Retrospective Effects in Human Causality Judgment

M.E. Le Pelley (mel22@hermes.cam.ac.uk)
D.L. Cutler (dlc29@hermes.cam.ac.uk)
I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Site
Cambridge CB2 3EB, England

Abstract

The phenomenon of retrospective revaluation has been a challenge to many associative learning theories as it involves a change in the associative strength of a cue on trials on which that cue is absent. The present experiment combines several retrospective learning contingencies in a single, within-subjects experiment, allowing for valid comparisons between contingencies. One of the most popular models of retrospective revaluation, Dickinson & Burke's (1996) modification of Wagner's (1981) SOP theory, fails to explain the full pattern of results. A connectionist model that explains retrospective revaluation in terms of changes in retrievability in memory, rather than as new learning about absent cues, is shown to provide a better account of the results.

Introduction

Perhaps the biggest challenge to traditional theories of associative learning in recent years has come from studies of retrospective effects in cue competition. Such effects have overturned the central tenet of many of the most influential learning theories (e.g. Rescorla & Wagner, 1972; Wagner, 1981) – that only cues present on a given trial may engage the learning process.

Consider a typical retrospective revaluation study, as shown in Table 1. Stage 1 involves training of the cue compounds AB and CD to predict some outcome. In Stage 2 one of the cues (the competing cue) from each compound is selected for either further training (in what is known as the backward blocking condition) or extinction (unovershadowing). The typical result of such studies is that, following stage 2 training, when the cues that have not received any further training in stage 2 (the target cues) are tested, D is now rated as a better predictor of the outcome than B. Thus the perceived predictive validity of a cue can be altered after initial compound training with that cue, either by training the other cue of the compound pair as a valid predictor of the outcome (as in backward blocking) or by extinguishing it (as in unovershadowing). The inference from this is that the associative strength of the target cue representation (B or D above) to outcome representation association can change on trials in which that cue is not presented (A+ and C-).

Retrospective revaluation has now been reliably demonstrated in a number of experiments with humans, using causal judgments of a cue→outcome relationship as indicators of the strength of the association between their representations (e.g. Chapman,1991; Dickinson & Burke, 1996; Shanks, 1985). We will consider in some detail here one of the more popular theories of associative learning; Wagner's (1981) SOP model, and Dickinson & Burke's (1998) modification allowing it to explain retrospective effects.

Table 1: A typical retrospective revaluation design.

Condition	Stage 1	Stage 2	Test
Backward Blocking	AB+	A+	B?
Unovershadowing	CD+	C-	D?

SOP proposes that stimuli are represented by nodes in an associative memory that are composed of a number of elements. These elements can be in one of three states at any instant; one inactive state (I) and two active states (A1 and A2). Presentation of a stimulus excites the elements representing that stimulus into A1. These elements then decay back to I via A2. Exciting a node via an associative connection, however, causes a transition from I directly to A2. Changes in the associative connection between two nodes depend on temporal overlap of the states of their elements. Whenever the elements of two nodes are in A1, there is an increment in the excitatory strength between them. When the elements for one node are in A1 and those of another are in A2, there is an increment in the strength of an inhibitory connection. Critically, SOP states that only cue elements in A1 will engage the learning process (i.e. learning will only accrue to cues that are physically present on a trial). Hence, as there can be no learning about absent cues, SOP is unable to explain the results of retrospective revaluation studies.

Dickinson & Burke (1996) proposed a modification to SOP to allow it to explain retrospective revaluation (Table 2). They suggested that CS elements in A2 could engage learning, with an increment in excitatory strength whenever there was an overlap in activation states (be this in A1 or in A2) and an increment in inhibitory strength whenever elements were in different states. Thus they specified the sign with which learning occurs to be a symmetrical function of elemental activation states.

Consider now the contingencies shown in Table 1. During the first stage both target and competing cues, and the US, are presented, and so all will have elements in A1. Hence target and competing cue elements will form excitatory connections to US elements, and within-compound associations will form between target and competing cues. In the unovershadowing contingency cue C is now presented, and will retrieve D elements into A2 via the within-compound link. The US will also have elements in

Table 2: Modified SOP.

		US Element	
		A1	*A2*
CS Element	*A1*	**E**	**I**
	A2	**I**	**E**

E : Excitatory connection strengthened
I : Inhibitory connection strengthened

A2 (retrieved via the C→US connection). As both D and the US have elements in the A2 state, modified SOP predicts an increment in the excitatory strength between them. Hence D's rating is predicted to increase as a result of the C- trials, even though it is absent.

The case for the backward blocking contingency is less clear. Presentation of A in stage 2 will retrieve B elements into A2. The outcome is presented, so some of its elements will be in A1, but it is also predicted by virtue of the A→US connection, so it will also have elements in A2 (these elements cannot go straight from A2 to A1 when the US is presented; they must pass through the I state first). Thus any inhibitory A2-A1 learning between B and the US will be offset to some extent by excitatory A2-A2 learning. Whether the model predicts a net increase or decrease in the rating of B depends on which of the processes engaged by congruent and incongruent elemental states is stronger.

The overall result, though, is that after training D will be rated higher than B. According to modified SOP, then, the driving force behind retrospective revaluation is unovershadowing; backward blocking has a smaller role. This is supported by Larkin, Aitken & Dickinson (1998), who tried to measure the effects of unovershadowing and backward blocking separately by comparing each to a control contingency, EF+ X+ (for which neither target cue nor competing cue is trained in Stage 2). As predicted by modified SOP, they found evidence for a significant effect of unovershadowing, but the evidence for backward blocking was weaker and fell short of significance.

Backward blocking and unovershadowing are not the only retrospective effects that have been found in human causal learning. It has long been known that following A+, AB- training, B will typically become established as an inhibitor of the US, able to counteract the excitatory potential of A. Chapman (1991) reversed this procedure, to give an AB-, A+ design. This procedure was sufficient to establish B as an inhibitor of the US (i.e. it received a lower rating on test than C or D from a CD-, X- control contingency). The inhibitory properties of B must have been assumed in retrospect, as A was only established as a good predictor of the US following AB- trials.

Note that the phenomenon of backward-conditioned inhibition is in line with the predictions of modified SOP. During the first stage a within-compound association is learnt between A and B. A then retrieves B elements into A2 in stage 2. The outcome is presented, and so has elements in A1. The resulting A2-A1 activity will result in formation of an inhibitory link between B and the US.

Thus modified SOP is well equipped to deal with some of the major findings of retrospective learning studies with humans. In the present experiment we use these retrospective effects as a benchmark from which to provide a more critical assessment of the mechanism for retrospective revaluation proposed by modified SOP.

The design of the experiment is shown in Table 3. We used an allergy prediction paradigm, as employed by Dickinson & Burke (1996) and Larkin et al. (1998). Participants play the role of a food allergist trying to judge the likelihood that various foods will cause an allergic reaction in a hypothetical patient (Mr. X). The foods, then, constitute the cues, and the allergic reaction is the outcome. Following training, subjects rated how strongly each of the foods predicted the occurrence of an allergic reaction.

Table 3: Design of the experiment.

Condition	Pre-exp	Cond 1	Cond 2
B. Block		AB+	A+
Unover		CD+	C-
L,A&D Control		EF+	G+
PR Control		HI+	H+/H-
BCI		JK-	J+
BCI Control 1		LM-	L-
BCI Control 2		NO-	P-
BB Pre-exp	QR		Q+
BB Pre-exp Control 1	ST		S-
BB Pre-exp Control 2	UV		
Fillers		WX-	

These ratings were taken as a measure of the associative strength of a connection from cue to outcome.

We also follow Dickinson & Burke (1996) and Larkin et al. (1998) in using a large number of cues. 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 the causal efficacy of each cue.

The first two rows of Table 3 (B. Block and Unover) are the contingencies of a standard retrospective revaluation experiment, as shown in Table 1. Retrospective revaluation is demonstrated if D is rated higher than B on test.

The "L,A&D" contingency is a control of the kind used by Larkin et al. (1998). Following compound training in Cond 1, neither cue is presented in Cond 2, and so no revaluation will occur. Thus backward blocking and unovershadowing can be assessed independently relative to this control. Backward blocking would be evidenced by a lower rating of B than E or F; unovershadowing by a higher rating of D than E or F.

The "PR Control" is a second control that might allow us to dissect out the effects of backward blocking and unovershadowing. Following compound training in Cond 1, the competing cue receives partial reinforcement. Thus there are an equal number of H+ and H- trials. Suppose that unovershadowing is much stronger than backward blocking. On each H- trial in stage 2 there would be an unovershadowing effect, with I's association to the US becoming stronger. On each H+ trial there would be little effect, as backward blocking is weak. The contingency becomes, in effect, HI+ H-, i.e. unovershadowing, and so we expect I's rating to be similar to D (from an actual CD+ C- contingency). The opposite would be true if backward blocking were stronger than unovershadowing. In general, the PR Control target cue will receive a rating closer to the target cue of the retrospective revaluation contingency having the stronger effect.

The next three rows show a backward-conditioned inhibition contingency and two controls respectively. As described earlier, backward-conditioned inhibition will be demonstrated if K is rated lower on test than M (from Control 1) and N or O (from Control 2).

"BB Pre-exp" is short for backward blocking pre-exposure. This involves compound pre-exposure during the first stage (cf. compound training for backward blocking), followed by excitatory training of the competing cue in Cond 2. Modified SOP predicts that unovershadowing will have a larger effect than backward blocking in retrospective revaluation, because in a backward blocking con-

tingency the US is strongly predicted in the second stage, such that it has elements in A2. The excitatory A2-A2 learning then offsets the effect of inhibitory A2-A1 learning. As a result of using compound pre-exposure in the "BB Pre-exp" contingency, though, the US will not be expected on Cond 2 trials. Hence when it is presented all of its elements should be free to enter A1. R will be retrieved into A2 by Q via the Q–R association developed during pre-exposure. The resulting A2-A1 overlap should produce strong inhibitory conditioning. Modified SOP thus makes the clear prediction that R will be rated lower than T (from Control 1) and U or V (from Control 2).

The Filler trial was used so that there were an equal number of positive and negative trials during Cond 1.

Method

Participants Twenty-four Cambridge University students (14 female, 10 male; age 19-23) took part in the experiment.

Apparatus The experiment was run on a Power PC Macintosh with a 14" monitor.

The foods used were: Oranges, Tomato, Cheese, Lobster, Rice, Peaches, Banana, Grapes, Yoghurt, Melon, Broccoli, Aubergine, Eggs, Potatoes, Carrots, Lentils, Sardines, Gammon, Dates, Mushrooms, Raspberries, Jam, Onion, Steak. These foods were randomly assigned to the letters A to X in the experimental design for each subject.

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 looking over records of foods eaten at the clinic by Mr. X, but would not be told whether or not allergic reactions occurred, while in the second and third blocks they would be asked to make predictions based on the foods eaten. 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.

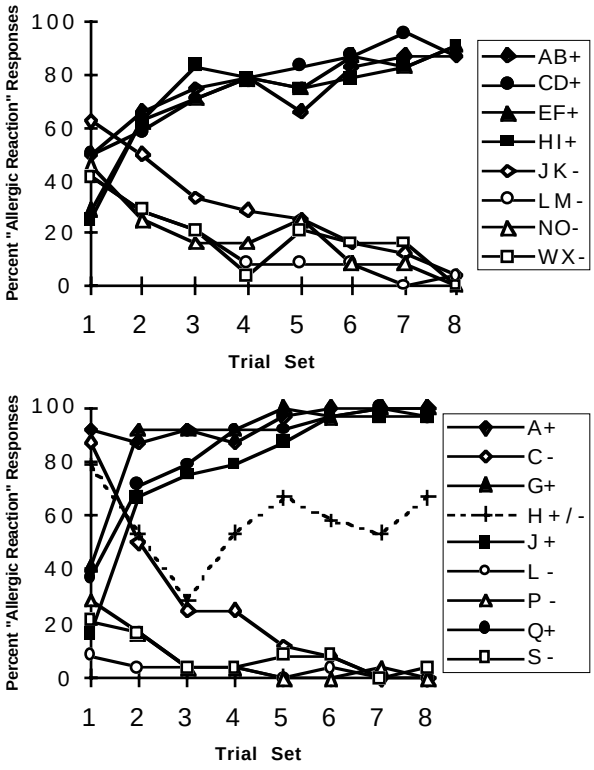

Figure 1. Acquisition of discriminations in (A) Stage 1 and (B) Stage 2.

On each pre-exposure trial, the words "Meal [meal number] contains the following foods:" followed by the two foods appeared on the screen. Subjects were then 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 three trial types in this stage: the order of trials was randomized over each set of three with the constraint that there were no immediate repetitions across sets. Participants saw each pair of foods eight times in this stage. The order of presentation on the screen (first/second) within each compound pair was randomized.

The same message appeared on the screen on Cond 1 and Cond 2 trials. However, now the subjects were 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. There was an explicit break between Cond 1 and Cond 2, when subjects were told that they would now see a new set of meals, some of which contained foods they had seen earlier and some of which didn't. There were eight trial types in Cond 1, and nine in Cond 2.The order of trials was randomized over each set of eight or nine. Participants saw each meal eight times in Cond 1 and Cond 2. Four of the eight H trials in Cond 2 were positive, and the other four were negative, in random order.

In the final rating stage subjects were asked to rate their opinions of the effect of eating each of the foods on a scale from -10 to +10. They were to use +10 if the food was very likely to cause an allergic reaction in Mr. X, -10 if eating the food was very likely to prevent the occurrence of allergic reactions which other foods were capable of causing, and 0 if eating the food had no effect on Mr. X (i.e. it neither caused nor prevented allergic reactions).

All of the foods seen in training were then presented in random order for rating. For clarification, participants also had access to a card on which the instructions on how to use the rating scale were printed. Once a food had been rated it disappeared from the screen and the next appeared, so that participants could not revise their opinions upon seeing later foods.

Results

Figure 1 illustrates the percentage of trials on which subjects thought an allergic reaction would be caused by the food(s) shown in each of the 8 trial sets of Cond 1 and 2. Subjects' responses were clearly appropriate to the relevant underlying contingencies by the end of each stage, with all of the positive trial types eliciting more "Allergic Reaction" responses, negative trial types receiving more "No Reaction" responses, and the H+/H- trials receiving about 50% positive and negative responses.

Of more interest are the ratings of the causal efficacy of each of the foods. The mean rating given to each of the 24 foods is shown in Figure 2. 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 food [$F(23,529) = 22.46$, $p < 0.001$]. Retrospective revaluation was seen, in that the target cue of the backward blocking contingency (B) was rated significantly lower than that of the unovershadowing contingency (D) [$F(1,23) = 7.24$, $p < 0.01$]. Hence it appears that our experimental paradigm is sensitive to changes in the perceived causal efficacy of cues on trials on which those cues are not present.

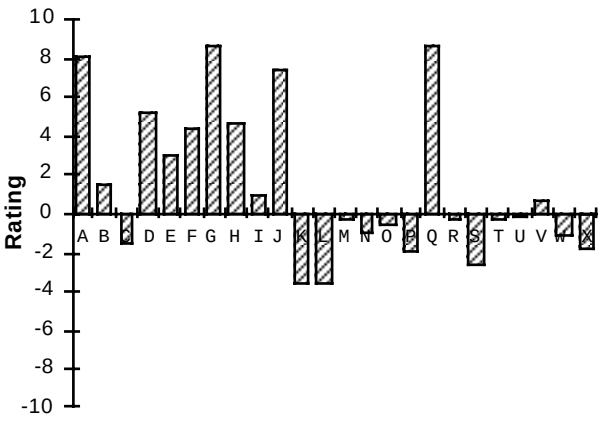

Figure 2. Mean ratings given to the 24 foods.

The result for the L,A&D Control contingency is given by the average of cues E and F, which are equivalent. This does not differ significantly from B or D [$F(1,23)$=2.78 and 1.23 respectively, ps > 0.05]. Given this failure to reach significance, our results neither confirm nor contradict those of Larkin et al. (1998).

We now turn to the results of the PR Control contingency. The rating of the target cue from this contingency (I) is very similar to that of the backward blocking contingency, but quite different from that of the unovershadowing contingency. This is supported statistically: B and I do not differ significantly [F < 1], whereas the difference between D and I is highly significant [$F(1,23) = 7.44$, $p < 0.01$]. It was stated earlier that the rating of the target cue of the PR Control contingency should be more similar to the target cue of whichever retrospective revaluation process (backward blocking or unovershadowing) is stronger. Hence, given that the rating of I is more similar to B than D, the PR Control contingency indicates that backward blocking is a stronger process than unovershadowing.

We also have evidence for backward-conditioned inhibition in this experiment. Cue K is rated lower than its equivalents in the two control contingencies (M, and the average of N and O, none of which differs significantly from each of the others). These differences are significant [$F(1,23) = 7.45$, $p < 0.01$ and $F(1,23) = 6.98$, $p < 0.01$ respectively]. There is no evidence for any retrospective learning in the BB Pre-exp contingency, however. The target cue of this contingency is R. The two controls here are T and the average of U and V (none of which differ from one another). In the former case, the means are identical; in the latter the difference is not significant [$F < 1$].

Discussion

Looking at the results above, we can see that modified SOP is successful in explaining some aspects of this experiment (the occurrence of retrospective revaluation and backward-conditioned inhibition). However, it also has important failures.

The fact that the PR Control indicates a stronger role of backward blocking than unovershadowing in this experiment is a great problem for modified SOP. It implies that the simplistic approach taken to the associative processes occurring in these contingencies is insufficient to provide a full account of human behaviour with respect to learning about absent cues, as the model predicts that unovershadowing will be more influential than backward blocking.

Note that this result also argues against subjects' using a rational, Bayesian approach to the contingencies seen. According to this idea, subjects would integrate the information experienced in the two stages to derive the most likely cause of the US. For example, A- trials following AB+ trials (unovershadowing) indicate that it *must* have been B that caused the US on the AB+ trials. Hence B's rating will increase as a result of A- trials. Less information is given by A+ trials following AB+, though: B *could* still be a cause of the US on AB+ trials. Hence this rational approach predicts that unovershadowing will be stronger than backward blocking, whereas the results of the PR Control indicate the opposite.

In addition, modified SOP predicts a large difference between the target cue of the BB Pre-exp group and its controls, but no difference is seen. The BB Pre-exp contingency should be the situation in which the inhibitory A2-A1 process, proposed to underlie backward blocking, should be most prominent, and yet no effect was seen. This is particularly noticeable as the BCI contingency, which on the surface appears very similar, did show an effect. The failure to find an effect in one of the two contingencies is hard to reconcile with modified SOP.

In summary, then, it seems that modified SOP is able to explain the existence of retrospective revaluation and associated phenomena, but that the mechanics of the explanation offered do not agree with our empirical findings. We now offer an alternative class of model which seems better able to cover the known facts with respect to human studies of retrospective revaluation.

APECS: A model of associative learning

It is possible to explain the results of the above experiments using a version of McLaren's (1993) APECS model. The mechanics for learning in APECS are similar to standard backpropagation (Rumelhart, Hinton, & Williams, 1986), but differ in that once the weights appropriate to a mapping have developed, the learning represented in those weights is protected. This is achieved by reducing the learning rate parameters for the hidden 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 hidden unit occurs only if that hidden unit has a negative error value, i.e. *if it is part of a mapping that predicts an incorrect outcome for the current input*. This reduces the interference arising as a result of subsequent experience of similar (but not identical) input patterns. Indeed APECS was originally designed as a solution to the problem of catastrophic interference in learning (McCloskey & Cohen, 1989).

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.

In addition, in our instantiation of APECS each different pattern of stimulation is represented by its own hidden unit, similar to Pearce (1987).

Consider what happens in the network on AB+ training. It will learn a mapping from A and B input units to the US output unit, mediated by a hidden unit that can be thought of as representing the configuration of A and B (AB_{hidden}). On each AB+ trial the excitatory connections to and from AB_{hidden} will grow stronger. Now consider the gap 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_{hidden}→US connection 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 AB_{hidden}. 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 on gaps between the AB+ trials, the AB_{hidden} unit will therefore develop a negative bias.

In a backward blocking contingency we now train on A+ trials. As explained above, in this instantiation of APECS, each different input→output mapping is assigned a new hidden unit. Hence a new unit is recruited to carry the A+ mapping. As the previous excitatory connection from A to the US (via the AB configural unit) is still useful in reducing the output error (i.e. there is a positive output error on A+ trials, which is propagated back to the AB_{hidden} unit), the learning rate for the A→AB_{hidden}→US connections will also remain high. Given the negative bias on the AB_{hidden} unit, and the fact that training was with A and B in stage 1, A alone will not succeed in fully activating the US at the start of A+ training. Hence the connections from A to both hidden layer units, and from the hidden layer units to the US node, will strengthen.

Now, when no stimuli are applied (during inter-trial interval) the hidden units will deliver some positive activation to the output unit. Thus the A_{hidden} unit will assume a negative bias. More importantly, the negative bias on the AB_{hidden} unit will also become increasingly negative to counter the extra positive activation feeding to the US.

What now happens when B is presented on test? B will provide AB_{hidden} with positive activation, but this may not effectively counter the unit's large negative bias, and hence the US will receive little activation. This, of course, assumes that AB_{hidden}'s gain in negative bias outweighs the strengthened AB_{hidden}→US connection. This is guaranteed, as the two bias changes must together counter the increased A→hidden and hidden→US weights of *both* routes to the US, whereas only the increased AB_{hidden}→US connection will facilitate B's ability to retrieve the US.

Hence APECS is able to explain backward blocking: as a result of A+ trials following AB+ training, B becomes less able to retrieve the US. Note that, unlike modified SOP, this is not as a result of new learning about B (the B→hidden connection is unchanged on A+ trials), but rather as a result of changes in the retrievability of a previously-learnt association.

Consider now the A- trials of the second stage of an unovershadowing contingency. Once again a new hidden unit is recruited to carry this mapping. In this case, however, the AB_{hidden} unit is not reused: it carries an inappropriate excitatory mapping and so will have a negative error. Hence its weights are frozen, and it takes on an increased negative bias. In addition, an inhibitory mapping from A to US will develop via the A_{hidden} unit in order to counter the positive activation flowing to the US via the original mapping.

Between A- trials, when no inputs are presented, the US will receive excess negative input as a result of this new, unsuppressed inhibitory mapping. The network counters this problem in two ways. One is for the A_{hidden} unit to develop a negative bias. The other is for the negative bias on the AB_{hidden} unit to reduce, allowing through more positive activation.

The upshot of the decrease in negative bias on the AB_{hidden} unit is that presentation on B will now excite the US more effectively than following initial AB+ training: this is the standard unovershadowing effect.

Thus the features of APECS that prevent it from suffering from catastrophic interference also allow it to explain retrospective revaluation. Indeed, a backward blocking contingency can be seen as an interference design, where two different pathways (via A_{hidden} and AB_{hidden}) compete to activate the same outcome. Hence A+ trials interfere with memory of the AB+ mapping, causing this pathway to be suppressed. In unovershadowing the situation is reversed: the two pathways have opposite outcomes (AB+ and A-), and so on A- trials the AB+ pathway need not be suppressed, and can even become stronger to counter the influence of the new negative pathway.

On performing initial simulations of retrospective revaluation using APECS it was found that unovershadowing consistently showed a larger effect than backward blocking. This sits well with the results of Larkin et al.'s (1998) study. But how then could this model explain the PR Control, which indicates that backward blocking has the greater effect?

The answer lies in the nature of the AB+ A+/A- design used. It was mentioned earlier that each new input→output mapping recruits a new hidden unit. Hence the occurrence of A+ and A- trials in stage 2 will lead to the recruitment of two new hidden units, one carrying an excitatory mapping, the other an inhibitory mapping. Thus there are now two excitatory pathways to the US (via the AB_{hidden} and $A+_{hidden}$ units) as opposed to only one inhibitory pathway (via the $A-_{hidden}$). This means that any influence of the inhibitory pathway on each excitatory pathway (i.e. unovershadowing) will be relatively slight, as the effect is shared between, and countered by, both excitatory pathways. The effect of one excitatory pathway on the other (backward blocking), though, is relatively unaffected, as it is still a one versus one situation. Hence backward blocking is relatively preserved in this contingency, whereas unovershadowing is greatly reduced.

Figure 3 shows simulation results for the retrospective revaluation contingencies of our experiment, along with the empirical results. The simulation results are the average of 24 simulations run with APECS, each representing one subject, with exactly the same trial order as experienced by the real subjects. 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 0.85 when a hidden unit is active and has a positive error, and 0.001 when it is not. The parameter for bias–hidden changes is 0.25

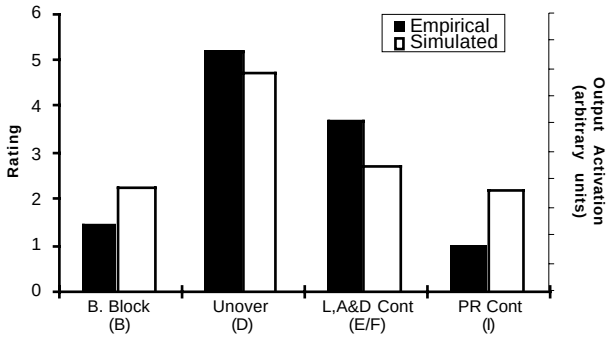

Figure 3. Empirical and simulated data for the retrospective revaluation contingencies.

when a hidden unit is active, 0.001 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.

What predictions does APECS make for other contingencies used in our experiment? Easiest to understand is the BB Pre-exp contingency. On the pre-exposure trials, there is no error on the output unit (whether or not the outcome occurs is not known). Given that it is output error that drives learning in error-correcting networks, this lack of error means that there is no drive to form associations. Hence the cues involved in these trials remain unconnected to the US following pre-exposure. As these cues have no connections to the output, their associative status cannot change on any subsequent trials on which they are not presented, and so no differences will be seen amongst these groups (as observed empirically).

The situation is slightly more complex for the BCI contingency. Typically a context in which outcomes occur will become a weak excitor of the outcome itself. Thus cues presented on negative trials in this context will become weak inhibitors in order to overcome this excitation (demonstrated by the negative ratings given to cues W and X). Hence J and K will develop an inhibitory link to the US via a JK_{hidden} unit. This unit will take on a slight negative bias to prevent its expression when no inputs are presented. The network is now presented with J+ trials. A new hidden unit will be recruited to carry this excitatory mapping. There will be a slight increase in the negative bias on the JK_{hidden} unit, but given that the inhibitory influence of this pathway is slight, the drive to suppress it will also be slight. On the gap between J+ trials, the US will receive excess positive input. One way to decrease this is for the J_{hidden} unit to develop a negative bias to suppress the excitatory mapping just learnt. A second way to reduce the activation of the US is to decrease the suppression of the JK_{hidden} unit, allowing more activity to flow through the inhibitory pathway. This release in suppression between trials (driven by the strong excitatory pathway) more than compensates for the increase in suppression on J+ trials (driven by the weak inhibitory pathway), and so overall the suppression of the JK_{hidden} unit (which carries the inhibitory mapping) decreases over J+ trials. Hence the ability of K to retrieve the US decreases as a result of J+ training following JK- trials: backward-conditioned inhibition is seen.

This is again confirmed by simulation. Figure 4 shows the results for the relevant contingencies from the simulation of this experiment described above.

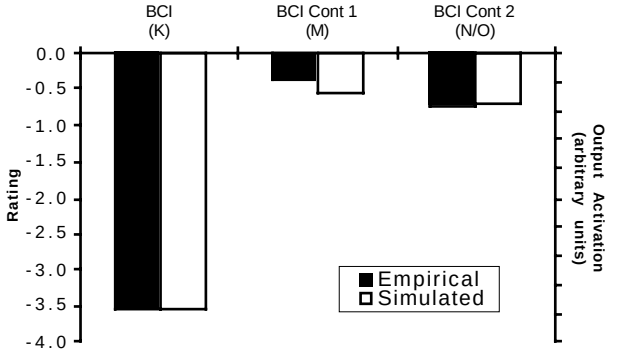

Figure 4. Empirical and simulated data for the backward-conditioned inhibition contingencies.

In conclusion then, it seems that modified SOP is able to explain certain retrospective effects in human causality learning on a coarse scale, but that the explanation offered for these effects (novel learning about absent cues following retrieval via within-compound associations) does not stand up to closer scrutiny. A memory-based explanation, with retrospective effects manifest as changes in retrievability rather than new learning about absent cues, shows better agreement with empirical data, and may prove a more fruitful approach for future investigation.

References

Chapman, G. B. (1991). Trial-order affects cue interaction in contingency judgement. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17,* 837-854.

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

Larkin, M. J. W., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgements under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24,* 1331-1352.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation, 24,* 109-166.

McLaren, I. P. L. (1993). APECS: A solution to the sequential learning problem. *Proceedings of the XVth Annual Convention of the Cognitive Science Society* (pp. 717-722). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgements. *Quarterly Journal of Experimental Psychology, 37B,* 1-21.

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.