Published in: *Learning & Behavior 2010, 38 (1), 1-26 doi:10.3758/LB.38.1.1*

An Attention-Modulated Associative Network

Justin A. Harris & Evan J. Livesey

The University of Sydney

Abstract

We present an elemental model of associative learning that describes interactions between stimulus elements as a process of competitive normalization. Building on the assumptions laid out in Harris (2006), stimuli are represented as an array of elements that compete for attention according to the strength of their input. Elements form associations amongst each other according to temporal correlations in their activation but restricted by their connectivity. The model moves beyond its predecessor by specifying excitatory, inhibitory, and attention processes for each element in real-time, and describing their interaction as a form of suppressive gain control. Attention is formalized in this model as a network of mutually inhibitory units which moderate the activation of stimulus elements by controlling the level to which the elements are suppressed by their own inhibitory processes. The model is applied to a range of complex discriminations and related phenomena that have been taken as evidence for configural learning processes.

A century of experimental study of associative learning has inspired the development of several very successful models that describe learning as a computational process. The most popular form of these models identifies the content of associative learning with the strengthening or weakening of a link that connects some representation of the events that are temporally correlated. In this way, the learned response arises by a process of stimulus substitution within a network. For example, a positive temporal correlation between the presentation of a conditioned stimulus (CS) and occurrence of an unconditioned stimulus (US) leads to the strengthening of connections between those components of the network that are identified with perceptual or behavioral responses to each stimulus. As a consequence of this strengthening, the CS is able to activate those components corresponding to the US and thus elicit some subset of the responses that were previously only elicited by the US.

While much has been made of computational models that describe the manner in which connection strength changes during associative learning, the accuracy with which these models map onto behavior depends on the nature of the components that they connect. One fundamental issue that has been extensively debated concerns the distributed versus unitary nature of the way that stimuli are represented within the associative network. On one side of this distinction are theories that assume individual stimuli to be represented by multiple elements distributed across the network; on the other side are theories in which whole stimulus patterns are represented by a single configural unit. Adherents to strong versions of each view can be found in the literature, as well as examples of a “hybrid” approach which combines configural and elemental representations in the same associative network. This review will begin by describing the primary examples of the different theoretical approaches, as well as discussing empirical evidence held to support the existence of configural representations. We will then move on to the primary objective of this paper, to present an

Address for correspondence:
School of Psychology
The University of Sydney
Sydney, 2006
Australia
Email: justinh@psych.usyd.edu.au ; evanl@psych.usyd.edu.au

extension of an elemental model recently proposed by Harris (2006).

Elemental and configural models

The notion that stimuli have distributed representations goes back more than half a century. Some of the earliest models of associative networks considered stimuli to be comprised of multiple elements, and associative learning involved the strengthening of connections between the elements of the CS and those of the US (Atkinson & Estes, 1963; Bush & Mosteller, 1951; Estes, 1950). One virtue of these models is that they provide a clear description of generalization and discrimination, based on the proportion of elements shared in common between stimuli. For example, conditioning of stimulus X involves the strengthening of associations between the set of X's elements and those of the US, and therefore conditioned responding to X will generalize to Y as a function of the number of X's elements that are common to Y. The explanatory power of this simple approach was greatly enhanced when it was combined with a learning rule that used a common error term to compute changes in associative strength (Rescorla & Wagner, 1972). An important feature of this learning rule is that it provides a mechanism for the development of inhibition whenever the summed associative strength (ΔV) of all elements active on a trial exceeds the asymptotic associative strength (λ) supported by the US on that trial. This enables elemental models to anticipate errorless performance on even difficult discriminations. For example, in a discrimination in which one stimulus, X, is followed by the US and a second stimulus, Y, is not, generalization from X to Y will be large if X and Y share many elements in common (i.e., they are similar). However, across continued training associative strength will progressively shift away from the common elements towards the elements unique to X, reducing generalization to Y. At the same time Y's unique elements will develop inhibitory associative strength that will cancel whatever excitatory strength generalizes via the common elements.

There are, however, discriminations that these elemental models remain unable to solve, in particular conditional discriminations in which the presence (+) versus absence (-) of the US is not correlated with any identifiable stimulus. The best known of these is negative patterning, in which two CSs, A and B, are consistently reinforced with the US when presented individually but are never reinforced when presented together as a compound. There are many demonstrations that animals from numerous species can solve this discrimination, in that they learn to respond less on AB- trials than on A+ and B+ trials (Bellingham, Gillette-Bellingham, & Kehoe, 1985; Harris & Livesey, 2008; Harris, Livesey, Gharaei, & Westbrook, 2008; Kehoe & Graham, 1988; Pavlov, 1927; Rescorla, 1972, 1973; Whitlow & Wagner, 1972; Woodbury, 1943). However this discrimination poses an insoluble challenge to simple elemental models because every CS element that is present on AB- trials is also present on A+ or B+ trials, and every element present on A+ or B+ trials is present on AB- trials. Thus, because there is no single element that is present on reinforced but not non-reinforced trials, or vice versa, there is no mechanism to prevent generalization of learning from A+ and B+ trials to AB- trials, or to cancel that generalization with inhibition on AB- trials. An equally challenging problem for simple elemental models is posed by biconditional discriminations in which four stimuli are presented as four different pairwise compounds, two of which are reinforced and two not. The discrimination is insoluble by simple elemental models if each single stimulus is reinforced in one compound and not in the other (e.g., AB+ and CD+ versus AC- and BD-), ensuring that no single element is correlated (positively or negatively) with the US. Despite this, there are numerous demonstrations that animals can solve these discriminations (Harris et al., 2008; Rescorla, Grau, & Durlach, 1985; Saavedra, 1975), albeit with considerable difficulty.

Evidence that animals can learn the aforementioned conditional discriminations is often held up as proof that associative learning operates on more than simple elemental units. Successful performance on these discriminations is taken as evidence that learning mechanisms must represent the conjunction of stimuli (e.g., that A and B occur together on AB trials) and this conjunctive representation can itself serve as the argument of an association. This notion of configural representations itself has a long history. Since Spence (1952), many have assumed that, within learning systems, stimuli are represented by their elemental components plus a representation of the configural nature of the way those elements are combined. Thus, on AB trials, the conjunction of A and B is represented as a separate unit in addition to the composite elements of A and B. When incorporated with the Rescorla-Wagner (1972) learning rule, training on a negative patterning discrimination will imbue the AB configural unit with inhibitory strength because it is reliably present on non-reinforced trials but never on reinforced trials. This allows the configural unit to cancel much of the excitatory associative strength that generalizes from what has been learned on A+ and B+ trials (Rescorla, 1972; Whitlow & Wagner, 1972). By similar logic, during training on a biconditional discrimination, configural units for AB and CD compounds will acquire excitatory associative strength with the US, while separate configural units for AC and BD compounds will acquire inhibitory strength (Saavedra, 1975).

The approach described above assumes that a compound of two or more stimuli is represented by the sum of elements of each stimulus plus a configural representation of their conjunction. In this sense, it combines a configural description of stimulus representation with an elemental one. The explanatory power of this approach has been further enhanced by recent theoretical extensions proposed by Wagner and Brandon (2001). According to their "Replaced Elements" model, configural units that are activated when stimuli are presented as a compound are not simply *added* to the array of elements representing each component stimulus, but rather they *replace* specific elements of the individual stimuli (see also Wagner, 2003, 2008). An associative network of this form not only has units that are uniquely present on trials with compound stimuli, but has units that are uniquely present on single-stimulus trials, so doubling the network's opportunity to learn to discriminate between a compound and its components. This innovation is also important in allowing elemental models with configural elements to account for a variety of data that are otherwise troubling for the traditional "added" configural element approach. In large part, the advantage of this replaced elements approach is that it reduces the amount of associative strength that generalizes from individual CSs to their compound.

Notwithstanding the popularity and success of the "hybrid" elemental-configural approach described so far, there are alternative models that can solve complex problems like negative patterning and the biconditional discrimination within a purely configural or purely elemental framework. A prime example of a pure configural model is that proposed by Pearce (1987, 1994, 2002). In striking contrast to elemental models, Pearce's model assumes that each conditioning episode involves strengthening or weakening of just one association that connects a single configural unit representing the entire pattern of CS inputs (including the context) with the US. Negative patterning and biconditional discriminations can be solved because responding on each trial type is governed by different associations arising from distinct configural units. Indeed, such a configural model is at risk of under-estimating the difficulty posed by these discriminations. To take account of this Pearce has proposed that each stimulus pattern fully activates its own configural unit but also partially activates configural units corresponding to other stimulus patterns, and the degree of this generalized activation depends on the similarity (featural overlap) between the current stimulus pattern and the pattern coded by each configural unit. For example, in negative patterning, trials with the AB compound will fully activate the AB configural unit but will also partially activate the A and B configural units (due to the overlap in their input patterns), and trials with A or B individually will fully activate the A or B configural unit but also partially activate the AB configural unit. While learning (the change in associative strength) is confined to the fully activated unit on any trial, the partial activation of other units will produce responding that will interfere with correct performance on the discrimination. Thus an animal trained on a negative patterning schedule will respond on AB- trials due to generalized activation of the A and B units, but will eventually stop responding on those trials when the AB configural unit develops sufficient inhibitory strength to cancel any excitatory strength arising from the generalized activation of A and B units.

Configural units that contain information about the specific conjunctions of stimuli have proved popular tools in enabling models of associative learning to solve conditional discriminations of the sort described thus far. However, the configural units themselves bear some explanatory burden. A configural unit specifically codes for the conjunction of stimuli, and is therefore storing a form of associative information that would normally be the purview of the associative process itself. But configural theories do not use associative mechanisms to acquire or store the associative information that is contained within a configural representation. Thus, these models are left needing to specify what mechanisms are responsible for coding and storing the associative information that is represented by configural units. The need for an adequate specification of these mechanisms is particularly apparent for models that combine elemental and configural representations, since otherwise such models are at risk of a combinatorial explosion created by the potential that configural units exist for all possible combinations of individual elements.

There are alternatives to invoking configural solutions to complex conditional discriminations. The key to solving the discrimination is to provide a means by which some of what is learned on reinforced trials does not generalize to non-reinforced trials, or vice versa. Configural models achieve this by ensuring that the associative change that accrues to each configural unit does not generalize effectively to trials with the single CSs or with different compounds. Two recent models have been developed that show how the reverse asymmetry can be achieved within a purely elemental associative network (Harris, 2006; McLaren & Mackintosh, 2002). Both models represent stimuli within a distributed network of elements, following the principles developed in Stimulus Sampling Theory (Atkinson & Estes, 1963; Bush & Mosteller, 1951; Estes, 1950). Two key features enable the McLaren-Mackintosh (2002) model to solve complex conditional discriminations. The first is the assumption that all stimuli have broadly distributed representations, in that any stimulus activates half of the total number of elements in the network. This ensures that any two stimuli, even if from very different modalities, will share a substantial proportion of their elements in common. The second feature of the model concerns the activation of the elements. McLaren and Mackintosh assume that the activation function is steeply sigmoid to the point of approximating a step function. Thus, even though sensory input to the elements is graded along a continuum, the state of the elements themselves is to a large extent binary: they are either fully activated when the input exceeds a threshold, or else inactive. It is the non-linearity of the activation function, particularly when applied to the common elements, that provides the means by which the McLaren-Mackintosh (2002) model solves negative patterning and biconditional discriminations.

The second purely elemental model to be discussed here was developed by Harris (2006). Like the McLaren-Mackintosh (2002) model, Harris's model relies on non-linearity in the activation of elements, but unlike McLaren-Mackintosh it does not rely on common elements to solve complex conditional discriminations. Rather, it assumes that non-linearity in element activation is introduced by the operation of a limited-capacity attention buffer that boosts element activation. According to the model, elements compete for attention based on the rate of rise in their activation (before the attention boost). More strongly activated elements can enter the attention buffer and thus receive a further boost to their activation strength, whereas weaker elements do not. Because access to attention is strictly competitive, individual elements can receive the boost to their activation in some instances (when pitted against weaker or fewer elements) but not under other circumstances (when pitted against stronger or more elements). For example, in negative patterning, some elements of A and B will enter the attention buffer on trials with A or B alone, but will be unable to enter the buffer on AB compound trials because the larger number of elements will increase the competition for access to attention. Thus these elements provide a solution to negative patterning because their activation is positively correlated with reinforcement – they are strongly active (by virtue of receiving an attention boost) on A+ and B+ trials, but are weakly active (being denied the attention boost) on AB- trials. In a similar way, elements can solve the biconditional discrimination due to incidental differences in the competition for access to attention between different compounds. For example, if stimulus B has a greater number of strong elements than stimulus C, this will mean that A's elements face stiffer competition for attention in the AB compound than in the AC compound, and thus some of A's elements will receive the attention boost on AC trials but not AB trials. In this way, activation of these A elements is correlated (negatively) with reinforcement, and their acquisition of inhibitory associative strength will contribute to solving the discrimination. It is important to note that counterbalancing the identities of the different stimuli used in this discrimination will equate stimulus properties at the group level, but incidental differences between stimuli at the individual level will remain, and it is at this level that the discrimination is solved.

Comparison of the different configural and elemental models described above has inspired many experimental investigations aimed at discerning which theoretical approach provides a more accurate picture of learning in animals. However, before discussing those investigations, we first focus our attention on the last of the models described above. The next section presents a detailed extension of the elemental model proposed by Harris (2006). The need for this is twofold. First, in the original model, the non-linear influence of attention on element activity arose by virtue of attention's limited capacity, which meant that attention would selectively boost activation of some elements but not others. The capacity of attention was defined in simple numerical terms – it could only "hold" a fixed amount of element activity, rather like a bucket able to hold a fixed volume of rocks of varying sizes. While this is an intuitively useful and computationally tractable conceptualization of attention, it is difficult to see how a mechanism of this sort could be meaningfully operationalized within a neural network. Therefore, in the section that follows, we specify in greater detail the attention system itself, refining the notion of a limited capacity buffer by framing attention as a network in which gain control normalizes overall activity (thus effecting a form of capacity limitation).

The second aspect of the model we develop here is its temporal resolution. In the previous version of the model (Harris, 2006), each operation was defined at the level of the individual trial. However, this trial-based operation doesn't capture all of the dynamic potential of the model's behavior – the way that stimuli interact through the acquisition of excitatory and inhibitory connections between their elements, and the way that competition for attention evolves over time within a trial. Accordingly, another objective here is to provide a more continuous description of the model, simulating its operations in "real time". This is important for any model to account for the influence of temporal variables such as CS-US contiguity and inter-trial intervals. But, as just noted, it is particularly important for the present model in order to adequately capture the dynamics of inter-element interactions. As we describe below, this is achieved by iterating each operation over many discrete moments within the trial, effectively dividing the trial into many mini-trials; an approach that has proved successful in numerous other "real time" elemental models (e.g., McLaren & Mackintosh, 2002; Sutton & Barto, 1981; Wagner, 1981).

The current model: An Attention-Modulated Associative Network

The current model retains a number of characteristics of the earlier proposal by Harris (2006). Like the previous model, it considers stimuli to have distributed representations within the associative network in that each stimulus excites a population of elements. Each element has a fixed probability of being connected to every other element, such that each element is connected to a subset of all elements, and associative learning depends on changes in the strength of those connections. Finally, the activation strength of an element is boosted by attention. Access to attention is subject to competition between elements, and the size of the boost that an element receives is not linear (it is largely all-or-none).

The current proposal advances the model in two important ways:

> First, whereas the previous model was trial-based, the current proposal describes the behavior of the network "in real time". That is, the model's operations are specified as differential equations describing continuous changes across time, and each operation can be implemented at a temporal resolution that can be defined within a single trial.
>
> Second, gain control plays a critical role in regulating (normalizing) the behavior of the network as a whole. In general terms, this means that the response of each element in the network to incoming stimulation is attenuated by the activity of other elements in the network. All units in the network are subject to at least one source of gain control that divisively normalizes their activation. Gain control is responsible for both the competitive aspect of the interactions between elements in the network and the non-linear nature of each element's activation (see also Harris, in press). Some of the functional consequences of the gain control process are similar to the replacement process that operates in the Replaced Elements model (Wagner, 2003; Wagner & Brandon, 2001), especially because the

effects of both gain control and replacement increase between stimuli that are perceptually similar. It should also be noted that the normalization process described here makes the function relating input strength to element activation strength sigmoid, a property of the element activation function used in the McLaren-Mackintosh (2002) model.

Network structure

Figure 1 illustrates the structure of the network proposed here. Elements (E) are activated by sensory input (S) to varying degrees depending on their tuning to the spatial and featural properties of the sensory environment. In this sense, each element can be said to have spatial and featural receptive fields because they respond preferentially to features of a particular quality (e.g. orientation or color) and in a given spatial location. For the purpose of modeling the representations of simple stimuli of the kind used in Pavlovian conditioning experiments (lights, tones, etc.) it is assumed that a given physical stimulus provides S input to a collection of E elements with similar spatial and/or featural receptive fields. The strength of activation of a given E depends on S and on "internal" input from other E elements to which it is connected. Learning depends on changes to the strength of these connections between Es. The strength of these connections is initially zero, but increases or decreases as a consequence of temporal correlations between elements' activations.

We assume that connectivity between elements is not uniform across the entire network, in that a given E is not connected equally to every other E. To introduce variability in the network in our modeling, we have assumed that the existence of a connection between any two E elements is probabilistic. In our modeling, we have set the likelihood of E_i being connected to E_j at 0.5 (consistent with Harris, 2006), which means that any E is on average connected to 50% of the rest of the E network. We have chosen this binary form of connection variability because it is simple and intuitive. However, an alternative would be to vary the "plasticity" of each connection, such that every E is connected to every other E but the plasticity of each connection (in Equation 6, below) would vary (when = 0, the connection would be effectively absent). As explained by Harris (2006), the assumption of variability in the network's connectivity provides an effective means to explain a variety of phenomena, such as evidence that some CS-US associations are acquired more quickly than others. For instance, the fact that animals learn a taste-illness association much faster than a noise-illness association (Garcia & Koelling, 1966) could be explained as extensive connectivity between gustatory and gastrointestinal elements and relatively poor connectivity between auditory and gastrointestinal elements. Variations in connectivity can also explain Rescorla's (2000, 2001, 2002a, 2002b) demonstrations that the rate of learning about two CSs differs, even when they are conditioned together in compound, if their initial associative strengths differ (see Harris, 2006).

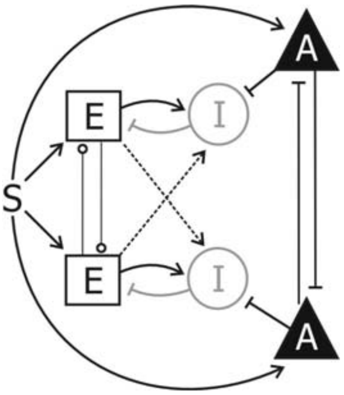

Figure 1. Illustration of the basic structure of the associative network proposed here. Sensory input (S) activates a set of elements (E; just two elements are shown here for clarity of exposition). Learning depends on changes in the strength of connections between the E elements, which by default have an initial strength of zero. Each E activates a paired inhibitory unit (I) and weakly activates the inhibitory units of other elements that have similar spatial and featural receptive fields. The inhibitory units reduce activation of their paired E element. Thus I units exert gain control, normalizing the activation of each E depending on the extent of activation of surrounding Es. The activity of each I unit, and thus of each E element, is regulated by a network of attention units (A). Each A unit is excited by the same sensory input that activates the corresponding E element, and that A unit inhibits the corresponding I unit, thus increasing activation of the E element (releasing E from inhibition). Finally, inhibition among A units normalizes their activity.

Each E element is paired with one inhibitory unit (I) and one attention unit (A). The activation of each E is normalized by its paired I; each I is excited by its paired E (and to a lesser extent by other Es with similar receptive properties). Thus I units exert suppressive gain control of E elements, normalizing activity within the network. The attention "buffer" of the previous model (Harris, 2006) is replaced by a network of attention units (A) which act to inhibit I units and thereby diminish their suppressive effect on E elements. Thus attention increases activation of a given E element by releasing it from inhibition by I. Activity in the attention units is also driven by sensory input: each A unit receives the same S as its paired E element. Finally, the activation of each A unit is normalized by activity in other A units in the attention network. Thus the ability of a given sensory input to excite its attention unit is diminished by other sensory inputs that compete for attention.

Computations

Above we have described the basic structure of the proposed network. What follows is a detailed description of the computational operations performed by each unit at each moment. We present a series of equations that determine the activation strength, *E*, *I*, and *A* of each element, inhibitory unit and attention unit, based on "external" (sensory) input and "internal" excitatory and inhibitory inputs from other units in the network. (Note, we use italics to refer to activation strength of each E, I, and A unit.) The general form of these equations derives from computational rules that have been used in numerous existing models of sensory systems that incorporate a gain control process of normalization (e.g., Grossberg, 1973; Heeger, 1992; Reynolds & Chelazzi, 2004; Reynolds & Heeger, 2009), and the application of this gain control mechanism to element activation has been discussed recently by Harris (in press).

At any given moment, each E, I, and A unit has a response potential (R_{pot}) which drives changes in that unit's activation (its actual response *R*). Changes in R_{pot} are assumed to be instantaneous, whereas changes in *R* are more gradual. It is convenient to define R_{pot} and *R* separately (and make a clear distinction between them) because R_{pot} reflects the level of activity that a unit would eventually reach if all other variables in the system remained exactly the same, whereas R reflects the actual activity of the unit at a given instant. R thus approaches R_{pot}. The general form we use to calculate R_{pot} is shown below in Equation 1.

$$R_{\text{pot}} = \frac{\mathrm{Input}^{\mathrm{p}}}{\mathrm{Input}^{\mathrm{p}} + \mathrm{N}^{\mathrm{p}} + \mathrm{D}} \tag{1}$$

In Equation 1, the response potential (R_{pot}) of a unit is given by the sum of its inputs (from S and E elements), divided by those same inputs plus normalizing inputs (N) and a constant D. For E elements, the normalizing inputs come from the I units, for I units the normalizing inputs come from A units, and for A units the normalizing inputs come from other A units (as illustrated in Figure 1, and defined in Equations 3, 4, and 5 below).

Equation 1 is a monotonically increasing function that asymptotes at 1. If N is zero (i.e., there is no normalizing influence on the unit's activity), R_{pot} reaches half height (R_{pot} = 0.5) when $\mathrm{Input}^{\mathrm{p}}$ = D. As N increases, the function is effectively shifted to the right such that the strength of $\mathrm{Input}^{\mathrm{p}}$ must increase by N^{p} in order for R_{pot} to reach half height. The constant D scales the range of Input values over which critical changes in R_{pot} occur and prevents the denominator of the equation from equaling zero. In all simulations presented here, D = 0.04 in all equations. The power, p, determines the slope of the function. As illustrated in Figure 2, when p = 1, the function is a simple monotonically increasing curve. Higher values of p make the function sigmoid, and increasing p increases the maximal slope.

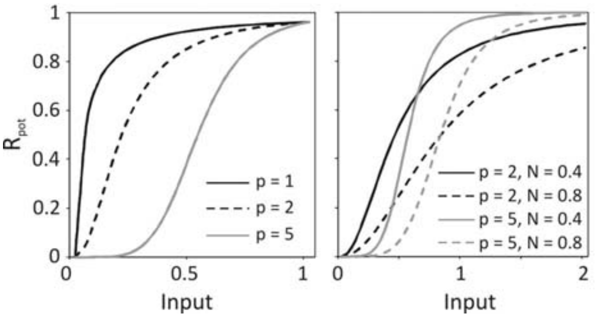

Figure 2. The relationship between input and response potential (R_{pot}) according to Equation 1 (with D = 0.04). The panel on the left show that, when p = 1, the response is a monotonically increasing but continuously decelerating function of input that approaches an asymptote of 1. When p > 1, the function becomes sigmoid, but still approaches an asymptote of 1. The panel on the right shows how the function is affected by the size of the normalization input (N): the function is shifted to the right as N increases from 0.4 to 0.8 (this corresponds approximately to the change in value of N between presentation of a single stimulus of 20 elements and a compound of two such stimuli).

We assume that p takes values greater than or equal to 2 based on psychophysical and neurophysiological evidence suggesting that the responses of sensory units follows a sigmoid function. The rate of change in perceived magnitude of a stimulus decreases as the absolute magnitude of the stimulus increases, as captured by the Weber-Fechner law and Stevens' power law (Stevens, 1962). However, the opposite relation has also been observed frequently for the lowest end of many stimulus dimensions. That is, for stimuli near detection threshold, observers become more sensitive in discriminating the relative magnitudes of stimuli as their absolute magnitude increases (Arabzadeh, Clifford, & Harris, 2008; Solomon, 2009). These two contrasting psychophysical effects indicate that the relationship between the physical intensity of a stimulus and its perceived intensity reverses as the stimulus intensity increases – the function is positively accelerated at low intensities and negatively accelerated at higher intensities. This sigmoid form of the function has been confirmed in numerous experiments using electrophysiological recordings in cats or primates to determine the relationship between the intensity of a stimulus (e.g., the contrast of a visual grating) and the response magnitude of neurons tuned to that stimulus (Crowder et al., 2006). In keeping with such evidence, we use a form of equation that gives a sigmoid shape to the element activation function.

In all simulations we present here, p = 2 for the activation functions of E and I units, and p = 5 for A units. This makes the behavior of A units more non-linear, in that their activation is closer to being all-or-none.

Equation 1 defines a unit's response potential to its inputs at a given instant in time. However, we assume that the unit cannot change instantaneously from its existing activation level to the new level specified by this response since this would imply an infinite rate of change. Rather, the response potential defined in Equation 1 gives the activation strength that the unit would ultimately reach if the current inputs were maintained at a constant level for an indefinite period. We assume that the real response of each unit approaches this potential in a gradual manner in real time. This is defined in Equation 2.

$$\frac{\mathrm{dR}}{\mathrm{dt}} = \delta \times \left(\mathrm{R}_{\mathrm{pot}} - \mathrm{R}\right) \qquad (2)$$

The rate of change of the activation (R) of a unit is proportional to the difference between its existing activation and the unit's response potential ($\mathrm{R}_{\mathrm{pot}}$) to the current inputs (as defined in the equations below). In all our calculations, the rate parameter, ⍰, differs when the unit's activity is rising versus falling. In our simulations, we have set ⍰ = 0.5 when ($R_{\mathrm{pot}} - R$) > 0, and ⍰ = 0.2 when ($R_{\mathrm{pot}} - R$) < 0. Thus unit activity rises more rapidly than it decays.

Equations 1 and 2 describe the general form of the function relating the response of a unit to its input. Below we give the specific equations for each E, I, and A unit.

Activation of E. Input to each element is given by the sum of its external (sensory) input (S) and internal associative input from other elements. In real time, the input to an element x combines the external input (S_x) which changes instantly, reflecting the onset and offset of external stimuli, with the current associative input to x. The associative input is the product of each element's activation and the strength (V) of its associative connection to E_{x}. The equation for calculating inputs to E is shown below (3.2). In these calculations, we have set at zero the lower limit on total summed input to any element. This constrains the activity in any unit to be non-negative. Equations 3 and 3.1 follow the form specified in Equations 1 and 2 to define the response of element E_{x} based on these inputs and its normalization by the paired I unit. Note that we have set the power, p = 2.

$$\frac{\mathrm{d}E_{\mathrm{x}}}{\mathrm{dt}} = \delta \times \left(E_{\mathrm{pot}} - E_{\mathrm{x}}\right) \qquad (3)$$

$$\text{where} \qquad E_{\mathrm{pot}} = \frac{\mathrm{Input}(\mathrm{E_x})^{\mathrm{p}}}{\mathrm{Input}(\mathrm{E_x})^{\mathrm{p}} + I_{\mathrm{x}}^{\mathrm{p}} + \mathrm{D}} \qquad (3.1)$$

$$\mathrm{Input}(\mathrm{E_x}) = \mathrm{S_x} + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{V_i} \cdot E_{\mathrm{i}} \qquad 3.2)$$

$$\text{if} \qquad \left[\mathrm{S_x} + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{V_i} \cdot E_{\mathrm{i}}\right] < 0, \quad \mathrm{Input}(\mathrm{E_x}) = 0$$

Activation of I. The activation of I units is determined by a function of the general form of Equations 1 and 2. The input to each I unit is a weighted sum of activities in all E elements, as shown below in Equation 4.2, and that input is normalized by inhibition from attention units (A), as shown in Equation 4.1. The inputs to I units from E elements can be thought of as a network of connections, with each connection being weighted (z) in the range from 0 to 1. For the input to a given inhibitory unit, I_x, the weighting ($z_{i,x}$) applied to each E_i is determined by the similarity of E_i's spatial and featural receptive field with the receptive field of the unit, E_x, undergoing normalization by I_x. Smaller values of $z_{i,x}$ reflect less similarity between E_i and E_x, and therefore E_x receives stronger suppressive normalization from those E_i with more similar sensory tuning. Thus, $z_{x,x} = 1$ (such that E_x provides the largest input to I_x) and every other $z_{i,x}$ is less than 1. For any E_i that has no overlap in receptive field with E_x (e.g., an element that responds to stimulation in a different sensory modality) $z_{i,x} = 0$, thus providing no normalizing influence.

$$\frac{\mathrm{d}I_x}{\mathrm{dt}} = \delta \cdot (I_{\mathrm{pot}} - I_x) \qquad (4)$$

$$\text{where} \quad I_{\mathrm{pot}} = \frac{\mathrm{Input}(\mathrm{I_x})^{\mathrm{p}}}{\mathrm{Input}(\mathrm{I_x})^{\mathrm{p}} + (\mathrm{k_a} \cdot A_x)^{\mathrm{p}} + \mathrm{D}} \qquad (4.1)$$

$$\mathrm{Input}(\mathrm{I_x}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{z_{i,x}} \cdot E_{\mathrm{i}} \qquad (4.2)$$

The input to each I unit is normalized by input from a specific A unit. This input is scaled by a factor, k_a, which allows the attention units to strongly suppress I units. In the simulations presented here, $k_a = 4$.

Activation of A. Activity in the attention units is computed using Equation 5 which, as for E and I units, takes the general form of Equations 1 and 2. Equation 5.1 defines the normalized response potential of A. Unit A_x receives the same sensory input as E_x (i.e., S_x). This sensory input is normalized by activity in every *other* A unit. Equation 5.2 defines how this normalizing input, *A*', is calculated from the sum of scaled activations across all A units except A_x. Thus the attention field functions as a competitive network of fully connected units, where all connections are suppressive and have the same fixed strength (w). In order for the attention field to act appropriately, w should be inversely proportional to the number of elements. In the simulations we present here, w was set to 0.04.

$$\frac{\mathrm{d}A_x}{\mathrm{dt}} = \delta \cdot (A_{\mathrm{pot}} - A_x) \qquad (5)$$

$$\text{where} \quad A_{\mathrm{pot}} = \frac{\mathrm{S_x^p}}{\mathrm{S_x^p} + {A'_x}^{\mathrm{p}} + \mathrm{D}} \qquad (5.1)$$

$$A'_x = \mathrm{w} \cdot \left(\left[\sum_{i=1}^{n} A_i \right] - A_x \right) \qquad (5.2)$$

Associative change. In the previous model (Harris, 2006) the connections between elements changed strength (V) according to a summed error term (Rescorla & Wagner, 1972), reflecting the difference between external input to US elements (λ) and the sum of internal inputs to the US elements from CS elements. This requires that the associative mechanism explicitly distinguish between external and internal inputs and represent information about their difference. Such a requirement violates the local activity principle according to which associative change in the connection between two elements is determined solely by activity in those two elements (McLaren & Dickinson, 1990). However, by formalizing associative processes in real time, it is possible to specify an algorithm that functions in very similar fashion to the summed error term rule, but in which associative change is determined by the activation state of the recipient element, rather than requiring that the recipient element compare between its different sources of input. This is possible because, when operating in real time, the summed error term is generally proportional to the instantaneous change in activation of the recipient element. For example, when the external input to a US element is greater than the sum of internal inputs from CS elements, activity in the US element rises. In the present model, we use a rule for determining associative change that is similar to an idea first suggested by Konorski (1948; see also McLaren & Dickinson, 1990) and later incorporated into the real-time associative model developed by Sutton and Barto (1981). According to this rule, changes in the strength of a connection are determined by the change in activation strength of the recipient element. However, the rule used here differs from those previous proposals in one respect: Konorski and Sutton and Barto assumed that the strength of a connection increased when activity in the recipient unit rose, and decreased when activity in the recipient unit fell; the model proposed here assumes that the strength of a connection between two elements increases when the recipient element's activity rises, but the strength of the connection decreases when activity rises in the inhibitory unit (I) of the recipient element. A fall in either E or I does not produce any change in V. Thus rather than tying the direction of associative change to the direction of change in activity of the recipient element, our rule ties the direction of associative change to the rise in excitation versus inhibition of the recipient element. The rise in E and I (dE/dt and dI/dt) are calculated exactly as in Equations 3 and 4. As in other models, the change in $V_{x,y}$ is also scaled by a third parameter (α_x) which reflects the associability of the signal element and is determined by the activation of the signal element (E_x). Thus changes in associative strength from element x to element y are *proportional* to a function, $\Delta_{x,y}$, of the co-activation of elements x and y:

$$\Delta_{x,y} = \alpha_x \cdot \left(\beta_E \cdot \frac{dE_y}{dt} - \beta_I \cdot \frac{dI_y}{dt}\right) \qquad (6)$$

$$\beta_E = \begin{cases} 0.02, & \text{if } \frac{dE_y}{dt} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\beta_I = \begin{cases} 0.1, & \text{if } \frac{dI_y}{dt} > 0 \\ 0, & \text{otherwise} \end{cases}$$

The strength of a connection between E_x and E_y ($V_{x,y}$) increases as a proportion of dE_y/dt and decreases as a proportion of dI_y/dt. The amount that $V_{x,y}$ changes in response to increments in E_y and I_y is scaled by rate parameters, β_E and β_I. In testing the model we set the rate parameter for the I unit higher than for the E element (in all simulations presented here, β_E = 0.02 and β_I = 0.1). β_E = 0 if the change in E_y is negative, and likewise β_I = 0 if the change in I_y is negative, so that learning is unaffected by falls in activation. Associative changes are only driven by increases in element activation.

One further aspect of the model is the assumption that changes in α and V are gradual, reflecting internal processes that are not immediately responsive to changes in element activation. Therefore, instead of setting α_x to equal E_x, the rate of change of α_x is governed by the discrepancy between α_x and E_x (Equation 7). Likewise, rather than the rate of change of $V_{x,y}$ being governed by the rises in E_y and I_y, the rate of change of V_{xy} accelerates and decelerates according to rises in E_y and I_y (Equation 8).

$$\frac{d\alpha_x}{dt} = k_\alpha \cdot (E_x - \alpha_x) \qquad (7)$$

$$\frac{d^2V_{x,y}}{dt^2} = k_v \cdot \left(\Delta_{x,y} - \frac{dV_{x,y}}{dt}\right) \qquad (8)$$

The parameters k_α and k_v are constants set to values between 0 and 1, which govern the rate of change of associability and of learning respectively. The gradually changing weights and element associabilities allow learning to proceed in a relatively stable fashion across erratic fluctuations in activation. The gradual change in α also provides some scope for conditioning to occur across intervals between the CS and US, beyond that which is possible from residual activation of CS elements after the CS is no longer present. The simulations here use k_α = 0.33 and k_v = .05. However, for the sake of parsimony, it is worth noting that under most conditions the model does a more than adequate job with k_α and k_v both set to 1, in which case Equations 7 and 8 can be replaced with:

$$\alpha_x = E_x \qquad (7.1)$$

$$\frac{dV_{x,y}}{dt} = \Delta_{x,y} \qquad (8.1)$$

Modeling learning across a single trial

We will now give a brief explanation of the basic workings of the model before moving on to simulations of experimental data. Figure 3 shows changes in activation over single trials to illustrate the operations of the model. Here, the model consists of only one context element, one CS element, and one US element, with full connectivity between each. In each case the US onset coincides with the CS offset.

Figure 3A plots activation of E, I and A units for both the CS and US across a simulated conditioning trial where the elements each receive a boost from attention. The E_{CS} and E_{US} elements reach a high level of activation, while the I_{CS} and I_{US} units are largely suppressed by activation of A_{CS} and A_{US}. To illustrate the effect of attention, Figure 3B shows the same trial but this time simulated with attention removed. There are two obvious consequences of this removal. First, both I units show high and sustained activity. Second, activity in both E units is reduced, being partially suppressed by the I units. As a direct consequence of both of these changes, the formation or strengthening of positive associations will be limited by the loss of E activity, and the rise in I will drive inhibitory learning. Figure 3C shows a typical learning trial after extensive conditioning, where performance is near asymptotic levels. In comparison with Figure 3A, the I_{US} unit is activated relatively early in response to associative activation of the E_{US} element by E_{CS}. The rise in I_{US} leads to a period of inhibitory learning between E_{CS} and E_{US} before the US is presented. When the US is presented, activation of the A_{US} unit suppresses the I_{US} unit, leading to a period of growth in associative strength. Learning reaches asymptote when these two changes cancel each other out. When the US is omitted after extensive conditioning (Figure 3D), the associative strength that is lost during the presentation of the CS is not regained, leading to extinction.

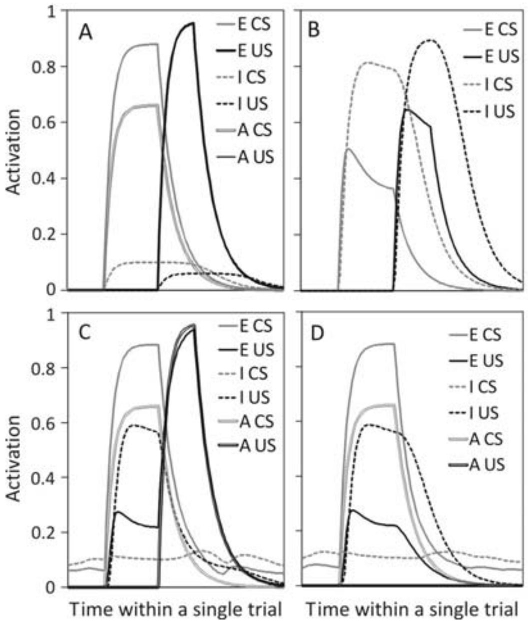

Figure 3. Activation profiles across time within a single trial for single units representing a CS and a US in the current model. Panel A: *E*, *I* and *A* responses of one CS element and one US element on the first presentation of CS and US. B: An identical scenario to Panel A but with attention removed from the simulation. Panel C: *E, I* and *A* responses of one CS element and one US element over a CS-US pairing after extensive conditioning has occurred. Panel D: Those same *E*, *I*, and *A* responses on a trial where the US is omitted.

Competition for attention between elements

The Harris (2006) model incorporated a limited capacity attention buffer as a means of providing non-linear changes in element activation when stimuli were presented in compound. In the current model, this is achieved by a competitive network of attention units. Increasing the "competition" in the network (increasing the total amount of activation in all A units) affects learning via its impact on the activation of E and I units. As the number of elements stimulated by S increases, so too does the normalizing input *A′*. Consequently *A* is suppressed to a greater extent, allowing *I* to increase, which in turn suppresses E. All else being equal, for any element x, E_x will be smaller and I_x greater when the element is activated as part of a compound of two stimuli than when activated as part of a single stimulus. However, the differences are not uniform. Elements with strong external input or weak external input will remain relatively unaffected by the addition of an extra stimulus because activation of their attention units will change very little (they will either remain active if they receive strong input, or remain inactive if they receive weak input). Elements with intermediate external input are more affected because their attention units suffer a sizeable loss of activity when part of a compound. It is these elements whose activations change the most when additional stimuli are presented in compound. These effects are illustrated in Figure 4.

Normalization within and between stimuli

As described above, activity in the network is normalized by summed input from E elements to each I unit, and this normalizing influence between elements is weighted as a function of their similarity. The weighting is expressed by the parameter z_i in Equation 4.2. Input from each E_i to its own I_i unit is unscaled ($z_{i,i} = 1$), whereas input from each element E_i to I_j is scaled by $z_{i,j} < 1$. To simplify this aspect of the model, we have not attempted to implement a graded change in the values of z_i across elements. Rather, we have set $z_{j,i} = 1/20$ (specifically 1 ÷ the number of elements in the stimulus) for all E_j elements that are activated by the same stimulus as E_i. Thus each element is normalized by all other elements that represent the same stimulus, and each element contributes 5% of its activity to the normalization of the other elements.

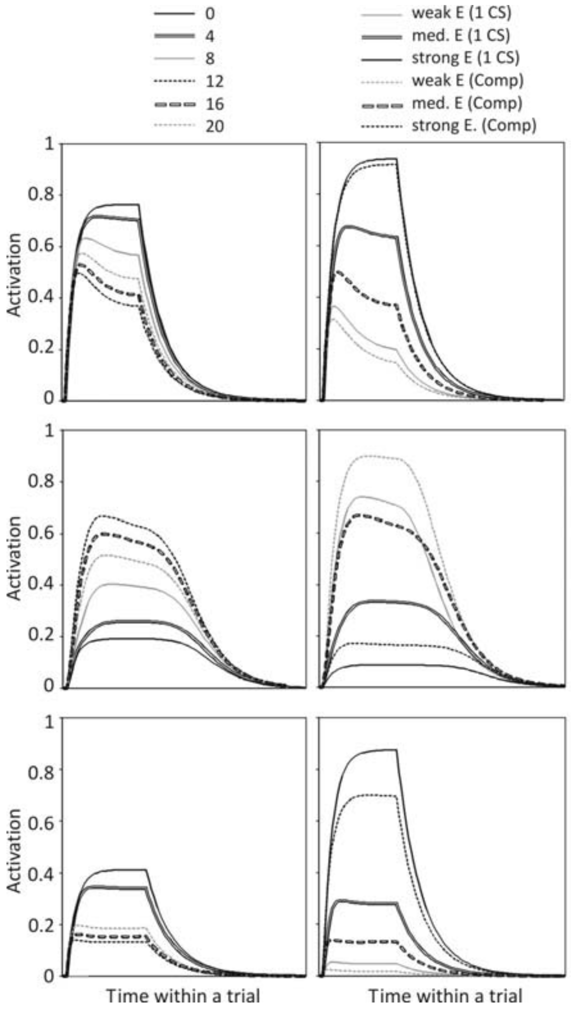

Figure 4. Left Panels: E, I and A responses of an element across the course of a stimulus presentation, as a function of the number of elements simultaneously activated. All elements have been given equal input from S. Right Panels: The E, I and A responses of 3 CS elements (receiving the weakest, median, and strongest S respectively) across the course of a stimulus presentation, as part of a single CS presentation and compound CS presentation.

In most simulations that we have conducted here, we assume that elements representing different stimuli do not contribute to each other's normalization. Thus, we have set z_i = 0 for all connections from the E elements of one stimulus to the I units of another stimulus. This assumes that those stimuli are very distinct, which is certainly an appropriate assumption for stimuli from different sensory modalities. However, this property of the model does provide a mechanism whereby stimuli from the same modality may interact given appropriate featural and spatial proximity. That is, two stimuli from the same modality may contribute to normalization of each other's elements by allowing z_i to be greater than zero (though less than 1/20). In simulations conducted here, whenever the design involved different stimuli from the same modality, we have set $z_{i,j}$ = 1/40. We acknowledge that this is a relatively arbitrary value, and the systematic manipulation of this parameter could be important in testing the model's performance against empirical findings. As discussed in several places below, the ability of the model to account in a principled way for stimulus similarity has served it well in enhancing its explanatory power.

Other details of the simulations

To reveal the model's behavior under a variety of situations, we have run simulations using the operations defined in Equations 1 through to 8. Except where specified, the results of the simulations that we present here are the average of at least 10 simulated runs – each run produces a slightly different result by virtue of the probabilistic nature of the connectivity between E elements in the network. Details that have not been specified already are described below.

In all simulations described here, each stimulus (including the US and context) has 20 elements. The strength of sensory input to each element within a stimulus is varied uniformly across a predefined range. The strongest US input is set to 2, the strongest CS input to 1, and the strongest context input to 0.5. The weakest input in each case equals the strongest input divided by the number of elements (20).

Each trial is broken into 150 moments. The onset of each CS or compound is at t = 20 and its offset at t = 50. On reinforced trials, CS offset coincides with the onset of the US which ends at t = 60. Context sensory inputs remain on for the entire trial. Every computational operation is performed at every moment and provides the input for computations at the next moment.

We will concentrate first and foremost on a range of complex discriminations which, like negative patterning, have no linear solution based on a simple summation of the associative strength of each individual stimulus. This type of discrimination has historically provided the greatest challenges to elemental learning theory, and certainly there is still a prevailing view that this type of discrimination is unsolvable by elemental learning mechanisms alone (see for instance Melchers, Shanks, & Lachnit, 2008, and replies).

Simple conditioning and extinction

Figure 5 shows the outcome of a single simulation with one CS that is continuously reinforced for the first half of the simulation (left panel) and then non-reinforced for the second half (right panel). Both panels plot the activation of US elements over the duration of the CS. Learning is negatively accelerated: initial increases in the activation of the US during the CS are relatively large, but diminish towards an asymptote. However, careful inspection of the earliest stage of conditioning reveals that the learning curve has, at this point, an upward inflexion (i.e., the increments in US activation from one trial to the next initially get larger, before showing the familiar decrease in growth rate towards the asymptote). This sigmoid function is consistent with general evidence for the development of conditioned responding across the course of simple conditioning (cf. Mackintosh, 1974), but it is unlike linear-difference models of learning in which the learning curve is negatively accelerated right from the outset of conditioning (e.g., Rescorla & Wagner, 1972). The initial upward inflexion of the curve here does not reflect an accelerated learning rate, indeed increments in associative strength are constant over these initial trials (i.e., learning is linear). Rather, the upward inflexion is a type of performance effect, being a direct consequence of the accelerating non-linearity in the element activation function itself (see Figure 2). That is, as associative input from CS elements to US elements increases from zero, the response of the US units increases supralinearly. This effect is somewhat analogous to the “response threshold” notion espoused by Spence (1956), according to which associative strength must surpass a threshold before its effect can be expressed in behavior. The decelerated component of the curve reflects the progressive acquisition of inhibition in each trial during the period when the CS activates the E_{US} elements. Here, the E_{US} elements activate their I_{US} units, leading to inhibitory associative change within the trial. The asymptote of learning is reached when the inhibitory associative changes driven by I_{US} units match the excitatory changes supported by the E_{US} elements (i.e., the term inside the brackets in Equation 6 equals zero).

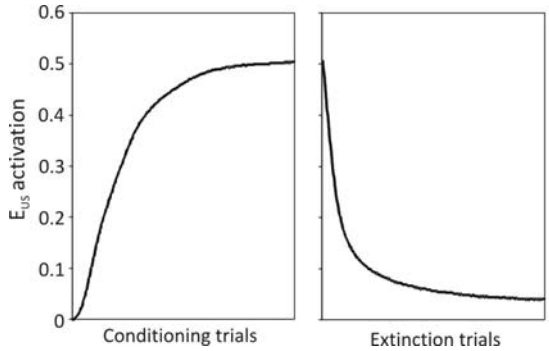

Figure 5. Left: Activation of US elements during the period of CS input across each trial of a simulation of simple conditioning of a single CS. Right: Activation of the same US elements during the same CS across each trial of an extinction phase when the CS was no longer followed by the US.

Like conditioning, extinction is negatively accelerated: US activation declines quickly when extinction commences but this decline slows as US activation approaches a floor (see right panel of Figure 5). Note, however, that this is floor is above zero. In the figure, extinction all but stops when activation of US elements falls below 0.05, rather than showing the uniform exponential decay function common to learning models that equate learning with a linear difference term (e.g., Rescorla & Wagner, 1972). This dramatic slowing of extinction occurs in the present model when the associative activation of US elements is sufficiently low that the E_{US} elements are barely able to activate their I units enough to generate further inhibitory learning. This threshold effect for activation of I units is a product of the non-linear activation function shown in Figure 2.

Negative patterning

The model we present here was developed to solve complex conditional discriminations. Therefore it should come as no surprise that, like its predecessor (Harris, 2006), the model can solve negative patterning and the biconditional discrimination, as confirmed by simulations shown in Figure 6. To explain how the model solves negative patterning, we must consider how normalization of activity in the attention network changes between compound and single CS trials. The increased number of sensory inputs on compound trials results in increased gain control. Because of the sigmoid shape of their activation function, attention units with strong input (A_{high}) remain relatively unaffected by the increased gain control, whereas units with weaker input (A_{mid}) are suppressed. For example, referring to the gray curves in the right hand plot of Figure 2 (p = 5, as for the attention units in the present model), a unit with an input value of 1 will suffer a relatively small decline in its activation if the normalization value (N) increases from 0.4 (the approximate value for a single 20-element stimulus) to 0.8 (the approximate value for two such stimuli), whereas units with inputs between 0.5 and 1 will suffer much greater loss of activity. This differential loss of activity across attention units translates into a differential decline in activation of the corresponding E elements because their inhibitory I units are less effectively suppressed by attention. Moreover, the rise in I_{mid} activity leads to the gradual development, across trials, of inhibitory associations from the E_{high} elements, which remain active in the compound, to the E_{mid} elements. Thus the two processes combine to effectively suppress activity in these E_{mid} elements on compound trials. It is this difference in activation of E_{mid} elements between single CS and compound presentations that provides a solution to negative patterning because the difference is correlated with the occurrence of reinforcement (see also Harris, in press, for discussion of these processes).

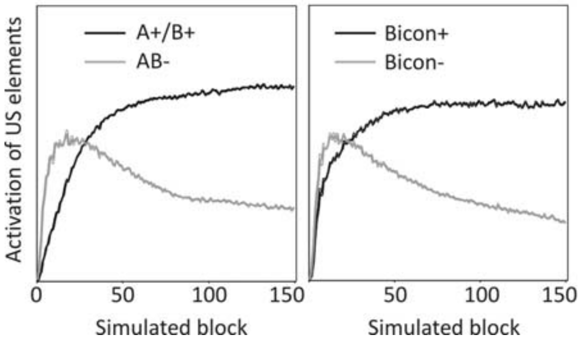

Figure 6. Activation of US elements during presentation of single or compound CSs across 150 simulated blocks of a negative patterning discrimination (left; one block comprises one A+, one B+, and two AB- trials) and a biconditional discrimination (right; one block comprises one trial of each compound). Each plot is the average of 20 simulations generated using the model proposed here. In each case the model eventually learns to perform correctly in that its US elements are activated more strongly on trials that are followed by the US (+) than on trials without the US (-).

The biconditional discrimination

The role played by attention in solving negative patterning is less effective in solving a biconditional discrimination. This is because all trials involve presentations of compounds, and thus normalization within the attention network is more closely equated for reinforced and non-reinforced trials. Indeed, the previous version of the present model could not solve biconditional discriminations except by assuming incidental differences in the salience of each stimulus that would create differences between compounds in the competition for access to attention (Harris, 2006). The present model, however, can solve biconditional discriminations (see Figure 6) without appealing to pre-existing differences in stimulus salience, but by relying on incidental differences in connectivity between elements. Its means for doing so depends on the acquisition of inhibitory associations between elements. As described above for negative patterning, across trials stronger E elements will develop inhibitory associations against weaker E elements because the inhibitory units of the latter elements will not be effectively suppressed by attention. However, incidental differences in connectivity between E elements means that a given element will be more strongly inhibited in one compound than in another compound. Any such element would acquire excitatory associative strength with US elements if it is more active in the reinforced compound than the non-reinforced compound, but would acquire inhibitory strength if more active in the non-reinforced compound than the reinforced compound. In other words, specific elements will emerge as virtual configural units for different compounds as a consequence of incidental differences in connectivity between elements, making the compounds themselves more distinct due to these variations in the pattern of their elements' activation.

Empirical Evidence

In this section, we review findings from key experimental investigations into the way that animals solve negative patterning discriminations. In line with the ongoing debates about the merits of elemental and configural representations (e.g., Pearce, 1987, 1994; Wagner, 2003, 2008), we extend this review to include studies of response summation when two or more CSs are combined as a compound.

Negative patterning and biconditional discriminations

While all the models described here are equipped to solve negative patterning and biconditional discriminations, most of these models tend to solve the biconditional discrimination more quickly (McLaren & Mackintosh, 2002; Pearce, 1987, 1994; Rescorla, 1972; Wagner & Brandon, 2001; Whitlow & Wagner, 1972). In light of this, we recently compared the rate at which rats and humans learn negative patterning and biconditional discriminations. Rats were trained in a magazine approach paradigm (Harris et al., 2008), and human participants were trained in a causal judgment task (Harris & Livesey, 2008). In each case, the negative patterning discrimination was mastered more quickly than the biconditional discrimination.

For the currently proposed extension of the Harris model, the prediction about the relative difficulty of biconditional and negative patterning discriminations is less clear than for its predecessor. Certainly the simulations shown in Figure 6 seem to solve the negative patterning and biconditional discriminations at similar rates. However, there are differences in the way the model solves the two discriminations, and two relevant factors can be identified. First, the solution to negative patterning benefits from competition for attention because some elements (those with intermediate activation strength) will be less active in the compound than in single stimuli. Second, both discriminations benefit from the formation of inhibitory inter-stimulus connections because these lead to elements becoming active only in the compound and not in the single CSs, or only in one compound but not another. However, the opportunity for inhibitory inter-stimulus connections is greater for the biconditional discrimination because each trial type involves the compound of two stimuli. The relative contributions of these two factors will vary according to how much competition there is for attention (i.e., how much the A units normalize each other) and the amount of connectivity between the elements. Thus negative patterning will be easier than the biconditional discrimination when the normalizing effect of attention contributes significantly to the non-linear behavior of the elements relative to the effect of inter-stimulus inhibitory links. Conversely, when attention has a relatively minor impact on the *E* and *I* responses of the elements (either because competition is too weak or too strong), biconditional discriminations should be easier than negative patterning.

In assessing different accounts of how animals solve negative patterning discriminations, Redhead and Pearce (1995b) trained animals (pigeons and rats) on a negative patterning discrimination in which one of the two stimuli (A) was more salient than the other (b). These authors pointed out that elemental models that invoke an added configural cue to solve these discriminations (Rescorla, 1972; Whitlow & Wagner, 1972) predict that responding to the salient A stimulus should emerge much sooner than to the weaker b stimulus, and as such the animals should show better discrimination of A+ versus Ab- than of b+ versus Ab-. In contrast, the configural model proposed by Pearce (Pearce, 1987, 1994) predicts that the b+ versus Ab- discrimination should emerge sooner than the A+ versus Ab- discrimination because the greater salience of A than b would lead to greater generalization between A and Ab than between b and Ab. The experimental results reported by Redhead and Pearce were consistent with this latter prediction, and contradicted the prediction of the elemental models with added configural cues. The ability of more recent elemental models (McLaren & Mackintosh, 2002; Wagner & Brandon, 2001) to account for this finding will depend on the manner in which they represent stimulus salience. In the present model, the physical salience of a stimulus determines the strength of sensory input to E and A units. Therefore, we incorporate differences in salience between two stimuli by arranging that the sensory input is lower in one than the other. Figure 7 shows the result of simulations in which sensory input from the weaker b stimulus was 90% of that for the stronger A stimulus. It is clear that the model anticipates superior discrimination between b+ and Ab- trials than between A+ and Ab- trials, just as Redhead and Pearce (1995b) found. The reason it performs in this way is relatively straight-forward. The weaker b elements will fare much worse than the stronger A elements in the competition for attention on Ab- trials, and thus many b elements will be activated strongly on b+ trials but not on Ab- trials, whereas many A elements will be strongly activated on both A+ and Ab- trials. As a result, the loss of associative strength on each Ab- trials will be largely confined to A elements, with b elements suffering much less associative loss. This will drive down responding to A more than b, producing poorer discrimination between A+ and Ab- than b+ and Ab-.

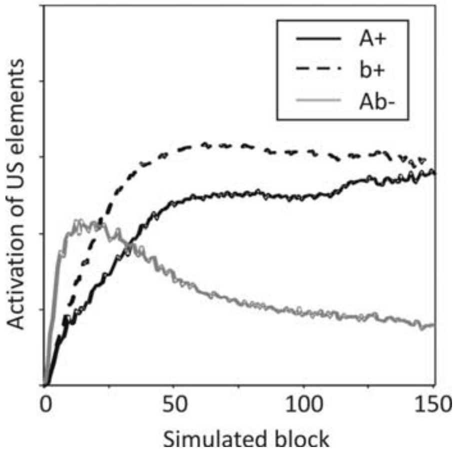

Figure 7. Simulation, generated using the present model, of a negative patterning discrimination in which one stimulus, A, is more salient than the other stimulus, b. The model anticipates better discrimination between b+ and Ab- trials than between A+ and Ab- trials, which is consistent with the results of an experiment by Redhead and Pearce (1995).

Complex negative patterning designs

Other empirical evidence that helps to distinguish the various models considered here comes from experiments that assess the impact of adding a third stimulus to a negative patterning discrimination (reviewed in Pearce, 1994). Pearce and Redhead (1993) reported an experiment that compared the performance of two groups of pigeons, one trained on a standard negative patterning discrimination (A+ B+ versus AB-) and the second trained on a discrimination that included a third (redundant) stimulus that was present on every trial (i.e., AX+ BX+ versus ABX-). They found that the inclusion of the redundant stimulus retarded mastery of the discrimination, in that the first group of pigeons learned the A+ B+ versus AB- discrimination more quickly than the second group learned AX+ BX+ versus ABX-. This difference is specifically predicted by Pearce's (1987, 1994) configural model because it assumes greater generalization from AX and BX to ABX than from A and B to AB. However, the difference in difficulty between the two forms of negative patterning is only anticipated by the McLaren-Mackintosh (2002) model when it is assumed that the amount of learning per trial is large, and is particularly problematic for the added (and replaced) configural element models (Rescorla, 1972; Wagner, 2003; Wagner & Brandon, 2001; Whitlow & Wagner, 1972) which predict that the addition of the redundant stimulus should facilitate learning of the discrimination, a prediction derived from their assumption that conditioning will proceed more quickly to AX and BX than to A and B. As shown by the left plot in Figure 8, the elemental model proposed here also tends to perform better on negative patterning discrimination with a redundant stimulus than on standard negative patterning, and thus the model would appear to be contradicted by Pearce and Redhead's findings.

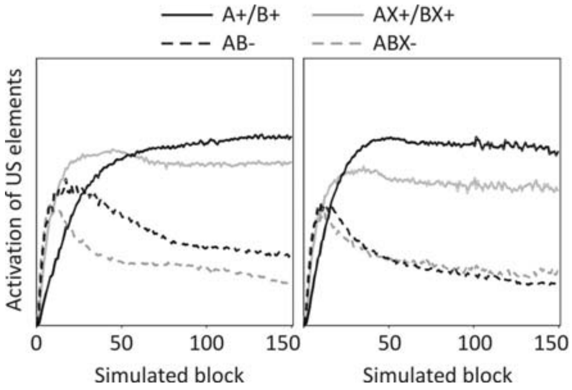

Figure 8. Simulations of a standard negative patterning discrimination (A+ B+ vs AB-; black lines) and discrimination with an added redundant stimulus (AX+ BX+ vs ABX-; grey lines), generated from the elemental model proposed here. The two plots show the performance of the model when simulated assuming that no stimulus contributes to the normalization of any other stimulus (left) and assuming that each stimulus contributes to the normalization of every other stimulus (right).

While the results of the experiment by Pearce and Redhead (1993) appear to discount the elemental models (including those with added configural elements), those findings may be a specific consequence of the choice of stimuli that were presented to the pigeons in the negative patterning discriminations. All three stimuli in those experiments were visual (red, green, and white coloured patches), and were presented together on a small screen (approximately 4.5 x 5.5 cm). As argued by Myers, Vogel, Shin and Wagner (2001), such similar and spatially close stimuli may well interact in ways that could influence their representation in an elemental network. The nature of these interactions can be specified in the elemental model we propose here in terms of the direct effect of each stimulus on gain control operating on the other stimuli. In the simulations shown on the left of Figure 8, none of the stimuli had any direct effect on the gain control mechanisms of any other stimulus (there was only an indirect effect via attention). However, if we allow stimuli to contribute to each other's normalization (i.e., if z_i is greater than zero for every E_i in calculating the input to I_x in Equation 4.2), the model does become slower to solve negative patterning when a redundant stimulus in included (see the plot on the right of Figure 8). This is important because, in the present model, different stimuli would be expected to contribute directly to each other's normalization if they were spatially and featurally similar. It is reasonable to suppose, as argued by Myers et al (2001), that the stimuli used by Pearce and Redhead were sufficiently similar to produce such normalization effects. Indeed a normalization-like effect between stimulus A and stimulus B was physically introduced by Pearce and Redhead by virtue of the manner in which they presented the stimuli together in compound. That is, both A and B were shrunk by 50% whenever they were presented in compound, relative to their size when presented separately. This physical change in the stimuli, as well as a likely sensory interaction between them, could change the nature of the prediction of elemental models. In contrast, Myers et al. (2001) describe an unpublished eyeblink conditioning experiment conducted by Bahçekapili in which rabbits trained on a negative patterning schedule with a redundant cue mastered this discrimination more quickly than rabbits given a standard two-stimulus negative patterning discrimination. This result is the direct opposite of that found by Pearce and Redhead (1993). An important difference between this experiment and the earlier one by Pearce and Redhead is that the experiment by Bahçekapili used three stimuli from different modalities (one visual, one auditory, and one tactile, cited in Myers et al., 2001).

As shown in Figure 8, such considerations of the effect of stimulus interactions do allow the present model to account for both the result reported by Pearce and Redhead (1993) and that obtained by Bahçekapili (cited in Myers et al., 2001). In the latter case, since one would expect little or no direct sensory interactions between such physically different stimuli, these stimuli should be treated by the present model as exerting no direct effect on one another's normalization. Without normalization between stimuli, the model performs better on negative patterning discrimination with a redundant stimulus than on standard negative patterning. That is, if A, B, and X do not contribute to each other's normalization, other than via attention, then the model solves the AX+ BX+ versus ABX- discrimination faster than the A+ B+ versus AB- discrimination (see the left panel of Figure 8). This is the result obtained by Bahçekapili (cited in Myers et al., 2001) using three stimuli from different modalities. In the case of Pearce and Redhead (1993), the physically overlapping visual stimuli are simulated by adding normalization between the stimuli, as outlined above. For each element i of one CS and element j in another, the input from E_i to I_j was given a weight $z_{i,j}$ = 1/40. Doing so yields the results shown in the right panel of Figure 8. Thus, the present model can account for both the finding reported by Pearce and Redhead (1993), and the opposite result by Bahçekapili, by assuming that sensory normalization affected the representation of the visual stimuli used by Pearce and Redhead, but that there was no normalization between the stimuli from different modalities used by Bahçekapili. The reason why stimulus normalization reverses the predicted order of these discriminations will become clear in the discussion of stimulus summation later in this article.

Two other negative patterning discriminations explored by Pearce and colleagues involve presentation of all possible combinations of three stimuli, and either reinforcing all trials except those on which all three CSs are presented together (i.e., A+ B+ C+ AB+ AC+ BC+ versus ABC-), or reinforcing all trials except those with two-stimulus compounds (i.e., A+ B+ C+ ABC+ versus AB- AC- BC-). Elemental models that use added configural elements to solve these discriminations (Rescorla, 1972; Wagner, 2003; Wagner & Brandon, 2001; Whitlow & Wagner, 1972) predict that responding to the reinforced compounds should be greater than responding to the single CSs, due to strong summation of associative strength when CSs are presented in compound. That is, in the first schedule these models predict that responding will be higher on AB+ AC+ and BC+ trials than on A+ B+ and C+ trials, and in the second schedule they predict greater responding on ABC+ trials than A+, B+ and C+ trials. On the other hand, Pearce's (1987, 1994) configural model predicts more responding on A+, B+ and C+ trials than on AB+ AC+ and BC+ trials in the first schedule, but predicts that the second schedule will be unsolvable because generalization to the two-stimulus compounds will be too strong.

Experiments in which pigeons were trained with the discriminations described above failed to confirm either of the predictions of the elemental models, but nor did they fully confirm the predictions of Pearce's configural model (Pearce, Esber, George, & Haselgrove, 2008; Redhead & Pearce, 1995a). That is, in the first design pigeons responded more on A+ B+ and C+ trials than on AB+ AC+ and BC+ trials (Redhead & Pearce, 1995a), consistent with the prediction made by Pearce (1994) but not that made by elemental models using configural cues to solve these discriminations. However, in the second design the pigeons did successfully solve the discrimination (Pearce et al., 2008), in contrast to the prediction of Pearce's model[i], and they responded more on A+ B+ and C+ trials than ABC+ trials, which is not consistent with the prediction of the elemental models with an added configural cue. It must also be noted that an experiment by Myers et al. (2001), in which rabbits were trained with the first type of negative patterning design, did produce results consistent with the prediction of the elemental models but not Pearce's configural model, in that the rabbits responded more on AB+ AC+ and BC+ trials than on A+ B+ and C+ trials (a similar result has been reported recently by Lachnit, Schultheis, Konig, Ungor, & Melchers, 2008, in a causal learning experiment with human subjects). As discussed above with respect to the effect of a redundant stimulus on negative patterning, Myers et al. (2001) point out that the contrasting results from their experiment and that of Redhead and Pearce (1995a) could be due to the fact that the rabbits in Myers et al's. experiment were trained with stimuli from three different modalities (one visual, one auditory, and one tactile) whereas the Redhead and Pearce's pigeons were trained with three visual stimuli.

The elemental model proposed here can account for all of the above findings if, once again, we assume that stimuli from different modalities do not contribute directly to one another's normalization (gain control) but that stimuli from the same modality will contribute to each other's normalization especially when presented in close spatial proximity. When trained with the A+ B+ C+ AB+ AC+ BC+ versus ABC- schedule, the model predicts stronger responding on AB+ AC+ and BC+ trials than on A+ B+ and C+ trials if the E elements of each stimulus do not provide input to the I units of any other stimulus (see simulation in left panel of Figure 9). Thus the model predicts the finding reported by Myers et al. (2001) when stimuli do not directly influence each other's normalization, as would be expected for the stimuli used by Myers et al. However, if the E units of one stimulus do provide input to the I units of other stimuli that are presented at the same time, then the predicted difference between the single CSs and two-stimulus compounds is reversed (see simulation in right panel of Figure 9). This reversal occurs because the opportunity for summation of associative strength between CSs is reduced when those stimuli contribute to each other's normalization. Thus the model can account for the results reported by Redhead and Pearce (1995a) if it assumes that each stimulus contributes to normalization of any other stimulus that is presented at the same time, as may well have occurred between the visual stimuli presented in close proximity by Redhead and Pearce. Indeed, Redhead and Pearce physically introduced such a normalization effect between the stimuli because the size of each stimulus was reduced as a fraction of the number of stimuli presented on a trial. That is, each stimulus was shrunk by 50% when presented as part of a two-stimulus compound, and was shrunk to 33% when presented in the triple compound.

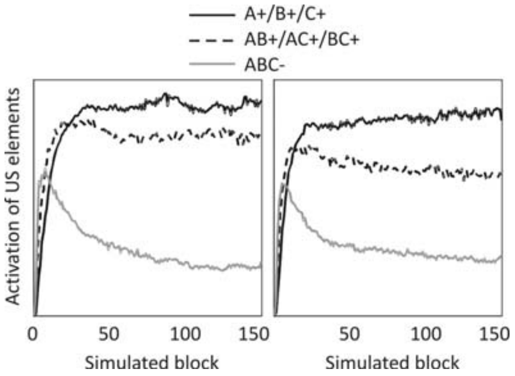

Figure 9. Simulated results, generated by the elemental model proposed here, of a complex negative patterning discrimination in which each of the three individual stimuli are reinforced (A+ B+ C+) and each of the three pairwise compounds are reinforced (AB+ AC+ BC+) but the triple compound is not (ABC-). The plot on the left shows the simulated performance when each element only contributes to normalization of other elements in the same stimulus, and not to normalization of elements of other stimuli (as would be expected for very different stimuli, such as from different sensory modalities). The plot on the right shows the simulated performance when elements contribute to the normalization of elements from other stimuli as well as their own (the cross-stimulus influences on gain control were set to half the strength of the within-stimulus inputs to gain control). Such cross-stimulus gain control effects would be expected between stimuli that are similar and presented in close spatial proximity.

The model also successfully accounts for the results of the A+ B+ C+ ABC+ vs AB- AC- BC- discriminations (Pearce et al., 2008). Simulations (not shown) demonstrate that it correctly produces weaker CS-generated activation of US elements on the double-CS compound trials than on single-CS or triple-CS trials. Moreover, the difference in simulated responding between the single- and triple-CS trials depends on whether there is normalization between stimuli. Without normalization between stimuli, the model predicts greater responding to ABC than to A, B, or C individually, in line with the prediction of other elemental models. However, when stimuli do contribute to one another's normalization, the predicted difference is reversed. Thus the model explains the result obtained by Pearce et al. if it assumes that the visual stimuli they used were sufficiently similar that normalization occurred between them. Once again, we point out that, in the experiment conducted by Pearce et al., a normalization-like interaction was physically introduced into the stimulus presentations in that part of each stimulus was displaced when it was presented with a second or third stimulus.

The current model explains many of the relevant results by appealing to a greater degree of stimulus normalization when the stimuli are spatially close or share similar features. This explanation will therefore be troubled by any data that are in line with those of Pearce & Redhead (1995a) but obtained usinge stimuli that do not contribute to each other's normalization. In this regard two experiments by Pearce and George (2002) would appear to be at odds with our exposition. These experiments produced similar results to those of Redhead & Pearce (1995a) using a discrimination in which A, B, and C were spatially distinct visual cues (shapes of differing color or pattern). If there were no normalizing interaction between these shapes then our model would not predict the results reported by Pearce and George. However, the details of the predictions from our model are somewhat unclear in this case. That is, although the stimuli did not overlap spatially, their proximity may still have effected some normalization within the pigeons' visual systems. Indeed, these stimuli may have affected one another's detectability at a more peripheral level if, by looking at one stimulus, the pigeons' gaze was directed away from other concurrently presented stimuli. This is of course speculation, but it highlights a limitation to the testability of the model we propose here since its predictions depend on details of the normalization process that cannot typically be specified a priori. However, these issues speak to the value of experiments that use stimuli from different modalities since only in this case can we be relatively sure that direct perceptual interactions are minimal or absent.

Perhaps the findings that are most difficult to reconcile with the account we offer here were those reported recently by Redhead (2007) who tested the effect of using within-modality and between-modality stimuli in the complex negative patterning design in a human contingency learning experiment. He found the same pattern of results in a group where A, B, and C were all visual, spatially overlapping stimuli (Experiment 1), and a group where A, B, and C where visual, auditory and tactile stimuli (Experiment 2). Both groups exhibited slightly better discrimination for the single stimuli than for the pair compounds, which is at odds with the current exposition. It is too early to say whether this finding constitutes general evidence against the application of our model to complex patterning discriminations, or whether it identifies a specific limitation of the model with regard to the manner in which human subjects consciously seek a solution to these discriminations. Nonetheless, it should be noted that in the same study (Experiment 3), Redhead did find an effect of stimulus modality on summation. After training with reinforced single stimuli (A+/B+/C+), the between-modality condition produced significant summation for test pair compounds (AB/BC/AC), where as the within-modality group did not. As will be discussed, this finding *is* consistent with the predictions of the current model.

Summation

Many of the difficulties faced by elemental models when simulating findings from complex negative patterning experiments arise because the models tend to predict strong summation of associative strength between different CSs. Indeed, summation between CSs is another point of divergence between elemental models and Pearce's (1987, 1994) configure model. The similarity rule used by Pearce to determine the amount of generalization between compounds and their component CSs tends to predict that, when two CSs are combined, the associative strength of the compound will equal the average associative strength of the individual CSs. This is because, in the simplest working of the model, only 50% of the associative strength of each individual CS will generalize to their compound. In contrast, a simple elemental approach like that of the Rescorla-Wagner (1972) model predicts complete generalization of associative strength between CSs and their compound, such that the associative strength of the compound will equal the sum of the associative strengths of the two CSs. In this section, we consider evidence from experiments that directly measure summation, and discuss how well different elemental and configural models can account for the variety of findings.

The empirical evidence for summation is equivocal. There are numerous demonstrations that animals respond to the compound of two CSs more than to each individual CS (e.g., Kehoe, 1982, 1986; Rescorla, 1997), consistent with the assumption that the associative strength of the compound equals the sum of strengths of the two CSs. However, there are also numerous reported failures to observe summation in Pavlovian conditioning paradigms. Most of these failures have occurred in autoshaping experiments with pigeons (e.g., Aydin & Pearce, 1995, 1997; Rescorla & Coldwell, 1995). Nonetheless, both successes and failures to observe summation have been reported in other paradigms, such as the conditioned nictitating membrane response in rabbits (Kehoe, Horne, Horne, & Macrae, 1994) and the conditioned magazine approach with rats (Pearce, George, & Aydin, 2002; Rescorla, 1997; Thein, Westbrook, & Harris, 2008). Such mixed evidence is troubling for any model that assumes that most of the associative strength of each individual CS generalizes to the compound.

The sensory relationship between CSs is one factor identified as relevant to the amount of summation observed when CSs are combined in a compound. Kehoe et al. (1994) observed summation of the conditioned nictitating membrane response in rabbits that had been trained with two CSs from different modalities (one auditory and one visual) but not when the CSs were both auditory. A similar result has been reported by Aydin and Pearce (1997) and Thein et al. (2008). If compounds composed of CSs from the same modality do not to produce more responding than the individual CSs themselves, this could explain the many failures to observe summation in autoshaping with pigeons (Aydin & Pearce, 1995, 1997; Rescorla & Coldwell, 1995), because the CSs used in those experiments were from the same (visual) modality. In light of this, Wagner (2003) has specified operations in the Replaced Elements model that capture this relation (also Myers et al., 2001). According to that model, some individual CS elements are replaced by configural units when stimuli are combined in compound, and the number of elements that undergo replacement increases as a function of the similarity between the stimuli. Because associative strength is lost whenever individual CS elements are replaced, an increase in the number of replaced elements means a decline in generalization of associative strength from the individual CSs to the compound.

A functionally similar operation affects summation between CSs in the elemental model proposed here, but the mechanism is normalization between elements, rather than replacement. Elements with overlapping receptive fields contribute to each others' normalization by providing weighted input to their inhibitory I units. Stimuli from the same modality will reduce activation of each other's E elements, thereby reducing the strength with which they associatively activate the US elements. In contrast, very distinct stimuli, such as from different modalities, do not contribute to each other's normalization, and thus there is no loss of the activation when presented together versus individually (other than via a loss of attention). Simulations of the model have confirmed this impression. Figure 10 plots the strength with which US elements are activated during presentation of a compound as a ratio of the activation strength of those same US elements during presentations of the individual CSs. When there is no normalization between the two CSs ($z_i = 0$ in Equation 4.2), their compound activates US elements twice as much as the individual CSs do. As normalization between the CSs increases (as z_i increases from 0 to 0.05, at which point normalization between stimuli is equal to that within a stimulus) the activation of US elements by the compound decreases systematically.

Differences in normalization between stimuli as a function of their similarity can also explain conflicting data concerning summation in a triple-stimulus compound following conditioning of each single CS or conditioning of the same stimuli as two-stimulus compounds. Pearce, Aydin, and Redhead (1997) compared responding to a triple compound,

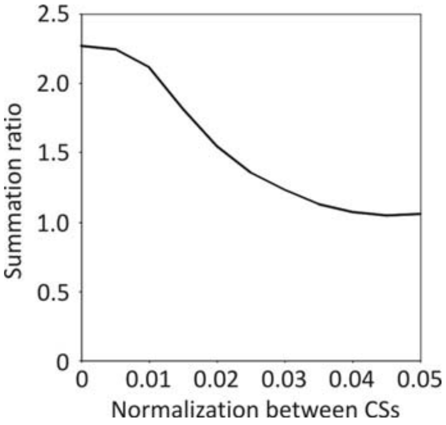

Figure 10. Summation as a function of the amount of normalization between CSs. The summation ratio is the activation of US elements across each presentation of the compound divided by the average activation of US elements across individual presentations of each of CSs. Normalization was set to vary between zero, when each CS contributes nothing to normalization of the other CS, and 0.05 (1 ÷ number of elements per stimulus), when normalization between CSs is equal to normalization between elements of the same CS.

ABC, between two groups of pigeons in an autoshaping experiment. One group had been trained with each of the CSs individually (A+, B+, and C+); the other group had been trained with the same three stimuli as three pairwise compounds (AB+, AC+, and BC+). They observed summation of responding in the second group of pigeons (i.e., responding to ABC > responding to AB/AC/BC), but not in the first group (responding to ABC = responding to A/B/C). This finding is consistent with the prediction of Pearce's (1987, 1994) configural model, but is problematic for elemental models that generally predict summation in both groups, and greater summation in the group conditioned with the single CSs. However, in an eyelid conditioning experiment with rabbits, Myers et al. (2001) obtained a result that is consistent with the prediction of elemental models but not Pearce's configural model. That is, they found much greater summation to ABC in rabbits trained with the single CSs than in rabbits trained with the two-stimulus compounds. Myers et al. argued that the difference between the two experiments was that their stimuli were from different modalities (one visual, one auditory, and one vibrotactile), whereas the stimuli used by Pearce et al. were all visual. Wagner (2003) has shown that the replaced elements model can account for both sets of findings by assuming that the stimuli used by Pearce et al. underwent substantial replacement when compounded whereas the stimuli used by Myers et al. underwent little replacement. In analogous fashion, the model we present here explains both sets of findings by assuming normalization between the three visual CSs used by Pearce et al. but no cross-modality normalization between the CSs used by Myers et al.. The other elemental model discussed here cannot accommodate both sets of data: The McLaren-Mackintosh model predicts greater summation to ABC after training with A+ B+ C+ than after training with AB+ AC+ BC+ (as found by Myers et al.), and while increasing overlap between stimuli reduces the amount of summation in each condition, the relative order of the two conditions is preserved. It remains a possibility that the model could appeal to some other variable, not related to stimulus similarity, that might also meaningfully distinguish the experiments by Myers et al. from those by Pearce et al., but no such difference is obvious to us.

Summation across compounds after negative patterning

The means by which different models solve negative patterning affects their predictions about how associative strength generalizes from the trained stimuli to novel compounds formed from the same stimuli. Rescorla (1972) and Whitlow and Wagner (1972) investigated this issue by training animals on a negative patterning discrimination with two stimuli (A+ B+ vs AB-) while concurrently conditioning a third stimulus (C+). They then tested the animals for responding to novel compounds formed between either of the negative patterning stimuli and the third stimulus (i.e., AC or BC). Responding to these new compounds was as high or higher than responding to any of the individual CSs, a finding consistent with the added configural cue hypothesis which anticipates summation of associative strength of each individual stimulus to the new compounds without the inhibitory strength of the AB configural cue. The evidence is, however, also consistent with the other models considered here. In simple form, Pearce's (1987, 1994) configural model predicts that responding to AC and BC should equal the average of the two constituent stimuli. But if one makes the reasonable assumption that the physical context is part of each stimulus configuration, and thus supports generalization across trials, the model does predict greater responding to AC and BC than to A, B, or C. The purely elemental models described by

McLaren and Mackintosh (2002) and Harris (2006) also predict greater responding to AC and BC than A, B, or C. Simulations of the model we propose here show that it too makes the same prediction.

Harris, Gharaei and Moore (2009) recently presented data from similar experiments that are less easily accommodated by the models considered here. Their experiments followed a similar logic to those described above: they trained rats on negative patterning before testing their responses to novel compounds composed of the trained stimuli. One experiment followed the design of Rescorla (1972) and Whitlow and Wagner (1972), but conditioned a compound, CD, instead of the single CS, C (i.e., rats were trained with A+ B+ vs AB- and CD+; A and D were auditory, B and C were visual). After extended training, the rats were tested for responding to the novel compounds AC and BD. The rats responded to these compounds significantly more than to AB and significantly less than to CD; their responses to AC and BD were above, but not significantly different from, their responses to A and B (Harris et al., 2009). In a second experiment, rats were trained on two concurrent negative patterning discriminations (A+ B+ vs AB-, and C+ D+ vs CD-), before being tested with the novel compounds AC and BD. On that test, the rate of responding to the new compounds was mid way between the rate of responding to the single CSs and the trained compounds. The final experiment trained rats on a negative patterning and a positive patterning discrimination, concurrently (A+ B+ vs AB-, and C- D- vs CD+), before once again testing the rats with the novel compounds AC and BD. In this case, responding on trials with the new compounds was marginally below, but not significantly different from, responding on the reinforced trials (A+ B+ and CD+ trials). The test data from all three experiments are presented in Figure 11. As discussed by Harris et al. (2009), each of the earlier models discussed here can account for the data from each of these experiments individually, but only by adopting different parameters in each case. That is, none of the models previously considered was able to simultaneously explain the data from all three experiments if constrained to adopt the same set of parameters across experiments – a reasonable requirement given that each experiment used the same set of stimuli, with the same strain of rat, and conditioned in the same paradigm using the same equipment.

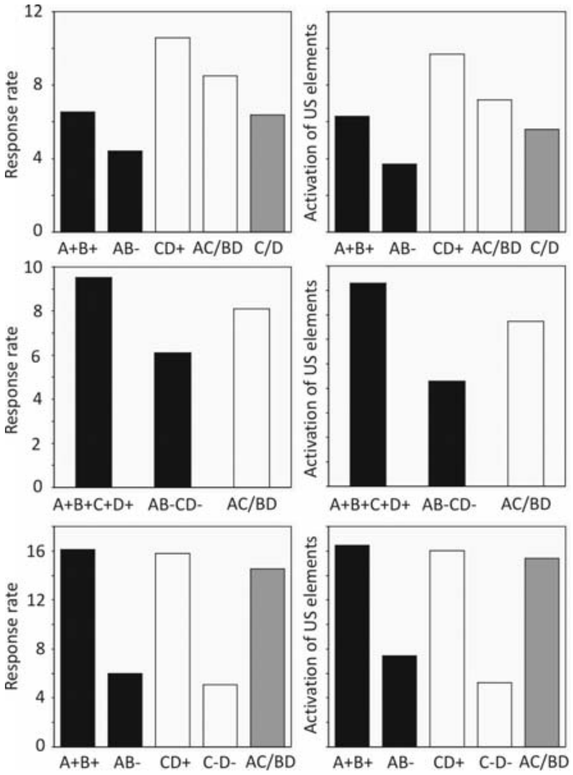

Figure 11. The three left hand panels (A, C, & E) show test results from three experiments reported by Harris et al. (2009) assessing rats responses to novel compounds AC and BD (grey columns) after training with these stimuli on different discriminations. In all experiments, A and B were trained in a negative patterning schedule. This was accompanied by either conditioning of the compound CD (panel A), training on a negative patterning schedule with C and D (panel C), or training on a positive patterning schedule with C and D (panel E). Panel A includes test performance to individual presentation of C and D, which had never been presented separately during training. The right hand panels (B, D, and F) show simulated results for each of the same designs, generated using the model presented here.

The model we propose here does a better job than its predecessors in accounting for the data presented by Harris et al. (2009). As shown in the right hand panels of Figure 11, when the model is trained on the discriminations to the point when the simulated activation of US elements matches the level of discrimination shown by the rats' behavioral responses, the model's prediction for the relative strength of responding to the novel compounds corresponds very closely to that observed empirically. In these simulations, we allowed stimuli from the same modality to contribute to one another's normalization, to be consistent with the approach to our simulation of the designs used by Pearce and colleagues described above (see also the section on Summation above).

The nature of the model's predictions can be understood as the product of two opposing factors: summation of associative strength across CSs and mutual inhibition between those same CSs. To explain, we will focus on one of the experiments – that in which rats were trained concurrently on two negative patterning discriminations (panels C and D in Figure 11). During the course of training, elements of each CS acquire excitatory associative strength with the US elements, but responding to the non-reinforced compounds during training is suppressed due to competition between the CSs for attention and due to the acquisition of inhibition from the strongest elements of each CS to the intermediate and weak elements of the other CS. For example, across repeated presentations, many A elements lose activation strength in the compound AB due to inhibition from B elements, and thus the associative strength of those A elements is not expressed. However, when A is presented with C as a new compound, those A elements that had been suppressed by B are no longer suppressed. The stronger activation of those elements means that their associative strength is expressed in the new compound, and thus responding to AC is higher than to AB. Indeed, given these processes the model tends to predict greater responding to the new compounds than was actually observed in the experiments by Harris et al. (2009). However the model's prediction for response rate is reduced if it assumes that stimuli from the same modality contribute to one another's normalization, as we have assumed in the simulations presented in Figure 11. This normalization between stimuli will establish some level of inhibition between the elements from the different stimuli in the new compound, such as between C and A, even though those stimuli had never been presented together during training. This inhibition develops because, on CD trials during training, D elements activate the I units of A (as a consequence of normalization between D and A), and this paired activity between C elements and I_A elements leads to the development of inhibition between C and A. This inhibition will have similar consequences as the inhibition between A and B – C will reduce expression of A's associative strength and A will reduce expression of C's associative strength.

Effects of stimulus pre-exposure

Before completing our discussion of the model proposed here, we wish to add a note regarding the effects of stimulus pre-exposure on subsequent conditioning. There are two very robust consequences if a stimulus is repeatedly presented prior to conditioning: (1) conditioning of that stimulus is retarded, relative to a non-pre-exposed CS (the so-called "latent inhibition" effect); and (2) generalization of responding to other non-conditioned stimuli is reduced (i.e., there is improved discrimination between the pre-exposed CS and other stimuli) (see Hall, 1991, for an extensive review and theoretical analysis of both effects; for recent reviews of latent inhibition, see Holmes & Harris, 2010; Lubow & Weiner, 2010).

The model we propose here does anticipate latent inhibition as a consequence of pre-exposure. Its mechanisms for doing so are somewhat similar to those expressed by Wagner (1981) and McLaren and Mackintosh (2002), inasmuch as activation of CS elements are reduced by virtue of context-stimulus and within-stimulus associations (Channell & Hall, 1983; Lovibond, Preston, & Mackintosh, 1984; Westbrook, Jones, Bailey, & Harris, 2000). This means that it can also account for evidence for context-specificity of latent inhibition. However, whereas both of those previous accounts attribute the change in stimulus associability to excitatory associations among stimulus elements (or from the context to the elements of the stimulus), in the present model pre-exposure reduces activation of many stimulus elements due to the development of inhibitory associations from the context and from other stimulus elements. These inhibitory associations develop because, during pre-exposure, the strong elements of the stimulus and of the context are active while the I units of the weaker and intermediate stimulus elements become active (these being less effectively inhibited by attention). This reduces the total associability of the pre-exposed stimulus because some of its elements are rendered inactive, and thus unable to participate in conditioning. This effect is shown in the left panel of Figure 12, which plots the simulated activation of US elements by CS elements on each conditioning trial of a pre-exposed CS and a non-pre-exposed CS. The figure also plots the simulated conditioning of a pre-exposed and non-pre-exposed CS when a "context shift" is introduced between pre-exposure and conditioning phases. The context shift is achieved in these simulations by resetting to zero all associative connections between the elements of the pre-exposed stimulus and the context. This dramatically reduces the difference between the pre-exposed and non-pre-exposed CSs. This effect of a context shift is, not surprisingly, sensitive to the salience of the context. In the simulations shown here, the maximum input strength, S, for the context was set to 1. Other simulations using a value of 0.5 (the value used for all other simulations presented in this article) revealed only a very modest effect of context shift on latent inhibition.

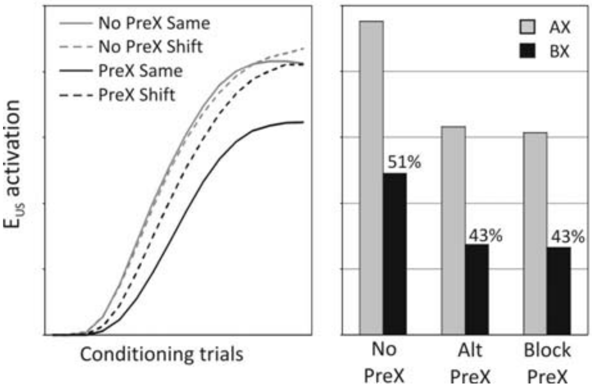

Figure 12. Left: Simulated conditioning of a pre-exposed (PreX) and non-pre-exposed CS. One set of simulations effected a context-shift (Shift) between pre-exposure and conditioning, whereas the other set maintained the same context throughout. Right: Simulated generalization of the conditioned response (CS-generated activation of US elements) between a CS, AX, and a non-conditioned stimulus, BX, which shared 50% of their elements in common (X). Activation of US elements is plotted for simulations that included alternating pre-exposures to A and B, blocked pre-exposures to A and B, or no pre-exposures to either stimulus. The values above each BX column represent the US activation on BX trials as a percentage of that activation on AX trials.

The other empirical consequence of pre-exposure is to reduce generalization between stimuli. This is achieved in the present model (see right-hand panel of Figure 12) because pre-exposure reduces the number of elements that undergo conditioning, effectively narrowing the stimulus' representation in the total population of elements, and thus reducing the number of elements that can support generalization to another stimulus. Experimental investigations of the effect have demonstrated that the schedule of pre-exposures is important in determining the degree to which generalization is reduced by stimulus pre-exposures. That is, if two stimuli, AX and BX (with shared X elements), are pre-exposed concurrently (i.e., presentations of AX and intermixed among presentations of BX) in advance of conditioning AX, the generalization of conditioned responding from AX to BX is lower than if AX and BX had been pre-exposed in different blocks (if all presentations of AX preceded BX, or vice versa) (Honey, Bateson, & Horn, 1994; Symonds & Hall, 1995). However, this difference in the effect of alternating versus blocked pre-exposure is not replicated by the present model. The right panel of Figure 12 shows the results of simulations in which AX was paired with a US after both AX and BX were given blocked or alternating pre-exposures, or no pre-exposure. In these simulations, AX and BX shared 50% of their elements in common. Both blocked and alternating pre-exposures have reduced generalization from AX to BX (from 51% to 43%), but have done so equally.

Other elemental models can explain the differential effect of the two pre-exposure schedules on generalization by appealing to the acquisition of mutual inhibition between the distinctive elements of each CS during alternating but not blocked pre-exposure (McLaren & Mackintosh, 2000). As such, it is perhaps surprising that the current model does not do likewise. Indeed, in the model, alternating exposure between AX and BX does establish mutual inhibition between A and B elements, whereas blocked exposure establishes only uni-directional inhibition (from the distinctive elements of the second stimulus to those of the first stimulus). However, mutual inhibition is only effective in reducing generalization from AX to BX if the presentation of BX would otherwise activate the distinctive A elements, via associative connections from the common X elements, and if these A elements in turn activate the US elements. In other words, the mutual inhibition between distinctive elements only serves to reduce second-order activation of US elements via A elements. In the present model, such second-order associations normally contribute little to generalization: associative activation of A elements on BX trials will be modest compared with the direct activation of X elements, and thus the contribution that A elements make to activation of US elements will be negligible.

Like many other accounts of latent inhibition, the means by which pre-exposure interferes with subsequent conditioning in the current model is related to decrements in processing of the CS. Indeed, the model resembles the mechanism described by Wagner (1981) and McLaren and Mackintosh (McLaren & Mackintosh, 2000) in that CS processing deficits are held to arise from associative connections from other elements to the CS elements. This particular approach gives no role for non-reinforcement during pre-exposure as an ingredient in latent inhibition. That is, these models are indifferent to the specific events that follow each stimulus presentation during pre-exposure, whereas most other accounts of latent inhibition place great importance on the fact that the stimulus is followed by no outcome during pre-exposure, suggesting for example that the animal learns to ignore the stimulus specifically because it signals nothing (e.g. Le Pelley, 2004; Lubow, Weiner, & Schnur, 1981; Mackintosh, 1975). As it stands, the model we propose here does not incorporate a mechanism by which the CS processing deficit is affected by the presence or absence of an event following each stimulus pre-exposure. However, we believe the model could be made to do so depending on the temporal dynamics of the rise and decay of activity in different units. Specifically, if activity in attention units were to decay much more rapidly than activity in E elements at the offset of a stimulus, this could cause a transient rise (or rebound) of activity in I units of the stimulus due to the sudden removal of their suppression by attention combined with continued input from residual activity in E elements. This rise in I unit activity would lead to the development of inhibitory associations from contextual and other elements to the E elements of the stimulus. That is, this process would increase the amount that CS elements are suppressed during subsequent conditioning, thereby enhancing evidence latent inhibition. Moreover, if during pre-exposure the stimulus were followed by another event, rather than by nothing, the elements of that second event would overshadow contextual and other CS elements in the formation of inhibitory associations with the CS elements. In other words, the second event would become an inhibitor of the pre-exposed stimulus, as a consequence of their backward pairing, and so protect the stimulus from becoming inhibited by the context and by itself.

While such changes can explain much of the evidence concerning latent inhibition, there are aspects of latent inhibition that are not explicable by such means. A clear example is the observation that latent inhibition can be lost (i.e., responding to the pre-exposed CS spontaneously recovers) across the course of a delay between the end of conditioning and the beginning of test (Aguado, Symonds, & Hall, 1994; Westbrook et al., 2000). Clearly, in such cases, latent inhibition must, at least in part, constitute a performance deficit, whereby something learned across the course of pre-exposure interferes with conditioned responding or with CS-primed retrieval of the US representation (e.g., Bouton, 1993). Indeed, in reviewing the literature, Hall (1991) concluded that latent inhibition is likely to be due to both a learning deficit, arising from changes in stimulus processing, and a performance effect due to interference with retrieval of the CS-US association. Models such as the present one can be pushed to predict some form of performance deficit, in addition to a learning deficit, if there remains some residual deficit in CS processing after conditioning which continues to depress element activation. However, such effects are likely to be small given that those elements suffering residual suppression after conditioning will also have been suppressed during conditioning, and thus acquired little associative strength. Therefore, the continued suppression of those elements at test will make relatively little difference to the potential of the CS to activate the US elements.

The possibility that learning during pre-exposure interferes with retrieval of the CS-US association poses a considerable challenge for neural network models of learning since typically there is no mechanism whereby associatively-activated events interfere with one another's representation in the network. Although the present formulation of the model does not achieve this, and it is not our intention to develop the model in this way here, it is instructive to consider how the present model implies a framework by which such interactions may occur between associatively-activated representations. Specifically, the normalization process used here to capture competition between the processing of stimuli is one reasonable means by which associative representations of events may interfere with one another. For example, if attention units could be driven by activity in E elements (rather than purely by external S inputs, as we have specified here), this would allow attention to modulate associatively-activated US representations and enhance conditioned responding. Moreover, if elements representing different outcomes were to be activated simultaneously by associative inputs from a CS, their competition for attention would provide a basis by which one could interfere with the other. For example, if pre-exposure established associations between the CS and the context (in light of the context elements regaining activity after each offset of the stimulus), associative activation of context elements by the CS might interfere with the associative activation of US elements by that CS across the course of conditioning.

Concluding remarks

The model we have presented here extends an elemental model of associative learning put forward recently by Harris (2006). Like that earlier model, it uses an "attention" mechanism to introduce non-linearity into the behavior of elements, and to provide a means by which elements can affect each other's activity. It extends that earlier model in three important ways. First, the attention mechanism is specified in more precise terms, defined here as a network of units that regulates activity in the network of sensory elements that represent each stimulus. Second, our description of the relationship between input to an element and the response of that element includes a gain control mechanism that normalizes activity in the local network. The normalization rules are taken from computational models of perceptual systems that have been developed to explain a large range of psychophysical and neurophysiological findings. In the current model, normalization provides an opportunity for competitive interactions between elements and introduces non-linearity in the response of elements. A similar normalization process operates within the network of attention units. This creates capacity-limitation-type effects within the attention system, and endows the behavior of the attention units with their own non-linearity that further increases the non-linear behavior of sensory elements. Finally, the operations of the model have been specified at much greater temporal resolution to function at a sub-trial or "real-time" level. This is important in order to adequately capture certain dynamic properties of the model, in particular the acquisition and expression of excitatory and inhibitory connections between elements of the same stimulus (or compound), and the normalizing interactions that evolve within the network of sensory elements and attention units. We show how this model is equipped to explain a large range of experimental findings, particularly findings regarding solutions to non-linear problems such as negative patterning and the biconditional discrimination. Indeed, this new model is superior to its predecessor in accounting for some recent findings concerning summation of associative strength between stimuli trained in negative patterning discriminations.

Melchers et al. (2008) have recently argued that neither elemental nor configural learning models can account for the full range of data on the learning of complex discriminations, and that the encoding of stimuli is flexible in the sense that the representations that engage in associative learning may shift from elemental to configural or vice versa according to the needs of the task at hand. Several authors, including ourselves, have defended the efficacy of the elemental approach in solving non-linear discriminations (e.g., Liljeholm & Balleine, 2008; Livesey & Harris, 2008; McLaren, 2008), and the current analysis clearly pursues the same goal. The model we present here reveals how relatively simple mechanisms of gain control and element competition provide solutions to discriminations which appear configural in nature.

Although the need for flexible representations in animal learning is still questionable, there are studies showing that discrimination and generalization are affected by previous experience and in a way that is at least suggestive of flexibility in whether learning proceeds in an elemental or configural fashion (Alvarado & Rudy, 1992; Urcelay & Miller, 2009). Should further evidence of such flexibility continue to emerge, it is clear that both purely configural and purely elemental models will have to account for such findings in some way. It is not our intention to provide such an exposition here, as there are too few results bearing on the issue to do so accurately. However, it is worth pointing out that flexibility in the representations of stimuli does not necessarily imply a shift from one kind of learning to another. The representations in the current model *are* flexible in the sense that the behavior of individual elements changes as a consequence of the other elements with which it competes and also as a consequence of the formation of associations between elements. Consequently, there are ways in which prior experience may lead to differences in the manner in which non-linear discriminations are solved.

One way in which the current model *could* become adaptive in the way it solves discriminations centers on how experience with certain discriminations affects the competitive processes that normalize attention. The influence of the attention units on both E and I activation allows for a wide range of non-linear discriminations to be solved. Thus gain control within the network of A units, which dictates which elements receive attention and which do not, is important in determining how effectively the model solves such discriminations. Two parameters that we have not manipulated in these simulations influence the degree to which the attention network modulates the representation of a stimulus, and the change in the stimulus' representation when presented alone versus in compound. First, the power parameter in Equation 5.1 (currently set at p = 5) determines the slope of the sigmoid function (see left panel of Figure 2) and thus dictates the degree to which each A unit operates in a threshold-like manner. When p is small, the function is relatively flat and all the A units of a given stimulus will be slightly more active on single-stimulus presentations than on compound presentations (compare functions in the right panel of Figure 2). When p is high, the function becomes more step-like, such that the A units with intermediate input are substantially affected by the total number of A inputs. Specifically, the intermediate units are much more strongly activated when the stimulus is presented on its own than when presented as part of a compound. Thus a shift from low to high values of p will influence the linearity of the attention network, and thereby determine how sensitive the stimulus elements (Es) are to the presence or absence of other stimuli.

The second variable that affects the behavior of the attention network is the parameter w in Equation 5.2. This weighting dictates the degree to which the A units suppress each other; in effect it controls the capacity of attention. When w is low, adding an extra stimulus (i.e. a compound presentation) has little effect on the activation of each A unit of a stimulus, consistent with an attention network with large capacity. Increasing w effectively shrinks the capacity of attention. Under such circumstances, adding an extra stimulus can have a profound effect, particularly on the activation of the intermediate A units of a stimulus. Once again, given that the activation of A units influences normalization of the E elements, changes in w will affect how dramatically the activation of certain elements changes between single-stimulus and compound presentations. Therefore, if the present model were to incorporate a mechanism by which experience produces longer-term changes in either p or w, such changes could easily lead to an increase or decrease in the linearity of generalization between single and compound stimuli, even for stimuli that have not previously been presented. We discuss these possibilities simply to highlight the fact that a model of this type, despite its purely elemental structure, has potential for the sort of flexibility in representational processes that are perhaps evident in human causal learning and may even be present in other animals.

Author Note

This work was supported by a grant to JAH (DP0771154) and a postdoctoral fellowship to EJL from the Australia Research Council.

Note—This article is one of an occasional series of reviews invited by the editors.

References

Aguado, L., Symonds, M., & Hall, G. (1994). Interval between preexposure and test determines the magnitude of latent inhibition: Implications for an interference account. *Animal Learning and Behavior, 22*, 188-194.

Alvarado, M. C., & Rudy, J. W. (1992). Some properties of configural learning: an investigation of the transverse-patterning problem. *Journal of Experimental Psychology: Animal Behavior Processes, 18*, 145-153. 10.1037/0097-7403.18.2.145.

Arabzadeh, E., Clifford, C. W. G., & Harris, J. A. (2008). Vision merges with touch in a purely tactile discrimination. *Psychological Science, 19*, 635-641. 10.1111/j.1467-9280.2008.02134.x.

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. B. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 121-268). New York: Wiley.

Aydin, A., & Pearce, J. M. (1995). Summation in autoshaping with short- and long-duration stimuli. *Quarterly Journal of Experimental Psychology, 48B*, 215-234.

Aydin, A., & Pearce, J. M. (1997). Some determinants of response summation. *Animal Learning & Behavior, 25*, 108-121.

Bellingham, W. P., Gillette-Bellingham, K., & Kehoe, E. J. (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Animal Learning & Behavior, 13*, 152-164.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning.

Psychological Bulletin, 114, 80-99. 10.1037/0033-2909.114.1.80.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review, 58*, 413-423. 10.1037/h0054576.

Channell, S., & Hall, G. (1983). Contextual effects in latent inhibition with an appetitive conditioning procedure. *Animal Learning & Behavior, 11*, 67-74.

Crowder, N. A., Price, N. S. C., Hietanen, M. A., Dreher, B., Clifford, C. W. G., & Ibbotson, M. R. (2006). Relationship between contrast adaptation and orientation tuning in V1 and V2 of cat cisual cortex. *Journal of Neurophysiology, 95*, 271-283. 10.1152/jn.00871.2005.

Estes, W. K. (1950). Towards a statistical theory of learning. *Psychological Review, 57*, 94-107. 10.1037/h0058559.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*, 123-124.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, 52*, 213-257.

Hall, G. (1991). *Perceptual and associative learning*. New York: Clarendon Press/Oxford University Press.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review, 113*, 584-605. 10.1037/0033-295X.113.3.584.

Harris, J. A. (in press). The arguments of associations. In N. A. Schmajuk (Ed.), *Computational models of conditioning*: Cambridge University Press.

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

Harris, J. A., & Livesey, E. J. (2008). Comparing patterning and biconditional discriminations in humans. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 144-154. 10.1037/0097-7403.34.1.144.

Harris, J. A., Livesey, E. J., Gharaei, S., & Westbrook, R. F. (2008). Negative patterning is easier than a biconditional discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 494-500. 10.1037/0097-7403.34.4.494.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*, 181-197.

Holmes, N. M., & Harris, J. A. (2010). Latent inhibition. In C. M. Mitchell & M. E. Le Pelley (Eds.), *Attention and learning*. Oxford: Oxford University Press.

Honey, R. C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 47(B)*, 83-103.

Kehoe, E. J. (1982). Overshadowing and summation in compound stimulus conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology: Animal Behavior Processes, 8*, 313-328. 10.1037/0097-7403.8.4.313.

Kehoe, E. J. (1986). Summation and configuration in conditioning of the rabbit's nictitating membrane response to compound stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 12*, 186-195. 10.1037/0097-7403.12.2.186.

Kehoe, E. J., & Graham, P. (1988). Summation and configuration: stimulus compounding and negative patterning in the rabbit. *Journal of Experimental Psychology: Animal Behavior Processes, 14*, 320-333. 10.1037/0097-7403.14.3.320.

Kehoe, E. J., Horne, A. J., Horne, P. S., & Macrae, M. (1994). Summation and configuration between and within sensory modalities in classical conditioning of the rabbit. *Animal Learning & Behavior, 22*, 19-26.

Kinder, A., & Lachnit, H. (2003). Similarity and discrimination in human Pavlovian conditioning. *Psychophysiology, 40*, 226-234. 10.1111/1469-8986.00024.

Konorski, J. (1948). *Conditioned reflexes and neuron organization*. Cambridge: Cambridge University Press.

Lachnit, H., Schultheis, H., Konig, S., Ungor, M., & Melchers, K. G. (2008). Comparing elemental and configural associative theories in human causal learning: A case for attention. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 303-313. 10.1037/0097-7403.34.2.303.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B*, 193-243. 10.1080/02724990344000141.

Liljeholm, M., & Balleine, B. (2008). It's elemental my dear Watson. *Behavioural Processes, 77*, 434-436. 10.1016/j.beproc.2007.09.007.

Livesey, E. J., & Harris, J. A. (2008). What are flexible representations?: Commentary on Melchers, Shanks and Lachnit. *Behavioural Processes, 77*, 437-439. 10.1016/j.beproc.2007.09.006.

Lovibond, P. F., Preston, G. C., & Mackintosh, N. J. (1984). Context specificity of conditioning, extinction, and latent inhibition. *Journal of Experimental Psychology: Animal Behavior Processes, 10*, 360-375. 10.1037/0097-7403.10.3.360.

Lubow, R. E., & Weiner, I. (Eds.). (2010). *Latent Inhibition: Data, theories, and applications to schizophrenia*. New York: Cambridge University Press.

Lubow, R. E., Weiner, I., & Schnur, P. (1981). Conditioned attention theory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 15, pp. 1-49). San Diego, CA: Academic Press.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London, UK: Academic Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298. 10.1037/h0076778.

McLaren, I. P. L. (2008). The flexibility thesis: A critique—Commentary on Melchers, Shanks and Lachnit. *Behavioural Processes, 77*, 440-442. 10.1016/j.beproc.2007.09.012.

McLaren, I. P. L., & Dickinson, A. (1990). The conditioning connection. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences, 329*, 179-186.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent Inhibition and perceptual learning. *Animal Learning & Behavior, 28*, 211-246.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*, 177-200.

Melchers, K. G., Shanks, D. R., & Lachnit, H. (2008). Stimulus coding in human associative learning: Flexible representations of parts and wholes. *Behavioural Processes, 77*, 413-427. 10.1016/j.beproc.2007.09.013.

Myers, K. M., Vogel, E. H., Shin, J., & Wagner, A. R. (2001). A comparison of the Rescorla-Wagner and Pearce models in a negative patterning and a summation problem. *Animal Learning & Behavior, 29*(1\), 36-45.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* (G. V. Anrep, Trans.). New York: Dover.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*, 61-73. 10.1037/0033-295X.94.1.61.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101*, 587-607. 10.1037/0033-295X.101.4.587.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior, 30*, 73-95.

Pearce, J. M., Aydin, A., & Redhead, E. S. (1997). Configural analysis of summation in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes, 23*, 84-94. 10.1037/0097-7403.23.1.84.

Pearce, J. M., Esber, G. R., George, D. N., & Haselgrove, M. (2008). The nature of discrimination learning in pigeons. *Learning & Behavior, 36*, 188-199. 10.3758/LB.36.3.188.

Pearce, J. M., & George, D. N. (2002). The effects of using stimuli from three different dimensions on autoshaping with a complex negative patterning discrimination. *Quarterly Journal of Experimental Psychology, 55B*, 349-364. 10.1080/02724990244000061.

Pearce, J. M., George, D. N., & Aydin, A. (2002). Summation: Further assessment of a configural theory. *Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology, 55B*, 61-73. 10.1080/02724990143000171.

Pearce, J. M., & Redhead, E. S. (1993). The influence of an irrelevant stimulus on two discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 19*, 180-190. 10.1037/0097-7403.19.2.180.

Redhead, E. S. (2007). Multimodal discrimination learning in humans: Evidence for configural theory. *Quarterly Journal of Experimental Psychology, 60*, 1477-1495.

Redhead, E. S., & Pearce, J. M. (1995a). Similarity and discrimination learning. *Quarterly Journal of Experimental Psychology. B, Comparative & Physiological Psychology, 48B*, 46-66.

Redhead, E. S., & Pearce, J. M. (1995b). Stimulus salience and negative patterning. *Quarterly Journal of Experimental Psychology. B, Comparative & Physiological Psychology, 48*, 67-83.

Rescorla, R. A. (1972). "Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative & Physiological Psychology, 79*, 307-317. 10.1037/h0032553.

Rescorla, R. A. (1973). Evidence for a unique stimulus interpretation of configural conditioning. *Journal of Comparative and Physiological Psychology, 85*, 331-338.

Rescorla, R. A. (1997). Summation: Assessment of a configural theory. *Animal Learning & Behavior, 25*, 200-209.

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 428-438. 10.1037/0097-7403.26.4.428.

Rescorla, R. A. (2001). Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *Quarterly Journal of Experimental Psychology. B, Comparative & Physiological Psychology, 54*, 53-68. 10.1080/02724990042000038.

Rescorla, R. A. (2002a). Comparison of the rates of associative change during acquisition and extinction. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 406-415. 10.1037/0097-7403.28.4.406.

Rescorla, R. A. (2002b). Effect of following an excitatory-inhibitory compound with an intermediate reinforcer. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 163-174. 10.1037/0097-7403.28.2.163.

Rescorla, R. A., & Coldwell, S. E. (1995). Summation in autoshaping. *Animal Learning & Behavior, 23*(3), 314-326.

Rescorla, R. A., Grau, J. W., & Durlach, P. J. (1985). Analysis of the unique cue in configural discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 11*, 356-366. 10.1037/0097-7403.11.3.356.

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. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience, 27*, 611-647. 10.1146/annurev.neuro.26.041002.131039.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61*, 168-185. 10.1016/j.neuron.2009.01.002.

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation, 6*, 314-326. 10.1016/0023-9690%2875%2990012-0.

Solomon, J. A. (2009). The history of dipper functions. *Attention, Perception, & Psychophysics, 71*, 435-443. 10.3758/APP.71.3.435.

Spence, K. W. (1952). The nature of the response in discrimination learning in animals. *Psychological Review, 59*, 89-93. 10.1037/h0063067.

Spence, K. W. (1956). *Behavior theory and conditioning*. New Haven: Yale University Press.

Stevens, S. S. (1962). The surprising simplicity of sensory metrics. *American Psychologist, 17*, 29-39. 10.1037/h0045795.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135-171. 10.1037/0033-295X.88.2.135.

Symonds, M., & Hall, G. (1995). Perceptual learning in flavor aversion conditioning: Roles of stimulus comparison and latent inhibition of common elements. *Learning and Motivation, 26*, 203-219.

Thein, T., Westbrook, R. F., & Harris, J. A. (2008). How the associative strengths of stimuli combine in compound: summation and overshadowing. *Journal of Experimental Psychology: Animal Behavior Processes, 34*, 155-166. 10.1037/0097-7403.34.1.155.

Urcelay, G. P., & Miller, R. R. (2009). Potentiation and overshadowing in Pavlovian fear conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 35*, 340-356. 10.1037/a0014350.

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

Wagner, A. R. (2003). Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology, 56B*, 7-29. 10.1080/02724990244000133.

Wagner, A. R. (2008). Evolution of an elemental theory of Pavlovian conditioning. *Learning & Behavior, 36*, 253-265. 10.3758/LB.36.3.253.

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, USA: Lawrence Erlbaum Associates, Inc.

Westbrook, R. F., Jones, M. L., Bailey, G. K., & Harris, J. A. (2000). Contextual control over conditioned responding in a latent inhibition paradigm. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 157-173. 10.1037/0097-7403.26.2.157.

Whitlow, J. W., & Wagner, A. R. (1972). Negative Patterning in classical conditioning: summation of response tendencies to isolable and configural components. *Psychonomic Science, 27*, 299-301.

Woodbury, C. B. (1943). The learning of stimulus patterns by dogs. *Journal of Comparative Psychology, 35*, 29-40. 10.1037/h0054061.

[i] Pearce et al. point out that the Pearce (1994) model can predict a successful solution to the discrimination if the model's similarity rule is modified to reduce generalization between input configurations (as suggested earlier by Kinder & Lachnit, 2003). However, such a change will also reduce the model's ability to account for summation of responding when two CSs are presented as a compound, something that the Pearce model already struggles to do (see the next section on Summation).