported using rational inference to derive their conclusions, so that after phase 1: BC+, phase 2: AB+, and phase 3: A+, they reasoned that if food A was responsible for the allergy, then food B must not have been, and if food B was not responsible, then food C was responsible.

Witnauer and Miller (2011) compared the second-order retrospective revaluation predictions that are made by the Van Hamme and Wasserman (1994) extension of the Rescorla–Wagner (1972) model with their own extension that involved more development of the role of within-compound associations. The Van Hamme and Wasserman extension modeled retrospective revaluation by updating an absent stimulus's associative strength with a negative learning rate, whenever a stimulus was presented with which it had a within-compound association. Witnauer and Miller's extension additionally multiplied this by the sum of within-compound associations between each of the present stimuli on a trial and the absent stimulus. Witnauer and Miller show that while both models demonstrate first-order retrospective revaluation effects, only their extension demonstrates the most commonly observed second-order retrospective revaluation effects, in which the first- and second-order associates move in opposite directions. It appears that the critical difference is that Witnauer and Miller's enhanced within-compound model encodes the sign of the within-compound associations (i.e., the inhibitory association between the nonshared elements of the two compounds), whereas the Van Hamme and Wasserman extension does not. Witnauer and Miller noted that Stout and Miller's (2007) SOCR model also predicts the second-order (and higher-order) effects and that Dickinson and Burke's (1996) modification of SOP does not. They concluded that all models that can explain the most commonly observed higher-order retrospective revaluation effects use within-compound associations.

In a second-order retrospective revaluation experiment with phase 1: AB+, phase 2: BC+, phase 3: A−, the present model predicts a different result than would be made by within-compound models. Because of lateral learning, recovery from overshadowing will occur to the shared element (i.e., B's associative strength will increase), but the model also predicts that the other, nonshared element (C) will also elicit more responding, when tested. The reason for this is that when BC is conditioned, it will gravitate toward using a relatively similar ensemble of neurons as the previously conditioned AB. As a result, as A is extinguished, C's somewhat similar ensemble will have its lateral inhibition lowered as well. This leads to greater positive neuron activity upon presentation of C and, thus, greater associative strength.

If retrieval by within-compound association is the mechanism by which retrospective revaluation occurs, we would expect that large within-compound associations should lead to greater retrospective revaluation than weak within-compound associations. Consider the following procedure: phase 1: AX+, phase 2: AX-−, phase 3: AY+, BX+, phase 4: A−, phase 5: X−, Y− (test). After the first two phases, the within-compound associations between A and X should be relatively large, despite the fact that responding to AX after the second phase should have returned to near initial conditions (i.e., low responding). In the third phase, A and X are separated but conditioned in compound with Y and B, respectively. Given that the AX within-compound association is stronger than the AY within-compound association after phase 3, then within-compound models predict that stimulus X should be revalued more than stimulus Y. The present model makes the opposite prediction, that Y will be revalued more than X. The first phase develops a neural ensemble for AX, but the second phase extinguishes this, essentially restoring the network to initial conditions. The third and fourth phases are then seen as a simple recovery from overshadowing paradigm, where Y is revalued more than X. Rational inference makes the same prediction as the present model because, at the end of phase 2, the inference would be that neither A nor X predicts the US. In this way, the third and fourth phases become a simple recovery from overshadowing paradigm.

The present model differs from other models of retrospective revaluation that do not employ within-compound associations. Although it revalues an absent stimulus according to associative mechanisms, it does so only when the stimulus presented in the second phase was previously paired with the absent stimulus to be revalued (i.e., unlike Ghirlanda, 2005, and Dawson, 2008). The present model also does not make use of memory retrieval, although this is another route apart from within-compound associations to explain the phenomena. For example, the APECS model (Le Pelley & McLaren, 2001; McLaren, 1993, 2011) takes this approach. It is a connectionist style approach that recruits a new hidden layer neuron for each unique trial it experiences (e.g., separate nodes for A+, A−, and AX+). Although a detailed description of the model is not feasible here, the bias of a node representing a compound behaves in much the same way as our lateral inhibition mechanism. In recovery from overshadowing, the first phase establishes a compound node ("AB+") and associates it with the US. During the intertrial intervals of this phase, the "bias" weight for this node is made negative, to offset the increased prediction made by the node when the inputs are absent. In the second phase, a new node is established ("A−"), and during the intertrial intervals of this phase, the "AB+" node's bias is increased. This increases the "retrievability" of node "AB+," which then leads to an increased response upon presentation of stimulus B (i.e., recovery from overshadowing). The bias of the APECS model functions like the lateral inhibition of the present model, except that it has the opposite sign: In our model, during extinction of stimulus A in the second phase, lateral inhibition is decreased, making the positive neuron response to B larger. In both models, the second phase does not change the input weights associated specifically with the absent (B) stimulus but, rather, the lateral weights for the present model and the bias for APECS.

Having two opposing pathways to compute associative strength is also a feature of the comparator hypothesis (Kasprow et al., 1987; Miller & Matzel, 1988). However, the comparator hypothesis uses the second pathway to evoke CS–CS associations and compare the associative strengths of different stimuli, while the present model simply uses the second pathway to help represent negative associative strengths. The dual-pathway structure also bears resemblance to the division of CS–US and CS–no-US associations discussed in Le Pelley (2004). A model of spontaneous recovery from extinction by Pan, Schmidt, Wickens, and Hyland (2008) uses positive and negative weights, which are changed in opposite directions and are summed to produce a measure of responding.

Application to other retrospective revaluation findings

In Fig. 11, we show that the present model can also explain the backward-blocking effect (Denniston, Miller, & Matute, 1996; Shanks, 1985; Wasserman & Berglan, 1998). Using the backward-blocking procedure in Shanks along with an additional control group (see Fig. 11 for details), we correctly simulate the effect ($p < .001$, Wilcoxon signed-rank test, for 15 different model initializations or stat rats) with the present model. In the figure, it appears that without lateral learning, the procedure increases, rather than decreases, responding to the blocked stimulus relative to a control group (BX). However, this is simply due to greater extinction of the context in the control group, which is overwhelmed when lateral learning is enabled.

Backward conditioned inhibition (Chapman, 1991; Urcelay et al., 2008) refers to the paradigm in which a nonreinforced compound is presented in the first phase (AX−) followed by a phase where one element is reinforced (A+). The result is that the other element becomes inhibitory relative to a control group. An experiment by Espinet, Iraola, Bennett, and Mackintosh (1995) preexposed compounds AX and

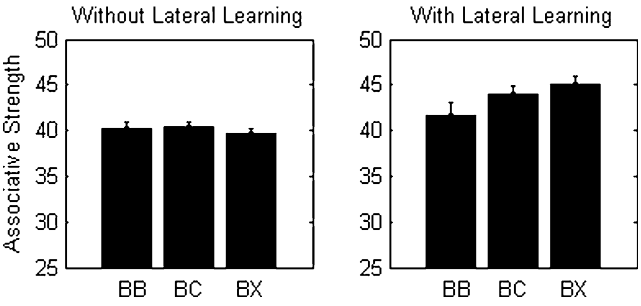

Fig. 11 Responding to stimulus B in the test phase of backward-blocking simulations when lateral learning is disabled (i.e., $\rho = 0$) and enabled (i.e., $\rho > 0$) using the paradigm of Shanks (1985) and an additional control group. With lateral learning enabled, the backward-blocking group (Group BB: phase 1 (50 trials): ABX+, X−, phase 2 (200 trials): AX+,X−, phase 3 (1 trial): BX− [Test]) expressed lower responding ($p < .001$, Wilcoxon signed-rank test, 15 differently initialized simulations or stat rats) than did both control groups, Group BC and Group BX. In Group BC, phase 2 trials reinforced a novel stimulus (phase 2: CX+ X−), while in Group BX, phase 2 trials did not involve any stimulus presentations (phase 2: X− X−). In the other phases, these groups received the same treatment and test as Group BB. Note that in phase 2, conditioning of A and the novel stimulus C reached asymptotic levels of responding in their respective groups. This simulation shows that lateral learning leads to a weak but significant backward-blocking effect

BX (AX−, BX−). In the second phase, conditioning to one of the nonshared constituents was conducted (A+). The result of these manipulations was that stimulus B's association with illness either was weakened or became inhibitory. This is called the *Espinet effect*. Formally, the Espinet paradigm is the second-order analogue of the backward conditioned inhibition paradigm. The present model does not explain these effects. For the same reasons as in the Rescorla–Wagner model, preexposure has no effect on subsequent conditioning phases. As a result, no sharing of neural representations occurs, unlike compound conditioning when a US is presented. If the present model were extended to develop similar neural representations during a preexposure phase (as occurs in a conditioning phase), then the model might also come to explain these two effects.

Reminder-induced recovery from overshadowing (Kraemer, Lariviere, & Spear, 1988) is the finding that presentation of the overshadowed stimulus somewhere between the conditioning sessions and test sessions (the day following conditioning and 2 days prior to test in Kraemer et al., 1988) enhances responding to an overshadowed stimulus. Corresponding reminder-induced recovery has also been discovered in the blocking (Schachtman, Gee, Kasprow, & Miller, 1983), relative validity (Cole, Denniston, & Miller, 1996), and latent inhibition (Kasprow, Catterson, Schachtman, & Miller, 1984) paradigms. One prominent interpretation of the reminder-induced recovery from overshadowing findings is that the overshadowed stimulus's associative strength is not reduced by being conditioned in compound, as the notion of cue competition suggests, since later we find that responding has "recovered." More formally, the interpretation says that overshadowing is due to a deficit in performance (e.g., memory retrieval failure in the test phase), rather than to a deficit in acquisition through cue competition in the conditioning phase. The question remains, however, as to what mental processes the reminder treatment might invoke. One remaining potential acquisition-deficit explanation (but see Schachtman et al., 1983) is that the reminder treatment strengthens a within-compound association between the overshadowed stimulus and the overshadowing stimulus. Then, when the overshadowed stimulus is later tested, the overshadowing stimulus is thereby retrieved and submitted as internal input to the associative learner, thereby generating a greater (or "recovered") level of responding. A similar mechanism might also explain spontaneous recovery from overshadowing (Kasprow, Cacheiro, Balaz, & Miller, 1982), which has also been thought to indicate a performance deficit rather than an acquisition deficit in learning. In this phenomenon, responding to the overshadowed stimulus is greater after a retention interval. The acquisition-deficit explanation would say that time, instead of a reminder treatment, may lead to stronger within-compound associations. There is some evidence within a sensory preconditioning paradigm, however, that within-compound associations degrade, rather than strengthen, when there is a delay between conditioning and test (Pineño, Urushihara, & Miller, 2005). Additional experimental work testing the strength of within-compound associations after reminder treatments and postacquisition delays may better discriminate between the performance-deficit and acquisition-deficit explanations.

Retrospective revaluation effects are not always observed (Dopson et al., 2009; Shevill & Hall, 2004). Within-compound association-based approaches can often explain this as a failure in within-compound association-based retrieval during either conditioning or test. In the present model, retrospective revaluation phenomena are reduced when the conditioned stimuli in the initial pairing are similar—that is, when there is significant overlap between their distributed input representations. Consider when stimuli are nearly identical. This will generate strongly similar input representations such that subsequent conditioning or extinction of one will similarly affect the other, because the model treats them as essentially the same stimulus. This is the opposite of retrospective revaluation behavior and is referred to as mediated conditioning, which has been found to occur when the paired stimuli are strongly similar (Liljeholm & Balleine, 2009). Then to observe neither mediated conditioning nor retrospective revaluation in the present model, one explanation is that the stimuli making up a compound stimulus have some middle-ground degree of similarity. From our model, another possible explanation of why retrospective revaluation is sometimes not observed is that the lateral learning rate (ρ) changes dynamically. In the present model, we set $\rho > 0$, which supports retrospective revaluation phenomena. As is noted in Fig. 9, when $\rho = 0$, no retrospective revaluation occurs.

Furthermore, if we set $\rho < 0$, the opposite of retrospective revaluation would occur (i.e., mediated conditioning) because lateral learning would change weights in the opposite direction. For example, recall the process in Fig. 10. Instead of reducing lateral weights, which increased the absent stimulus B's associative strength, the lateral weights would be increased, which would reduce activity and associative strength when B is presented. Thus, as stimulus A is extinguished, so would be its previously partnered stimulus B.

In a recovery from overshadowing experiment, Liljeholm and Balleine (2006) found that the extinction of the more salient element of the compound revalued the less salient element more than the other way around. In the present model, the more salient stimulus takes a larger share of the associative strength due to cue competition. This means that the more salient stimulus will have more associative strength to extinguish and its lateral weights will be reduced proportionally. Thus, extinction of a salient stimulus will lead to more revaluation of the absent stimulus than will extinction of a weakly salient stimulus. In more general terms, the larger the change in a present stimulus's associative strength, the more the absent stimulus is revalued.

The present model's relationship to the neurobiology of the striatum

The present model, although described in an abstract way, can be readily related to features of the neurobiology of the ventral striatum and basal ganglia. The positive and negative neurons map to the striatal projection neurons belonging to the direct and indirect pathways of the basal ganglia, respectively. The input weight learning rule (Equation 9) corresponds with experimental findings regarding the effects of dopamine ($\lambda - \sum V$) and pre- (S_i) and postsynaptic activity ($L(u_j)$) on cortico-striatal synapses (Frank & Fossella, 2011; Reynolds & Wickens, 2002; Schulz, Dayan, & Montague, 1997; Shen, Flajolet, Greengard, & Surmeier, 2008). The key feature of the present model that sets it apart from related neurobiological models of the basal ganglia (Frank, 2005; Houk, Adams, & Barto, 1995) is its lateral inhibition and lateral learning. The projection neurons of the striatum have inhibitory (Plenz, 2003; Tunstall, Oorschot, Kean, & Wickens, 2002) lateral connections that freely contact other projection neurons, regardless of pathway (Yung, Smith, Levey, & Bolam, 1996). There is, at most, about a one-third probability of one neuron sending an inhibitory connection to another (Taverna, Illijic, & Surmeier, 2008). As a result, these lateral inhibitory connections are usually one-way (at best only one in nine connections are reciprocated in a randomly connected network). Although the lateral inhibitory connections are believed to be weak, fast spiking inhibitory interneurons provide additional inhibition (Gruber, Powell, & O'Donnell, 2009; Tepper, Wilson, & Koos, 2008). Therefore, the lateral inhibitory component of this model could roughly be viewed as representing the contributions of both types of inhibition. Neurobiology concerning the lateral learning rule, however, is less clear than evidence concerning its input weight counterpart. Long-term learning has been found to occur in the lateral synaptic connections of striatal projection neurons (Rueda-Orozco et al., 2009), although how this relates to a dopamine-based error term is not clear. In the model, the learning rule for these lateral connections is similar to that of the input connections, except that the CS intensity, I, is replaced by a term related to the activity of one of its laterally inhibiting neurons.

Conclusions

The present model provides an elemental associative explanation for key retrospective revaluation phenomena apart from within-compound associations, based on novel mechanisms that correlate to some degree with neurobiology. The present model only allows stimuli previously conditioned in compound to revalue one another, rather than permitting the conditioning of a CS to substantially revalue a separately conditioned CS. This model thus not only matches experimental data that speak to this issue, but also agrees with the natural intuition that independently conditioned stimuli should have little effect on one another when one is revalued.

Author Note This work was financially supported by CIHR, NSERC, the Walter C. Sumner Foundation, and Dalhousie University.

References

Aitken, M., & Dickinson, A. (2005). Simulations of a modified SOP model applied to retrospective revaluation of human causal learning. *Learning & Behavior, 33*(2), 147–159.

Amundson, J., Escobar, M., & Miller, R. (2003). Proactive interference between cues trained with a common outcome in first-order pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 29*(4), 311.

Blaisdell, A. P., Gunther, L. M., & Miller, R. R. (1999). Recovery from blocking achieved by extinguishing the blocking CS. *Animal Learning Behavior, 27*(1), 63–76.

Brandon, S. E., & Wagner, A. R. (1998). Occasion setting: Influences of conditioned emotional responses and configural cues. In N. A. Schmajuk & P. C. Holland (Eds.), *Occasion setting: Associative learning and cognition in animals* (pp. 343–382). Washington, DC: American Psychological Association.

Chapman, G. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 837.

Cole, R., Barnet, R., & Miller, R. (1995). Effect of relative stimulus validity: Learning or performance deficit? *Journal of Experimental Psychology: Animal Behavior Processes, 21*(4), 293.

Cole, R., Denniston, J., & Miller, R. (1996). Reminder-induced attenuation of the effect of relative stimulus validity. *Learning & Behavior, 24*(3), 256–265.

Connor, P., & Trappenberg, T. (2011). Characterizing a brain-based value-function approximator. In C. Butz & P. Lingras (Eds.), *Advances in artificial intelligence* (Vol. 6657, pp. 92–103). Springer Berlin Heidelberg.

Dawson, M. (2008). Connectionism and classical conditioning. *Comparative Cognition and Behavior Reviews, 3,* 1–115.

Delamater, A. R., Sosa, W., & Katz, M. (1999). Elemental and configural processes in patterning discrimination learning. *The Quarterly Journal of Experimental Psychology: Section B, 52*(2), 97–124.

Denniston, J., Miller, R., & Matute, H. (1996). Biological significance as a determinant of cue competition. *Psychological Science, 7*(6), 325–331.

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 65–117). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology Section B, 49*(1), 60–80.

Dopson, J., Pearce, J., & Haselgrove, M. (2009). Failure of retrospective revaluation to influence blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 35*(4), 473.

Escobar, M., Pineño, O., & Matute, H. (2002). A comparison between elemental and compound training of cues in retrospective revaluation. *Learning & Behavior, 30*(3), 228–238.

Espinet, A., Iraola, J., Bennett, C., & Mackintosh, N. (1995). Inhibitory associations between neutral stimuli in flavor-aversion conditioning. *Animal Learning & Behavior, 23*(4), 361–368.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience, 17*(1), 51–72.

Frank, M. J., & Fossella, J. A. (2011). Neurogenetics and pharmacology of learning, motivation, and cognition. *Neuropsychopharmacology: Official publication of the American College of Neuropsychopharmacology, 36*(1), 133–152.

Ghirlanda, S. (2005). Retrospective revaluation as simple associative learning. *Journal of Experimental Psychology. Animal Behavior Processes, 31*(1), 107–111.

Gruber, A. J., Powell, E. M., & O'Donnell, P. (2009). Cortically activated interneurons shape spatial aspects of cortico-accumbens processing. *Journal of Neurophysiology, 101*(4), 1876–1882.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review, 113*(3), 584–605.

Harris, J. A., Gharaei, S., & Moore, C. A. (2009). Representations of single and compound stimuli in negative and positive patterning. *Learning & Behavior, 37*(3), 230–245.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of information processing in the basal ganglia* (pp. 249–270). MIT Press.

Houwer, J. D., & Beckers, T. (2002). Higher-order retrospective revaluation in human causal learning. *Quarterly Journal of Experimental Psychology: Section B, 55*(2), 137–151.

Jamieson, R., Crump, M., & Hannah, S. (2012). An instance theory of associative learning. *Learning & Behavior, 40*(1), 61–82.

Kasprow, W., Cacheiro, H., Balaz, M., & Miller, R. (1982). Reminder-induced recovery of associations to an overshadowed stimulus. *Learning and Motivation, 13*(2), 155–166.

Kasprow, W., Catterson, D., Schachtman, T., & Miller, R. (1984). Attenuation of latent inhibition by post-acquisition reminder. *The Quarterly Journal of Experimental Psychology, 36*(1), 53–63.

Kasprow, W., Schachtman, T. R., & Miller, R. R. (1987). The comparator hypothesis of conditioned response generation: Manifest conditioned excitation and inhibition as a function of relative excitatory strengths of CS and conditioning context at the time of testing. *Journal of Experimental Psychology. Animal Behavior Processes, 13*(4), 395–406.

Kaufman, M. A., & Bolles, R. C. (1981). A nonassociative aspect of overshadowing. *Bulletin of the Psychonomic Society, 18*(6), 318–320.

Kraemer, P., Lariviere, N., & Spear, N. (1988). Expression of a taste aversion conditioned with an odor-taste compound: Overshadowing is relatively weak in weanlings and decreases over a retention interval in adults. *Learning & Behavior, 16*(2), 164–168.

Kutlu, M., & Schmajuk, N. (2012). Solving pavlov's puzzle: Attentional, associative, and flexible configural mechanisms in classical conditioning. *Learning & Behavior, 40*(3), 269–291.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology Section B, 57*(3), 193–243.

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *The Quarterly Journal of Experimental Psychology Section B, 54*(4), 311–352.

Liljeholm, M., & Balleine, B. W. (2006). Stimulus salience and retrospective revaluation. *Journal of Experimental Psychology: Animal Behavior Processes, 32*(4), 481–487.

Liljeholm, M., & Balleine, B. W. (2009). Mediated conditioning versus retrospective revaluation in humans: The influence of physical and functional similarity of cues. *The Quarterly Journal of Experimental Psychology, 62*(3), 470–482.

Lysle, D. T., & Fowler, H. (1985). Inhibition as a "slave" process: Deactivation of conditioned inhibition through extinction of conditioned excitation. *Journal of Experimental Psychology. Animal Behavior Processes, 11*(1), 71–94.

Matzel, L. D., Schachtman, T., & Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning and Motivation, 16,* 398–412.

McLaren, I., Forrest, C., & McLaren, R. (2012). Elemental representation and configural mappings: Combining elemental and configural theories of associative learning. *Learning & Behavior, 40*(3), 320–333.

McLaren, I. P. L. (1993). APECS: A solution to the sequential learning problem. In *Proceedings of the xvth annual convention of the cognitive science society* (pp. 717–722).

McLaren, I. P. L. (2011). APECS: An adaptively parameterised model of associative learning and memory. In E. Alonso & E. Mondragon (Eds.), *Computational neuroscience for advancing artificial intelligence: Models, methods and applications* (pp. 145–164). Hershey: IGI Global.

Melchers, K., Lachnit, H., & Shanks, D. (2004). Within-compound associations in retrospective revaluation and in direct learning: A challenge for comparator theory. *Quarterly Journal of Experimental Psychology Section B, 57*(1), 25–54.

Miller, R. R., Barnet, R., & Grahame, N. (1992). Responding to a conditioned stimulus depends on the current associative status of other cues present during training of that specific stimulus. *Journal of Experimental Psychology: Animal Behavior Processes; Journal of Experimental Psychology: Animal Behavior Processes, 18*(3), 251.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. *Psychology of Learning and Motivation, 22,* 51–92.

Pan, W., Schmidt, R., Wickens, J., & Hyland, B. (2008). Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *The Journal of Neuroscience, 28*(39), 9619–9631.

Pineño, O., Urushihara, K., & Miller, R. R. (2005). Spontaneous recovery from forward and backward blocking. *Journal of experimental psychology Animal behavior processes, 31*(2), 172–183.

Plenz, D. (2003). When inhibition goes incognito: Feedback interaction between spiny projection neurons in striatal function. *Trends in Neurosciences, 26*(8), 436–443.

Redhead, E. S., & Pearce, J. M. (1995). Stimulus salience and negative patterning. *The Quarterly Journal of Experimental Psychology, 48*(1), 67–83.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Reynolds, J., & Wickens, J. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks, 15*(4–6), 507–521.

Rueda-Orozco, P. E., Mendoza, E., Hernandez, R., Aceves, J. J., Ibanez-Sandoval, O., Galarraga, E., Connor, P. C., LoLordo, V. M., & Trappenberg, T. P. (2009). Diversity in long-term synaptic plasticity at inhibitory synapses of striatal spiny neurons. *Learning & Memory, 16*(8), 474–478.

Schachtman, T., Gee, J., Kasprow, W., & Miller, R. (1983). Reminder-induced recovery from blocking as a function of the number of compound trials. *Learning and Motivation, 14*(2), 154–164.

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

Shanks, D. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B, 37*(1), 1–21.

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science, 321*(5890), 848–851.

Shevill, I., & Hall, G. (2004). Retrospective revaluation effects in the conditioned suppression procedure. *Quarterly Journal of Experimental Psychology Section B, 57*(4), 331–347.

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review, 114*(3), 759.

Taverna, S., Ilijic, E., & Surmeier, D. J. (2008). Recurrent collateral connections of striatal medium spiny neurons are disrupted in models of Parkinson's disease. *Journal of Neuroscience, 28*(21), 5504–5512.

Tepper, J. M., Wilson, C. J., & Koos, T. (2008). Feedforward and feedback inhibition in neostriatal GABAergic spiny neurons. *Brain Research Reviews, 58*(2), 272–281.

Tunstall, M. J., Oorschot, D. E., Kean, A., & Wickens, J. R. (2002). Inhibitory interactions between spiny projection neurons in the rat striatum. *Journal of Neurophysiology, 88*(3), 1263–1269.

Urcelay, G., Perelmuter, O., & Miller, R. (2008). Pavlovian backward conditioned inhibition in humans: Summation and retardation tests. *Behavioural Processes, 77*(3), 299–305.

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550.

Van Hamme, L., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*(2), 127–151.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spears & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5–47). Hillsdale, NJ: Earlbaum.

Wagner, A. R. (2003). Context-sensitive elemental theory. *The Quarterly Journal of Experimental Psychology: Section B, 56*(1), 7–29.

Wagner, A. R., & Brandon, S. E. (2001). A componential theory of Pavlovian conditioning. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 23–64). Mahwah, NJ: Earlbaum.

Wasserman, E., & Berglan, L. (1998). Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. *The Quarterly Journal of Experimental Psychology Section B, 51*(2), 121–138.

Witnauer, J. E., & Miller, R. R. (2011). The role of within-compound associations in learning about absent cues. *Learning & Behavior, 39*(2), 146–162.

Woodbury, C. B. (1943). The learning of stimulus patterns by dogs. *Journal of Comparative Psychology, 35,* 29–40.

Yung, K. K., Smith, A. D., Levey, A. I., & Bolam, J. P. (1996). Synaptic connections between spiny neurons of the direct and indirect pathways in the neostriatum of the rat: Evidence from dopamine receptor and neuropeptide immunostaining. *The European Journal of Neuroscience, 8*(5), 861–869.