Representation and Generalisation in Associative Systems

M.E. Le Pelley (mel22@hermes.cam.ac.uk)
I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Site
Cambridge CB2 3EB, England

Abstract

This paper examines the nature of stimulus representation in associative learning systems. Specifically, it addresses the issue of whether representation is elemental or configural in nature. We use a human causal learning paradigm, employing contingencies more commonly associated with studies of retrospective revaluation. Whereas most models of retrospective revaluation view it as an entirely elemental process, our results show that it has a configural component. However, the results also prove troublesome for simple configural theories employing fixed generalisation coefficients. It is possible to explain the data using an elemental theory employing configural representation. Our favoured explanation, however, involves a configural theory employing adaptive generalisation. We present such a theory, APECS, and show through simulation that it is well-equipped to deal with our findings.

Introduction

Recent years have seen a great deal of debate concerning the nature of stimulus representation in associative learning systems – more specifically, over how stimulus *compounds* should be represented, and how generalisation between similar compounds should be dealt with. Consider, for example, rewarding the compound AB. Elemental theories (e.g. Rescorla & Wagner, 1972; Mackintosh, 1975; Wagner, 1981) propose that such compounds are represented as being comprised of separable A and B elements that gain individual associative strengths. The conditioned responding shown to a particular stimulus compound is then found by simply adding together the individual associative strengths of each of the elements of that compound.

We sought to test this fundamental assumption of elemental theories, using a causal judgment procedure with human subjects. Our experimental design is shown in Table 1. We used an allergy prediction paradigm – participants play a food allergist trying to judge the likelihood that various foods will cause an allergic reaction in a fictional patient. The foods, then, constitute the cues; the allergic reaction is the outcome. Following training, subjects rate how strongly certain individual foods, and compounds of two foods, predict the occurrence of an allergic reaction. These ratings are taken as our measure of associative strength.

We train two stimulus compounds, AB and CD, to be predictors of the outcome, and then in Stage 2 one of the elements of each compound is extinguished. Cues A and C have an identical training history, as do B and D. The question is, what effect does this treatment have on the originally experienced compounds (AB and CD) as opposed to "elementally equivalent" compounds made up of identically trained cues that have never been seen in compound before (AD and BC)?

An elemental theory predicts no difference between the two types of compound. If the associative strength of a compound is given by adding together the strengths of all the separate elements contained in that compound, then whether or not those elements have been seen in compound before should have no effect. The Rescorla-Wagner (1972) model (R-W), for example, states that:

$$\Delta V_A = \alpha_A \beta_{US} \left(\lambda - \sum V \right) \qquad (1)$$

where ΔV_A is the change in associative strength of cue A, α_A represents the salience of cue A, β_{US} represents the salience of the US, λ relates to whether the US is actually present on a trial (taking a positive value if the US is present, 0 if it is not), and ΣV is the summed associative strength of all cues present on a trial. According to R-W, following Stage 1 all of cues A to F will have associative strengths of 0.5λ (ignoring the effect of α and β, which will be equivalent for all the different cues as a result of counterbalancing). In Stage 2, extinction trials will reduce V_A and V_C to 0. According to an elemental rule, the associative strength of a compound is found by summing the associative strengths of all of the elements of that compound. The associative strength of AB will be given by the total of the associative strength of A (0) plus the strength of B (0.5λ), i.e. 0.5λ. Of course, the compounds BC and AD are also both made up of one element with a strength of 0, and the other with a strength of 0.5λ, and so all of the compounds AB, CD, BC and AD should give rise to the same level of conditioned responding, as they are all elementally equivalent.

Subjects also received EF+ trials in Stage 1, with no further training of either cue in Stage 2. Given that neither E nor F is experienced in Stage 2, the associative strength of EF should remain at λ (as $V_E=V_F=0.5\lambda$). Hence R-W predicts that EF should receive a higher rating than the other

Stage 1			**Stage 2**	
AB+ **CD+** **EF+**			**A-** **C-**	
G+	H+	I+	GL+	Q-
J+	K+	L+	IO-	V+
KM-	KN-	LO-	HJ?	W+
LP-	Q-	R-	JP?	X+
S-	T-	U-		

Table 1. Experimental design. Important trials in bold. +: outcome; -: no outcome; ?: exposure trial.

compounds, which should all receive similar ratings.

However, it is important to note that the AB+, A- design we are using is more commonly associated with studies of retrospective revaluation (see Le Pelley & McLaren, in press, for a review). This term is used to describe changes in the associative status of previously trained cues in the absence of those cues. For example, it is typically found that A- trials following AB+ training lead to an increase in the causal efficacy of B, even though B itself is absent on these trials. This is the phenomenon of unovershadowing, and would be revealed in our experiment by higher ratings given to B and D than to E and F, which receive no such revaluation in Stage 2.

Findings of retrospective revaluation are problematic for many theories of associative learning. R-W, for example, states that α, the salience of a cue, is positive for a cue that is actually presented on a trial, and zero for all absent cues. Hence the theory incorrectly predicts that there will be no learning about absent cues. So V_B remains unchanged at 0.5λ during Stage 2, with R-W thus constrained to predict that B, D, E and F will all receive similar ratings on test.

It is possible, however, to adapt R-W to allow it to predict unovershadowing. Van Hamme & Wasserman (1994) proposed that absent cues, rather than having $\alpha=0$, should take on a negative value of α, thus engaging the learning process with a negative sign. So on Stage 2 A- trials, while A's association to the outcome becomes weaker, the association from the absent cue B to the outcome will become correspondingly stronger. Markman (1989) proposed that only absent *but expected* cues should take on negative α. Dickinson & Burke (1996) suggested that this expectancy arises as a result of within-compound associations formed during Stage 1 compound training. During AB+ trials, subjects learn not only that A and B predict the US, but also that A predicts the presence of B, and *vice versa*. Presentation of A on A- trials now creates an expectancy of the absent cue B, and it is this expectancy that imbues it with negative α.

Modified R-W now predicts that B will be rated higher than E and F (which are not revalued) following Stage 2. It also predicts that AB will receive a similar rating to EF. According to unmodified R-W, the rating of AB falls during Stage 2 as A is extinguished and B remains unaffected. Modified R-W, on the other hand, states that as V_A falls (asymptoting at 0), V_B will increase (asymptoting at λ). Given these opposing changes in associative strength, the overall associative strength of the AB compound (given by V_A+V_B) should remain roughly constant.

Note, however, that modified R-W is still an elemental theory. As such it is still constrained to predict that compounds AB and CD will receive the same rating as the elementally equivalent compounds BC and AD.

The other trial types listed in the experimental design are relevant to a different issue in associative learning theory: they are not discussed here. We were careful to ensure equal numbers of positive and negative trial types in each stage. Following Dickinson & Burke (1996), we also made sure that each subject encountered a large number of different trial types (16 in Stage 1, 8 in Stage 2). This creates a large memory load, hopefully preventing subjects from basing their ratings on inferences made from explicit episodic memories of the various trial types. Instead subjects should have to rely on associative processes to provide an "automatic" measure of causal efficacy for each cue. Using a large number of trial types makes us more confident that it is indeed associative, rather than cognitive, processes being tapped in our study.

Method

Participants Sixteen members of Cambridge University (9 female, 7 male; age 19-49) took part in the experiment.

Procedure At the start of the experiment each subject was given a sheet of instructions presenting the "allergy prediction" cover story for the experiment. They were told that in the first block they would be arrange for Mr. X to eat different meals on each day, and would monitor whether he had an allergic reaction or not as a result. In relation to the exposure trials (that do not bear on the issue at hand in this paper), subjects were told that occasionally the results of eating the foods had been lost. On these trials they would know the foods eaten in the meal, but not the result of eating those foods. They were also told that at the end of the experiment they would be asked to rate each of the foods according to how strongly it predicted allergic reactions. The 24 foods used were randomly assigned to the letters A to X in the experimental design for each subject.

On each conditioning trial, the words "Meal [meal number] contains the following foods:" followed by the two foods appeared on the screen. Subjects were then asked to predict whether or not eating the foods would cause Mr. X to have an allergic reaction, using the "x" and "." keys (counterbalanced). The screen then cleared, and immediate feedback was provided. On positive trials the message "ALLERGIC REACTION!" appeared on the screen; on negative trials the message "No Reaction" appeared. If an incorrect prediction was made, the computer beeped. On the exposure trials of Stage 2, the same message appeared, but now subjects were cued to enter the initial two letters of each of the foods. This was to ensure that they paid attention to the pairings of foods when no allergy prediction was required.

There were 16 trial types in Stage 1, and 8 in Stage 2. The order of trials was randomised over each set of 16 or 8. Participants saw each meal 8 times in Stage 1 and Stage 2. The order of presentation on the screen (first/second) within each compound pair was also randomised.

In the final rating stage subjects were asked to rate their opinions of the effect of eating a number of meals containing either one food or two on a scale from -10 to +10 (in fact, subjects were also given a short rating test at the end of Stage 1 – again the results of that test do not bear on the work presented here and so will be ignored.). They were to use +10 if the meal was very likely to cause an allergic reaction in Mr. X, -10 if eating the meal was very likely to prevent the occurrence of allergic reactions which other foods were capable of causing, and 0 if eating the meal had no effect on Mr. X (i.e. it neither caused nor prevented allergic reactions). For clarification, participants also had access to a card on which the instructions on how to use the rating scale were printed. Once a meal had been rated it disappeared from the screen and the next appeared, so that participants could not revise their opinions upon seeing later meals.

Results and Discussion

Figure 1 shows the mean rating of the casual efficacy of each of the meals of interest as judged on test. In this figure, and in the following analysis, the ratings of equivalent cues

(i.e. cues or compounds that have received an identical training history) have been averaged for each subject. Thus we averaged the ratings of AB and CD, BC and AD, A and C, B and D, and E and F. No significant differences existed between equivalent cues [$F_{max}(1,15)=1.97$, $p>0.1$].

A one-way, repeated measures ANOVA was carried out on these ratings as a preliminary to assessing the effects of interest by means of planned comparisons. There was a significant main effect of meal [$F(10,150)=13.98$, $p<0.001$].

In common with earlier studies of retrospective revaluation, we see that B and D are rated higher than E and F on test [$F(1,15)=8.72$, $p<0.01$]. Given that B, D, E and F all received exactly the same number of pairings with the outcome in Stage 1, this finding implies that the ratings of B and D have changed as a result of Stage 2 A- and C- trials. This demonstration of learning about absent cues violates the assumption of the original R-W model that learning can only proceed to cues presented on a trial. However, it is consistent with modified R-W, in which absent-but-expected cues engage the learning process with a negative sign.

The ratings for AB/CD are actually very similar to those received by EF: the difference between them is not significant [$F<1$]. This again disagrees with original R-W, which states that extinction of A should reduce the causal efficacy of the AB compound relative to EF. And again it is consistent with modified R-W, which proposes that as V_A falls on A- trials, V_B rises, such that the rating for AB will remain roughly constant. In further support of this idea we see that the ratings given to B/D do not differ significantly from those given to compounds AB/CD [$F<1$]. From the standpoint of an elemental theory, this finding implies that the associative strength of compound AB is almost entirely due to the strength of cue B, which is the prediction made by modified R-W.

Of most interest, though, is the finding that the ratings for AB/CD (the compounds actually experienced during training) are higher than those for AD/BC, even though all of these compounds are elementally equivalent. This is confirmed statistically [$F(1,15)=10.72$, $p<0.005$]. This finding is troublesome for any theory that proposes that stimuli in an associative network are represented in a wholly elemental manner, as such theories are constrained to predict that AB, CD, BC and AD will receive equal ratings on test.

Thus it is clearly insufficient to view a compound AB as simply being composed of separable A and B elements which gain associative strength independently. Instead it seems that the fact that A and B have been seen together before is important when determining the response to the AB compound, i.e. there is importance attached to the unique *configuration* of A and B cues. This heightened responding to previously experienced configurations will not apply to the BC compound, as B and C elements have never been experienced in configuration during training. Hence if configurations of cues are taken into account we can explain the finding that AB/CD are rated higher than BC/AD.

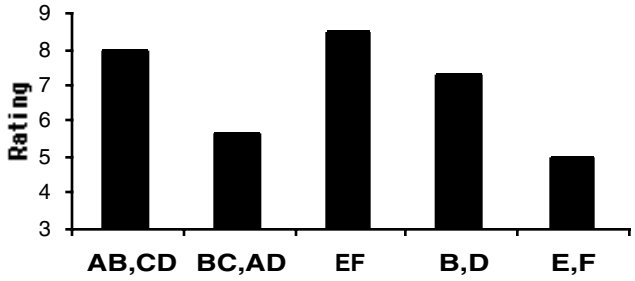

Figure 1. Mean ratings given to the cues of interest.

It is possible to modify R-W even further in order to encompass this importance of specific configurations of cues. Wagner & Rescorla (1972) suggested that, in addition to activating individual cue elements, presenting a compound stimulus also activates a unique element representing that configuration of cues (a "configural element"). Thus presentation of AB will activate elements for A and B, and also an AB element. As regards the learning process, all elements are treated in exactly the same way. If we assume that all elements have equal salience, then following AB+ trials, $V_A=V_B=V_{AB}=1/3\lambda$. Note that this is still very much an elemental theory in nature: the associative strength of a compound is given by summing the individual strengths of all of its elements, whether those elements represent compounds or single cues.

We can combine this with the idea of negative α. On A- trials following AB+ training, elements for both B and AB will have negative α: neither is present, but both are retrieved via within-compound associations. As A extinguishes (until $V_A=0$), B and AB will become more excitatory, such that the overall strength of the AB compound (given by $V_A+V_B+V_{AB}$) remains roughly constant across A- trials. There is no configural element for the BC compound, however, as this configuration has never been experienced during training. Thus the associative strength of the BC compound will be given by (V_B+V_C). Given that this compound does not receive the extra excitatory influence of a configural element, it is bound to receive a lower rating than AB. This "configural element" adaptation of Van Hamme & Wasserman's modified R-W therefore allows us to explain the finding that AB is rated higher than BC.

However, this "double modification" leads to further incorrect predictions. Firstly, it predicts that AB will be rated higher than B. Presentation of AB activates A, AB and B units. The latter two have excitatory connections to the US, and their influence will sum. Presentation of B only activates the B unit, so the excitatory influence will be less. In fact, the ratings for AB and B do not differ. Secondly, it predicts that B will receive a similar rating to BC. Given that $V_C=0$, both will rely on the B–US association for all their excitatory strength. In fact, B/D is rated significantly higher than BC/AD [$F(1,15)=5.64$, $p<0.05$].

Perhaps a consideration of context will help. Presenting USs in a context makes the context itself a weak excitor of the US. In terms of our experiment, subjects come to realise that the patient is quite prone to allergic reactions regardless of which particular foods he has eaten. Cues presented on nonreinforced trials (e.g. A and C) will become weak inhibitors of the US to counter this general excitatory influence. A and C do in fact receive negative ratings on test (mean –1.6), adding weight to this argument. If V_A and V_C are negative then we can resolve the problems outlined above. The predicted rating for AB will fall due to A's inhibitory influence: B's rating will not be affected in this way: AB and B will

receive similar ratings. Furthermore, BC will now be rated lower than B due to C's negative effect.

In summary, it would seem that an elemental rule (a) employing configural elements, and (b) allowing negative learning about absent-but-expected elements, might explain our empirical data, as long as the role of context is also taken into consideration. Figure 2 (black bars) presents simulation results for this experiment using just such a model. Comparison with the empirical data in Figure 1 shows close agreement (r^2=0.98). The exact parameters used are relatively unimportant – we simply note that it is possible for such a model to explain the patterns present in our data.

Another approach to explaining our data is to reconsider the contention of elemental theories that a stimulus compound is composed of separable A and B elements. Recently this idea has been challenged, notably by Pearce (1987), whose configural theory proposes that a compound stimulus is best viewed as a unitary event that is separate from its elements, but able to generalise to them. In other words, it would be a single, "AB" configuration that developed an associative connection to the outcome. Generalised responding to other stimuli occurs to the extent that these stimuli are similar to previously experienced configurations.

Specifically, Pearce's (1987) configural theory states that:

$$\Delta V_X = \beta_{US}(\lambda - V_X) \qquad (2)$$

where V_X (the associative strength of configuration X) is given by the sum of the *conditioned* responding to configuration X and the *generalised* responding to X as a result of its similarity to other trained configurations. The extent to which generalisation occurs between two configurations depends on their similarity:

$$S_{X,Y} = \frac{nc_X}{nt_X} \cdot \frac{nc_Y}{nt_y} \qquad (3)$$

Thus the similarity (S) between configurations X and Y is equal to the proportion of the total elements in configuration X that are common to the two configurations, multiplied by the proportion of the total elements in configuration Y that are common to the two configurations. Then:

$$\begin{matrix}\text{Generalised strength} \\ \text{from Y to X}\end{matrix} = S_{X,Y} \times V_Y \qquad (4)$$

Consider, for instance, compounds AB and BC in our experimental design. Each configuration has two elements, one of which (B) is common. Hence they will have a similarity of 0.25, so any conditioned responding to AB will generalise by a factor of 0.25 to BC.

In our Stage 1, a representation of AB will develop an associative strength of λ (again ignoring β, which will be equal for all configurations). In Stage 2 A is extinguished. A has a similarity of 0.5 to AB, and hence receives generalised strength of 0.5λ from it. In order to counteract this excitatory influence A must itself take on a strength of -0.5λ (to prevent conditioned responding when it is inappropriate).

On test, responding to AB is given by the sum of its own conditioned strength and its generalised strength from A (to which it has a similarity of 0.5). Hence:

$$V_{AB} = \lambda + 0.5(-0.5\lambda) = 0.75\lambda$$

The same holds true for CD. How about responding to AD? This configuration has never been seen before, and hence will receive only generalised strength. It has a similarity of 0.25 to AB and to CD, and a similarity of 0.5 to A. Thus:

$$V_{AD} = 0.25(\lambda) + 0.25(\lambda) + 0.5(-0.5\lambda) = 0.25\lambda$$

Hence this configural theory predicts that responding to AB will be greater than for AD.

This kind of configural theory thus seems tailor-made to explain the different ratings given to elementally equivalent old and new compounds. The problem for a configural theory such as Pearce's is that it cannot explain the occurrence of retrospective revaluation. In common with the original R-W model, Pearce predicts that B's rating should not change as a result of A- trials. Following Stage 1 AB+ trials, V_B should be 0.5λ (as B has a similarity of 0.5 to configuration AB, which will have developed an associative strength of λ). However, given that the compound AB is not seen again, its conditioned associative strength will not change, and so B's rating (which depends on generalisation of excitatory strength from the trained AB compound) will remain unchanged. Is it possible to modify a configural theory such as Pearce's to also explain the phenomenon of retrospective revaluation? The answer at present seems to be no, and we leave it for others to challenge this conclusion.

The problem for such configural rules seems to lie in their use of fixed, non-adaptive generalisation coefficients. The generalisation between two similar stimuli takes a set value that cannot change whether the two stimuli are reinforced or not. An alternative possibility is to use adaptive generalisation coefficients that vary dynamically, such that the generalisation between two similar stimuli can change on a trial-by-trial basis according to whether the two stimuli predict the same or different outcomes. For more on the value of adaptive generalisation coefficients, see McLaren (1993, 1994) and Le Pelley & McLaren (in press).

Consider the AB+, A- contingency used in our experiment. On Stage 1 AB+ trials, subjects learn that when A and B are presented together, the outcome is expected. Following these trials, they have no reason to believe that what holds for A in the presence of B should not hold true for A alone. Hence generalisation to other compounds containing A might be set high as a default. Stage 2 A- trials provide evidence against this idea, though. As a result generalisation between A- and AB+ should be reduced, in order to prevent new learning (that A alone is not reinforced) from interfering with old (that A and B in compound are reinforced). This leaves AB as a good predictor of the outcome, while simultaneously allowing complete extinction of A. The fact that the generalisation between the AB compound and its ele-

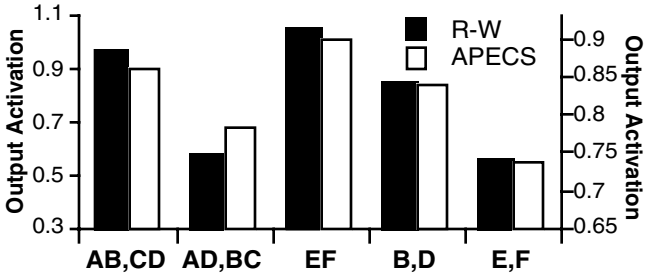

Figure 2. Simulation results for modified R-W with configural elements (black bars) and APECS (white bars).

ments is allowed to change as subjects move from Stage 1 AB+ trials to Stage 2 A- trials provides a mechanism for change in the causal efficacy of B as a result of these A- trials. As it turns out, a model employing configural representation with adaptive generalisation coefficients is indeed well-equipped to explain our empirical data.

A suitable candidate is the APECS model presented by Le Pelley & McLaren (in press). In this instantiation of APECS, each different pattern of stimulation is represented by its own hidden unit, which can equally well be termed "configural units". The mechanics of learning in APECS are similar to those of standard backpropagation (Rumelhart, Hinton & Williams, 1986), but differ in that APECS employs adaptive generalisation coefficients: once the weights appropriate to a mapping have developed, the learning in those weights can be protected against interference. This is achieved by reducing the learning rate parameter for the configural unit carrying the mapping. The effect is to "freeze" the weights to and from a certain configural unit at the value they hold immediately following experience of that configuration. Crucially, this freezing of weights to and from a certain configural unit occurs only if that configural unit has a negative error value, i.e. *if it is part of a mapping that predicts an incorrect outcome for the current input*. Specifically, APECS has different learning rate parameters for input–hidden and bias–hidden connections. The former are frozen to prevent interference; the latter remain high. Hence extinction (suppression of inappropriate responses) is achieved by an increase in the negative bias on the hidden unit carrying the inappropriate mapping, rather than by reduction of weights (which would cause the original mapping to be lost from the network). Given appropriate input cues, the negative bias on the hidden unit can be overcome and the original mapping retrieved.

Consider an AB+, A- contingency. During Stage 1, the network will learn associations from A and B input units to a hidden unit representing the AB configuration. It also learns an excitatory association from this hidden unit to the output: A and B in compound come to predict the outcome.

Now consider the inter-trial interval (ITI) between AB+ trials, when no inputs are presented. According to the logistic activation function employed with APECS, when no inputs are presented the hidden units will have an activation of 0.5 (see Rumelhart et al., 1986). This activation will feed along the AB+ pathway learnt on the preceding trial, and activate the US unit. This is obviously inappropriate when no inputs are presented. The US unit will take on a negative error, which is propagated back to the AB configural unit. As explained earlier, a negative error means that the weights to and from the hidden unit are frozen. In order to suppress the expression of the US during the ITI, the AB configural unit will therefore take on a negative bias.

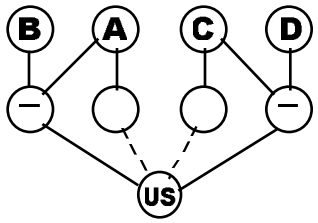

Figure 3. Associations developed by APECS following AB+, CD+ then A-, C- training. Excitatory connections are shown by solid lines, inhibitory associations by dotted lines. Negative bias on hidden units is indicated by a minus sign.

In Stage 2 the network experiences A- trials. Given that this configuration has not been seen before, a new hidden unit is recruited to carry the mapping. Of course, as a result of the associations built up during Stage 1, A has an excitatory connection to the US (via the AB configural unit). However, the US is not presented on these trials: the AB unit carries an inappropriate mapping, and so will take on a negative error. As a result its weights are frozen, and it will take on an increased negative bias in order to suppress expression of the US (i.e. to allow extinction of A). Thus the learning about the mapping from A and B to the output has not been lost from the network, it has simply become harder to retrieve. In addition, an inhibitory mapping will develop from the new A- hidden unit to the outcome in order to counter the positive activation flowing via the AB configural unit. The situation for cues A, B, C and D following Stage 2 training is shown in Figure 3.

Note that the negative bias taken on by the AB configural unit is a result of presentation of A alone leading to inappropriate output activation on A- trials. Now on test both A and B are presented together. Presentation of both cues (each with an excitatory connection to the AB configural unit) will be sufficient to overcome the negative bias built up by this unit, and so the mapping from the configural unit to the US will be expressed as before. In other words, the presence of the extra retrieval cue on test (B, compared to A alone in Stage 2) allows retrieval of the original AB+ mapping. Adaptive generalisation protects the AB+ mapping from the effect of extinction of A. Hence APECS predicts that extinction of A will have little effect on the rating received by AB, and thus that AB and EF will receive similar ratings (as is seen in Figure 1).

What if the compound BC is presented on test? As a result of the processes previously mentioned, C will be completely extinguished and so will not cause any activation of the output. Presentation of B will send some positive activation to the AB configural unit. However, without the additional positive influence of A, B alone will be unable to completely overcome the negative bias on this unit. As a result less positive activation will flow to the US than if A were also present. Thus APECS correctly predicts that AB will receive a higher rating than BC.

How about unovershadowing? Again, APECS explains the phenomenon as being a result of the attempt to minimise interference between old and new learning through adaptive generalisation. We saw that the AB configural unit starts Stage 2 with a reasonable negative bias (built up during ITIs following AB+ trials in Stage 1). This unit then takes on additional negative bias on the initial A- trials of Stage 2. However, if this negative bias is allowed to grow too much then the network will lose the information that B has in the past predicted the US, as presentation of B will be insufficient to impact on this negative bias. This would be an undesirable consequence of learning about A. In order to protect this learning, over the course of Stage 2 A- trials, as the inhibitory connection via the A-only hidden unit becomes stronger, the bias on the AB configural unit lessens.

Thus on initial A- trials, the network achieves extinction of A by suppressing the original excitatory pathway. This makes sense: given the limited evidence for the causal efficacy of A, its failure to predict the US may be a freak occurrence. It is undesirable to lose the information that A predicts the US on the basis of this limited evidence. Extinction by suppression of a pathway allows for rapid reactivation of that pathway should A now come to predict the US again. But with increasing evidence that A genuinely does not predict the outcome, the balance shifts. The original suppression is lifted to prevent loss of information about the other cues that A was trained with, which probably were the cause of the outcome originally. Extinction of A is now achieved more permanently by development of an inhibitory association to the outcome. This is sufficient to balance the increased excitation flowing from A to the US via the now less suppressed AB unit. This lesser suppression of the AB unit, meanwhile, reduces its negative bias to levels below that developed in Stage 1, meaning that presentation of B will now cause greater US activation than E or F (as the EF unit has not undergone this de-suppression). Unovershadowing is the result. For a more detailed discussion of APECS and unovershadowing, see Le Pelley & McLaren (2001).

There is a problem, however. As things stand APECS incorrectly predicts that BC should receive a rating similar to B. We can overcome this problem by considering context, as described earlier, so that A and C become weakly inhibitory. Figure 2 (white bars) shows the results of a simulation of this experiment using APECS. Again, comparing this to Figure 1 reveals close agreement between empirical and simulated data (r^2=0.98). The simulated results are actually the average of 16 simulations run with APECS, each representing a different subject. Each trial involved 1000 learning cycles. A hidden unit is defined as being "active" when it receives positive activation from the input layer. Thus if cue A is presented to the network, any hidden unit representing a configuration that includes cue A will be active. Activity extends into the period immediately following each trial, when no inputs are presented (again for 1000 learning cycles). The learning rate parameters for input–hidden and hidden–output units are both 0.8 when a hidden unit is active and has a positive error, and 0 when it is not. The parameter for bias–hidden changes is 0.3 when a hidden unit is active, 0 when it is not. Thus we make the reasonable assumption that changes due to learning take place faster than changes in memory, i.e. learning represents rapid acquisition, and memory represents a more gradual decline in retrievability. We also included an input unit representing context, that was active on every trial. Context will have a far lower salience than the foods used on each trial: we use a parameter of 0.028 for changes in weights from the context unit. The simulation results are robust under quite large variations in the parameters used.

Conclusion

Retrospective revaluation has typically been assumed to be best explained in terms of changes in the associative strengths of separable stimulus elements. Perhaps unsurprisingly, then, the most influential theories attempting to account for retrospective revaluation (e.g. Van Hamme & Wasserman's [1994] modification of R-W; Dickinson & Burke's [1996] modification of Wagner's [1981] SOP model) have been elemental in nature. The results presented here, however, suggest that this simple elemental view of retrospective revaluation is incorrect. Our data conflict with the fundamental assumption of simple elemental theories that a compound AB is best represented as being composed of separable A and B elements that gain strength independently. Instead there is a configural component involved in retrospective revaluation that is ignored in these earlier theories. It is possible to account for the data using an elemental theory modified to give a role to "configural elements". Alternatively, a model employing configural representation with adaptive generalisation also provides a good account of our results.

References

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

Le Pelley, M.E., Cutler, D.L., & McLaren, I.P.L (2000). Retrospective effects in human causality judgment. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 782-787). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Le Pelley, M.E., & McLaren, I.P.L. (2001). Retrospective revaluation in humans: Learning or memory? *Quarterly Journal of Experimental Psychology,* accepted subject to revision.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General, 118*, 417-421.

McLaren, I. P. L. (1994). Representation development in associative systems. In J.A. Hogan & J.J. Bolhuis (Eds.), *Causal mechanisms of behavioural development* (pp. 377-402). Cambridge: Cambridge University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*, (61-73).

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McLelland & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

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

Wagner, A.R., & Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R.A. Boakes & M.S. Halliday (Eds.), *Inhibition and Learning (pp. 301-336).* New York: Academic Press.

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