Published in: N. A. Schmajuk (Ed) *Computational models of conditioning*. Cambridge: Cambridge University Press. (pp 53-70) 2010.

The arguments of associations

Justin A. Harris

School of Psychology, The University of Sydney, Australia

Abstract

This chapter considers associative solutions to "non-linear" discrimination problems, such as negative patterning (A+ and B+ vs AB-) and the biconditional discrimination (AB+ and CD+ vs AC- and BD-). It is commonly assumed that the solution to these discriminations requires "configural" elements that are *added* to the compound of two stimuli. However, these discriminations can be solved by assuming that some elements of each stimulus are *suppressed* when two stimuli are presented in compound. Each of these approaches can solve patterning and biconditional discriminations because they allow some elements, as the arguments of associations, to have differential "presence" on reinforced versus non-reinforced trials, and thus differential associability and control over responding. The chapter then presents a more specific version of one of these models, describing how interactions between stimuli, particularly the competition for attention, provide a mechanism whereby some elements are more suppressed than others when stimuli are presented simultaneously as a compound.

Most computational models of conditioning adopt associative strength (V) as the variable that tracks learning about the association between a conditioned stimulus (CS) or action and the reinforcing unconditioned stimulus (US). Many of these models make very simple assumptions about the arguments of associations – the CSs and USs themselves. For example, the Rescorla-Wagner model treats these stimuli as singular units such that, during learning, a single connection strengthens between the CS unit and US unit (Rescorla & Wagner, 1972; Allan R. Wagner & Rescorla, 1972). While the Rescorla-Wagner model has proved a successful account of the algorithms that underlie many aspects of learning, its simple treatment of the stimuli involved in conditioning has not equipped it to explain a number of important findings. Two particular pieces of evidence that will be considered here are the demonstrations that animals can learn, albeit with difficulty, to master negative patterning and biconditional discriminations. In the simplest case of a negative patterning discrimination, two distinct CSs, A and B, are each individually paired with reinforcement (+), and these trials are intermixed among trials in which the compound of the same two CSs, AB, is presented without reinforcement (-). Many different species in many different paradigms have successfully learned this discrimination, responding more on A+ and B+ trials than on AB- trials (Kehoe & Graham, 1988; Pavlov, 1927; Rescorla, 1972, 1973; Whitlow & Wagner, 1972). The biconditional discrimination represents an even more difficult task in which four CSs are combined as two compounds (AB and CD) that are both reinforced, while on intermixed trials the same four stimuli are presented as different pairwise combinations (AC and BD) but these compounds are not reinforced. Again, there are demonstrations that animals can learn this discrimination, responding more on AB+ and CD+ trials than on AC- and BD- trials (Rescorla, Grau, & Durlach, 1985; Saavedra, 1975), although this appears to pose even greater difficulty than the negative patterning discrimination (Harris & Livesey, 2008; Harris, Livesey, Gharaei, & Westbrook, 2008).

The difficulty for models like Rescorla-Wagner derives from the way they treat the generalisation of associative strength between single CSs and their compounds, or, in the case of the biconditional discrimination, between different compounds composed of the same stimuli. The Rescorla-Wagner model makes the simple assumption that associative strengths are additive between CSs, an assumption at the heart of its common-error term that has been instrumental in providing an account of many conditioning phenomena, from blocking and overshadowing to conditioned inhibition (see Le Pelley, in press, Chapter X in current volume). Indeed, the numerous demonstrations of response summation (e.g., Kehoe, 1982, 1986; Rescorla, 1997; Thein, Westbrook, & Harris, 2008) support the assumption that most of the associative strength of two CSs generalises to their combined presentation as a compound. However, if associative strength reliably generalises between CSs and their compounds, there will always be a consistent ordering in the level of responding shown to single CSs and their compounds. Typically, responding to the compound will be greater than that to the single CSs because the summed associative strengths of the CSs in the compound will be greater than the strength of either individual CS. As such, animals could never learn to respond less to a compound than to its individual CS components in a negative patterning discrimination, or could never learn to respond differentially to the different compounds in a biconditional discrimination.

Configural solutions to non-linear discriminations

The solution to negative patterning and biconditional discriminations requires non-symmetrical generalisation of associative change between single CS and compound trials, or

between compounds composed of common stimulus components. For example, an animal learning a negative patterning discrimination must acquire associative strength during A+ and B+ trials that does not generalise to the AB compound, or else it must acquire inhibitory strength on AB- trials that does not generalise to A and B individually. In the biconditional case, some excitatory associative strength acquired on AB+ and CD+ trials must not generalise to AC and BD, or inhibitory strength acquired on AC- and BD- trials must not generalise to AB and CD. A computationally expedient way to achieve this is to allow an associative unit to be present during reinforced but not non-reinforced trials, or vice versa. Spence (1952) pointed out that this could be achieved by assuming that a compound of two CSs is more than the sum of the individual stimuli, in that the *configuration* of the two stimuli is itself represented as an added element (i.e., there is a computational unit that stands for the conjunction of two or more stimuli). Rescorla (1972, 1973) and Whitlow and Wagner (Whitlow & Wagner, 1972) showed how these added configural units can be incorporated into the Rescorla-Wagner model to provide a ready solution to negative patterning and biconditional discriminations. In negative patterning, the compound configural unit acquires inhibitory strength that suppresses responding on AB- trials, but this inhibition does not generalise to A+ and B+ trials because the configural unit is absent on those trials. Similarly, in a biconditional discrimination, configural units for AB and CD acquire excitatory associative strength on reinforced trials that does not generalise to AC and BD, whereas configural units for AC and BD acquire inhibitory strength on non-reinforced trials that does not generalise to AB and CD (Saavedra, 1975).

The configural hypothesis described above does not successfully accommodate findings from a number of more recent experiments on negative patterning discriminations (see Pearce, 1994, for review). Such evidence has informed alternative descriptions of configural representations. For example, Wagner and Brandon (2001) have proposed that compound configural units are not simply added to the arguments of associations, but that these units *replace* some elements of the component stimuli. Therefore, in negative patterning, some units are present in the compound but absent from single-CS trials, whereas other units are present on single-CS trials but absent from the compound. This facilitates learning because associative strength acquired to the latter units during A+ and B+ trials does not generalise to AB- trials, as well as allowing inhibitory strength to be acquired by the compound configural unit that does not generalise to A+ and B+ trials.

The configural approach has been incorporated in layered network models which place configural units within a hidden layer between input and output layers (e.g., Kehoe, 1988; Schmajuk & di Carlo, 1992). The behaviour of these hidden configural units differs from that of traditional models in which configural units are added to the representation of stimuli. In the earlier models, configural units have a fixed and predefined relation to the stimulus inputs (configural unit AB is necessarily activated by A and B in compound), whereas hidden configural units have adaptive relationships to the sensory input, allowing them to be "tuned" to specific combinations of inputs. This provides the hidden configural units with greater flexibility, allowing them to contribute to learning phenomena beyond solving non-linear discriminations, such a learning-to-learning (Kehoe, 1988) and occasion-setting (Schmajuk, Lamoureux, & Holland, 1998).

A quite different approach has been described by Pearce (1987, 1994). He has argued that all CSs, be they single stimuli or compounds, are represented by a single configural unit that codes for the entire pattern of sensory input. This model assumes that only one configural unit undergoes

associative change on a given trial, and therefore one configural unit acquires excitatory associative strength on reinforced trials, whereas a different configural unit acquires inhibitory strength on non-reinforced trials. Since the strength of activation of each configural unit on a given trial is proportional to the overlap between the current pattern of sensory input and the pattern of input coded by that unit, the associative strength acquired on reinforced trials generalises only partially to non-reinforced trials and the inhibition acquired on non-reinforced trials generalises only partially to reinforced trials.

The models described above have in common their use of configural representations to solve negative patterning and biconditional discriminations. Indeed, it has become common for students of associative learning to consider these discriminations as "proof" of the existence of configural representations because it is assumed that a solution to these discriminations can only be achieved by configural units. A key objective of this chapter is to show that this assumption is false. As described earlier, the solution to negative patterning and biconditional discriminations requires asymmetrical generalisation of excitatory and inhibitory associative strength between reinforced and non-reinforced trials. For negative patterning, this can be achieved by assuming that some elements, such as configural representations, are present on compound trials but not single-stimulus trials. However, it can also be achieved by assuming that some elements are present on single-stimulus trials but not on compound trials. Similarly, the biconditional discrimination can be solved if different configural units are present for each different compound, but it can also be solved if different stimulus elements are lost (or reduced) for the different compounds. Thus each approach can provide a successful computational solution to these discriminations. The remainder of this chapter will be given over to describing computational models that do not involve configural representations.

Elemental solutions to non-linear discriminations

Two formal computational models have been proposed that can solve negative patterning and biconditional discriminations without invoking configural representations. Both assume that the arguments of associations are arrays of elemental units that represent the multiple features of a given stimulus or stimuli. One of these models, proposed by McLaren and Mackintosh (2000, 2002), assumes that all stimuli share a large proportion (at least 50%) of their elements in common, and it is these common elements that provide the solution to negative patterning and biconditional discriminations. An important feature of this model is the non-linear relationship between elemental activation strength and the input strength of the corresponding feature (the relationship is modeled as a steep sigmoid or even step function; see Figure 1). The nature of this function means that features that are weakly present in both stimuli may become strongly represented in the compound, even though they are virtually absent from the representation of each stimulus individually. Consider, for example, a feature, X, with weak input in stimulus A and stimulus B, shown as points X_A and X_B in Figure 1. The element corresponding to X will be virtually inactive when either A or B is present on its own. However, when both stimuli are presented together, the inputs from each stimulus will sum, reaching point X_{AB} which is located on the ascending part of the function and thus achieves a significant level of activation of that element. This constitutes a mechanism whereby arguments of associations may be present during a compound but effectively absent during the individual stimuli. (Note that, while this mechanism does not describe these elements as configural,

in that they do not specifically code for the conjunction of two or more stimuli, their computational behaviour is equivalent to that of added configural units in the Rescorla-Wagner model.)

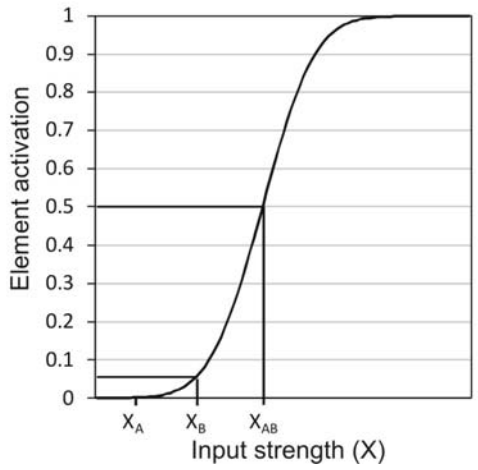

Figure 1. The sigmoidal relationship between the input (the physical intensity of a stimulus feature) and the resultant strength of activation of the corresponding elemental unit, as assumed in the McLaren-Mackintosh model (2002). Note that a feature (X) with low input strength in each of two stimuli (X_A and X_B in this example) provokes very little activation of the corresponding element when either stimulus is present alone. However, when the same two stimuli are presented together, the sum of those physical inputs (X_{AB}) can provoke strong activation of that element, far exceeding the simple sum of the activations from each single stimulus.

The McLaren-Mackintosh model has a second means for solving patterning and biconditional discriminations. Any element whose feature is strongly present in both individual stimuli will be activated during presentations of each individual stimulus and will also be activated during presentations of the compound. However, if the input in each case falls on the upper, flat portion of the curve, the strength of the element's activation will be the same for each single stimulus and for the compound. When associative strength of each element is updated using a common delta-rule, such as that proposed by Rescorla and Wagner (1972), such elements can contribute to solving patterning and biconditional discriminations because they have a higher rate of reinforcement than do other elements (those present in one stimulus but not in the other). For example, in a negative patterning schedule with equal numbers of A+, B+ and AB- trials, any element, X, that is common to both A and B will be present on all trials and reinforced on $2/3^{rds}$ of those trials. By contrast, any element, Y, that is present in only one of the two stimuli will be present on $2/3^{rds}$ of all trials but reinforced on only half of these trials. Consequently, because X has a higher reinforcement rate than Y, it will acquire greater excitatory associative strength and Y will eventually acquire inhibitory strength. Errorless performance is achieved on the discrimination when the associative strength of X elements is 2λ and the associative strength of Y elements equals $-\lambda$.

The second computational model that does not invoke configural representations was proposed by Harris (2006). At the heart of this model, non-linear activation of stimulus elements is

created by a limited-capacity attention mechanism that boosts the activation strength of those elements that have successfully entered attention. The multiple elements within and between stimuli compete for access to attention, with the more salient elements (those with high input) winning over the less salient elements. This competition for attention means that some elements in each individual stimulus lose activation strength when the stimulus is presented as part of a compound. Specifically, some elements benefit from the attention boost when their stimulus is present on its own, but lose that boost when the same stimulus is presented as part of a compound because the increase in total number of elements increases competition for attention.

The decline in activation of some elements when a stimulus is presented in a compound provides a mechanism for a variety of cue interaction effects, such as one-trial overshadowing (James & Wagner, 1980; Mackintosh & Reese, 1979), external inhibition (Brimer, 1970; Pavlov, 1927), and incomplete summation between CSs (Thein et al., 2008). In the latter two cases, the decrease in responding is attributed to a decline in activation of some elements because responding is modeled as the sum of each element's activation multiplied by its associative strength. Of course, in many such cases, decrements in responding could reflect performance interactions, such as when the orienting response to a novel stimulus interferes with performance of conditioned responses (CRs) to a CS. However, there is evidence that one CS may reduce the CR produced by another CS even when the CRs themselves do not compete for expression in behaviour. A particularly clear illustration is provided in an unpublished experiment by Robert Polewan (2006). In an eyelid conditioning experiment, rabbits were trained with two CSs, a light and tone, at different CS-US intervals: 300 ms for the light, 700 ms for the tone. Conditioning at these different latencies gives rise to temporally distinct CR waveforms, with the conditioned eyeblink response developing earlier after onset of the 300-ms CS than the 700-ms CS. After such CRs were established to the two CSs, the rabbits were given probe tests in which the light and tone were presented together as a compound. If the associative strength of each CS were effectively expressed on these compound trials, the waveform of the eyeblink CR should have two peaks, corresponding to the two original CR waveforms from each individual CS. However, Polewan found that the waveform on these compound probe trials showed only a single peak that was closer to the peak response to the tone, and in general the early CR (to the light) was suppressed. The is not due to a limitation in the temporal dynamics of the eyeblink response, because other experiments have shown that rabbits can produce bimodal CR waveforms, with both early and late peaks, when trained with a mixture of two CS-US intervals (Choi & Moore, 2003; Millenson, Kehoe, & Gormezano, 1977), including a mixture of 300 and 700 ms intervals (Polewan, 2006). As Polewan suggested, it was as if the "rabbits ignored the light and focused their attention to the tone on compound trials, resulting in TL- waveforms that resembled T- waveforms." (pp 90-91). This specific direction of this interaction is likely due to the greater salience of the tone than the light. The fact that one CS can interfere with the CR to another CS, even when the CRs to each CS do not themselves interfere with one another, is consistent with the model proposed by Harris (2006) in that the presence of the more salience tone would steal attention from the light, thereby reducing activation of the light elements and in turn reducing the ability of those light elements to associatively activate the US elements at the appropriate time.

The difference in activation strength of certain elements when a stimulus is presented alone versus a part of a compound, combined with a common-error-term learning rule, allows these elements to solve negative patterning. That is, because these elements are strongly activated on

reinforced trials but weakly activated on non-reinforced trials, they ultimately acquire most of the associative strength, while the elements that receive attention on both single-CS and compound trials become inhibitory. However, this mechanism is less readily equipped to solve the biconditional discrimination. To do so, it must assume there are some differences between the different compounds in the level of competition for attention. Such differences can arise from differences in the salience of the different stimuli themselves, even differences that are idiosyncratic to individual subjects due to variations in their sensitivity to the different stimuli. For example, if stimulus A is less salient than stimulus D, B will have steeper competition for attention on BD- trials than on AB+ trials. Therefore, some elements of B will receive the attention boost in activation on AB+ trials but will not receive this boost on BD- trials. As a result, these elements will acquire excitatory associative strength that will produce greater responding on AB+ trials than BD- trials. At the same time, C will have steeper competition for attention on CD+ trials than on AC- trials, and therefore some elements of C will receive boosted activation on AC- trials but not on CD+ trials. Thus these elements will acquire inhibitory strength that will reduce responding on AC- trials relative to CD+ trials. While such differences are sufficient for the model to solve the biconditional discrimination, it solves it much more slowly than a negative patterning discrimination. It is worth pointing out that humans and rats also find the biconditional discrimination more difficult than negative patterning (Harris & Livesey, 2008; Harris et al., 2008).

The model proposed by Harris (2006) relies on an attention system that is selective in its action. That is, more salient elements selectively enter attention and thereby receive a multiplicative boost to their activation, while less salient elements that do not compete effectively for attention receive no boost. In the rest of this chapter, I will describe an alternative formulation of the way that attention can modulate the activation of elements. Rather than assuming that attention selectively exerts its effect on some elements and not others, I propose that the elements vary in their sensitivity to attention as an inherent property of their own activation function. The theoretical processes underlying this are ones that have been developed already in the psychophysical and sensory neuroscience literatures. They capture the way that stimuli interact within sensory systems and how attention influences this interaction. Thus the formulation presented here has the advantage of being better grounded in sensory-perceptual research. It also specifies in greater detail the mechanism by which attention operates on the elemental network to create non-linear changes in element activation.

One crucial feature of the operations I describe below is the non-linear relationship between the input strength of a stimulus and the response of the sensory/perceptual system. It has been long known that 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). While this relationship holds for most of any stimulus dimension, 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). The two contrasting psychophysical effects indicate that the relationship between the physical intensity of a stimulus and its perceived intensity is sigmoid. This relationship has been confirmed in numerous experiments using electrophysiological recordings in cats or primates to determine the relationship between the intensity of a stimulus (eg, the contrast of a visual grating) and the response magnitude of neurons tuned to that stimulus (Crowder et al., 2006). As mentioned

earlier, this sigmoid relationship has already been used by McLaren and Mackintosh (2002) to predict non-linear summation between element input strengths and thereby provide a solution to non-linear problems such as negative patterning and biconditional discriminations. Here, however, I explore a very different means by which this relationship can effect the type of inter-stimulus interactions that are required to solve non-linear discriminations.

The formulation I present below derives in large part from computational models of sensory systems that incorporate a normalization model of gain control (Heeger, 1992; Reynolds & Chelazzi, 2004; Reynolds & Heeger, 2009). The approach used here is similar to the network normalisation rules used by Grossberg's (1975) model of attention and associative learning which uses on-centre off-surround shunting inhibition to constrain entire network activity at an upper bound and to quench noise in the network (see also Schmajuk & di Carlo, 1992). Such normalisation also has the advantage that it allows the number of elements (or stimuli) to be increased indefinitely without saturating the network. The present model achieves normalisation by defining the activation strength, R, of a given element as equal to the strength of the sensory input to which that element is tuned, S, subjected to a divisive normalization (or "gain control") that sums across all sensory inputs weighted according to a suppressive field, z. This relationship between S and R for element x is shown in Equation 1.

$$R_x = \frac{S_x}{\sum_{i=1}^{n} z_i S_i} \tag{1}$$

The properties of the suppressive field are similar to those of the receptive field to which an element is tuned, in that some inputs have a greater effect than others on the normalisation of S_x depending on how close they are to X in topographic and featural space. Inputs that are close to X have greater weighting in the inhibitory field, such that $z_x = 1$, and for any input Y, $z_y < 1$. The greater the difference between X and Y, the smaller z_y becomes, and if X and Y are very different sensory inputs (e.g., from different sensory modalities), $z_y = 0$. If all inputs apart from X are held constant, the summed weighted value for all inputs other than X is constant (C). Therefore, to describe how R_x changes across variations in S_x, we can simplify Equation 1 as:

$$R_x = \frac{S_x}{S_x + C} \tag{2}$$

Equation 2 is a monotonically increasing function with an asymptote of 1. It represents a specific instance of a more general relationship expressed in Equation 3, which, as shown in Figure 2 describes a sigmoid function between R_x and S_x, again with an asymptote of 1. The power, p, determines the slope of the curve, and C determines the position of the curve such that R_x equals half its maximum height (0.5) when S_x^p equals C.

$$R_x = \frac{S_x^p}{S_x^p + C} \tag{3}$$

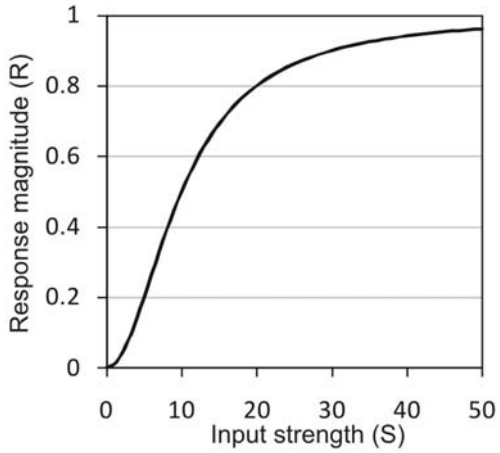

Figure 2. The function relating the strength of a sensory input (S) to the magnitude of the sensory response (R) as described in Equation 3. In this example, p = 2 and C = 100.

In Equation 3, increasing the amount of sensory input by, for example, adding a new stimulus, will increase the value of C. This will shift to the right the function relation R_x to S_x, as shown in Figure 3. The consequence of this will be a reduction in activation of each element, as *per* the normalisation effect. But more specifically, the amount that R_x decreases will depend on S_x, with the greatest drop in R_x for values of S_x^p close to C. Therefore, some elements with strong input will suffer relatively little change in their activation whereas other elements with intermediate input will suffer a substantial decrease in activation strength. It is this differential effect, whereby some elements suffer greater loss of activation than others when their stimulus is presented as part of a compound, which provides a solution to non-linear discriminations such as negative paternning.

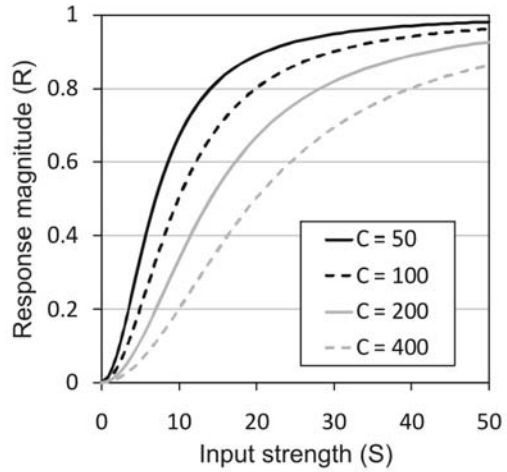

Figure 3. The relationship between the physical intensity of sensory input (S) and the magnitude of the sensory response (R) as described in Eq 3. In each of these examples, p = 2. Note how increasing C shifts the function to the right, and in each case R_x equals 0.5 when S_x^p equals C.

The normalisation process described above captures how stimuli can affect the activation of one another's elements in such a way as to provide a solution to non-linear discriminations. As described, this process relies on the features of each stimulus acting within the supressive field of the other, but this presents a limitation. The mechanism can operate for stimuli from the same modality, but it is less plausible that stimuli from different modalities should act to affect one another in this way. Attention, as an amodal mechanism, provides a means to explain how stimuli from different modalities can afffect one another's activation. Attention is modelled as a spatially and featurally selective field that multiplicatively increases the input strength, S, of a stimulus. Thus if a stimulus or feature, X, captures (or receives) attention its input strength is increased by a gain factor, γ. The consequence of this for R is shown in Equation 4.

$$R_x = \frac{\gamma S_x^p}{\gamma S_x^p + C} \tag{4}$$

This shows the attention gain exclusively applied to S_x. In practice, it is likely to affect some other elements close to X, and therefore have some effect on C. However, as long as the attention field is smaller (more selective) than the suppressive field, most elements in the suppressive field for X will not be in the attention field, and therefore C will increase less than S_x. Therefore, for simplicity, I will allow attention to increase S_x but not C. As such, we can re-write Equation 4 by dividing through by γ to give Equation 5.

$$R_x = \frac{S_x^p}{S_x^p + \frac{C}{\gamma}} \tag{5}$$

Equation 5 makes it clear that attention directed to S_x will effectively reduce C, and therefore shift to the left the function relating S_x to R_x. Conversely, a decrease in attention to S_x will result in the opposite shift.

The operations just described represent the key ingredients of the current proposal. Whenever two stimuli are presented as a simultaneous compound, attention directed to the features of one stimulus reduces (or removes) attention to the features of the other. In effect, attention is simply divided equally among all the stimuli that are present. As a result, the function relating R_x and S_x is effectively shifted to the right, resulting in an overall decrease in their activation strength (R). More importantly, because of the non-linear nature of this function, the rightward shift will have much greater impact on some elements (those for which S^p is close to C) than on others (those with higher values for S). Figure 4 illustrates this point, showing how a rightward shift effected by a decrease in attention can differentially reduce R_x depending on the magnitude of S_x. Thus, one stimulus changes the pattern of element activation of another stimulus, rather than simply scaling the activity uniformly across elements, making a compound qualitatively distinct from its component stimuli.

To confirm that this model can solve negative patterning and biconditional discriminations, Figure 5 plots the average of 50 simulations for both types of discrimination. In these simulations, all stimuli had 10 elements and the activation strength (R) for each element was calculated using Equation 5. The sensory input, S_x, was a random number between 1 and 10, the power, p, was set at 2 and the constant, C, was 4. The loss of attention when any stimulus was combined with a second

stimulus was simulated by defining γ in proportion to the sum of the initial (pre-normalised) values of the second stimulus. This is specified below in Equation 6 for a stimulus, X, when compounded with a stimulus Y containing 10 elements.

$$\gamma_x = \frac{20}{\sum_{i=1}^{10} S_{i,y}} \tag{6}$$

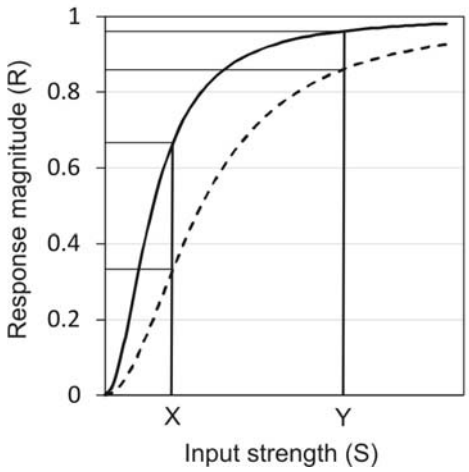

Figure 4. A rightward shift in the function relating S to R produces a large drop in R for values of S close to X (at which, S^p is close to C), but produces only a small change in R for high values of S (such as at Y) or for very low values of S. For points X and Y, the shift from the solid curve to the dashed curve corresponds to a decrease in R of 0.33 and 0.1, respectively. These represent 50% versus 10% reductions.

This simulates a process in which attention is shared between stimuli in proportion to the salience of their elements. Figure 5 shows the predicted conditioned response strength across all 10 elements for each single CS, or 20 elements of each compound. The term R'_γ , defined below in Equation 7, is the activation strength of a US element, y, in response to the summed "internal" input from every other element. As such, the aggregate of R'_γs for all US elements gives the estimated conditioned response strength. On each trial, the associative strength (V) between each CS element (x) and each US element (y) is updated according to a modified version of the Rescorla-Wagner (1972) rule, as defined in Equation 8 with β set at .02. It is worth noting that, by incorporating this learning rule, the model is equipped to deal with the range of empirical findings, such as cue competition effects like blocking, that are explained by the Rescorla-Wagner model.

$$R'_y = \sum_{i=1}^{n} R_i \cdot V_{i-y} \tag{7}$$

$$\Delta V_{x-y} = R_x \cdot \beta \cdot \left(R_y - R'_y\right) \tag{8}$$

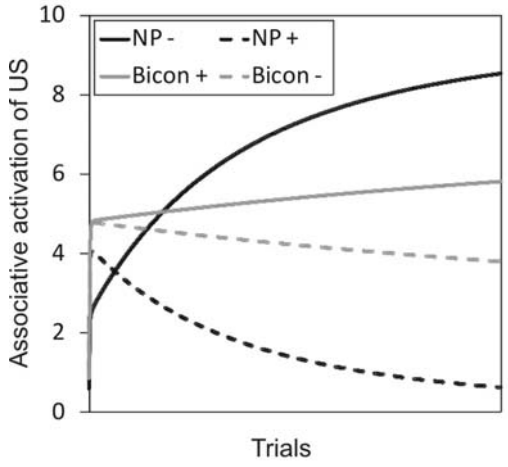

Figure 5. The average of 50 simulations of negative patterning and biconditional discriminations, showing the sum of the products of associative strength and activation strength (R) for each element (the maximum value is 10). Bicon+ and Bicon- refer to reinforced and non-reinforced trials of the biconditional discrimination (i.e., the mean of AB+ and CD+ versus the mean of AC- and BD- trials). NP+ and NP- refer to reinforced (A+ and B+) and non-reinforced (AB-) trials of the negative patterning discrimination.

The operations described above produce differential changes in the activation of elements depending on whether they are part of a single CS or compound. The proposal uses attention because this is a plausible mechanism by which one stimulus could influence the sensory response to one another even when those stimuli are very different, such as from different sensory modalities. However, the model does predict an even greater interaction between stimuli that are more similar due to the fact that such stimuli would not only compete for attention, but could also fall within the suppressive field and thus contribute directly to the normalisation process. That is, for two similar stimuli, the value of C may increase for each stimulus, shifting to the right the function relating S to R.

In conclusion, this chapter considers the nature of stimulus interactions that are required to explain how animals can solve non-linear discriminations such as negative patterning and the biconditional discrimination. While some researchers (e.g., Melchers, Shanks, & Lachnit, 2007) have assumed that these discrimination problems can only be solved by recourse to configural representations that uniquely code stimulus conjunctions, the modeling discussed in the present chapter shows that this is not correct. Non-linear discriminations are intractable to those associative models that assume a one-to-one relationship between the representation of an event and the separate components of that event (e.g., between a compound of two stimuli and the individual stimuli themselves) because these models predict effective generalisation of associative change between reinforced and non-reinforced trials. Viable models of associative learning must assume that stimulus representations involve a non-linear combination of stimulus elements. This can be achieved by adding configural elements to the representation of each compound, or by suppressing the activation of some stimulus elements when stimuli are presented in compound. As such, most complex discriminations can be solved relying solely on elemental representations. Of course the

model formulated here was designed with the express purpose of solving those non-linear discriminations. The mechanisms proposed do not equip the model with the means to explain a range of phenomena that extend beyond the scope of this chapter. Perhaps relevant among these phenomena are learning-to-learn and occasion-setting, given that these can be accounted for by layered network models that incorporating configural representations (Kehoe, 1988; Schmajuk & di Carlo, 1992).

Acknowledgements

The author thanks Nestor Schmajuk, Mike Le Pelley, and an anonymous reviewer for comments on an earlier draft of this chapter. He also thanks John Moore for discussions and for providing the PhD dissertation of Robert Polewan.

References

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

Brimer, C. J. (1970). Disinhibition of an operant response. *Learning and Motivation, 1*, 346-371.

Choi, J.-S., & Moore, J. W. (2003). Cerebellar neuronal activity expresses the complex topography of conditioned eyeblink responses. *Behavioral Neuroscience, 117*, 1211-1219.

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.

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology, 18*, 263-327.

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

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.

James, J. H., & Wagner, A. R. (1980). One-trial overshadowing: evidence of distributive processing. *Journal of Experimental Psychology: Animal Behavior Processes, 6*(2), 188-205.

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.

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.

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

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.

Le Pelley, M. E. (in press). The hybrid modeling approach to conditioning. In N. A. Schmajuk (Ed.), *Computational models of conditioning*: Cambridge University Press.

Mackintosh, N. J., & Reese, B. (1979). One-trial overshadowing. *Quarterly Journal of Experimental Psychology, 31*, 519-526.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent Inhibition and perceptual learning. *Animal Learning & Behavior, 28*(3), 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. (2007). Stimulus coding in human associative learning: Flexible representations of parts and wholes. *Behavioural Processes*.

Millenson, J. R., Kehoe, E. J., & Gormezano, I. (1977). Classical conditioning of the rabbit's nictitating membrane response under fixed and mixed CS-US intervals. *Learning and Motivation, 8*, 351-366.

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*(1), 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*(4), 587-607. 10.1037/0033-295X.101.4.587.

Polewan, R. J. (2006). *Physiological and behavioral studies of rabbit eyeblink conditioning under temporal uncertainty: Purkinje cell response and compound conditioning.* Unpublished doctoral dissertation, University of Massachusetts.

Rescorla, R. A. (1972). "Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative & Physiological Psychology, 79*(2), 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*(2), 200-209.

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.

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.

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

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation, 6*, 314-326.

Schmajuk, N. A., & di Carlo, J. J. (1992). A neural network approach to hippocampal functioning in classical conditioning. *Behavioral Neuroscience, 105*, 82-110.

Schmajuk, N. A., Lamoureux, J. A., & Holland, P. C. (1998). Occasion setting: A neural network approach. *Psychological Review, 105*(1), 3-32.

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

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

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

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.

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.

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: application of a theory. In M. S. Halliday & R. A. Boakes (Eds.), *Inhibition and learning* (pp. 301-336). San Diego, CA: Academic Press.

Whitlow, J. W., & Wagner, A. R. (1972). Negative Patterning in classical conditioning: summation of response tendencies to isolable and configural components. *Psychonomic Science, 27*(5), 299-301.