Supplementary Materials Supporting:

G. R. Esber

&

M. Haselgrove

Reconciling the influence of predictiveness and uncertainty on stimulus salience:

A model of attention in associative learning

S1. The Mackintosh (1975) model

Mackintosh [2] suggested that the change in the association (ΔV) between a cue (i) and the outcome proceeds according to the equation:

$$\Delta V_i = \alpha_i \ \theta \ (\lambda - V_i) \qquad \text{Equation 1.1}$$

In which θ is a parameter that determines the rate of learning, $(\lambda - V_i)$ is an error term that represents the discrepancy between the asymptote of learning that can be supported by the outcome (λ) and the current strength of the association between cue i and outcome (V_i). Most importantly, α_i is the attention paid to cue i, which varies according to the following rules:

$$\Delta\alpha_i > 0 \text{ if } \left|\lambda - V_i\right| < \left|\lambda - V_p\right| \qquad \text{Equation 1.2}$$

and

$$\Delta\alpha_i < 0 \text{ if } \left|\lambda - V_i\right| \geq \left|\lambda - V_p\right| \qquad \text{Equation 1.3}$$

in which V_p is the associative strength of all cues other than i present on that trial. Equation 1.2 ensures that attention to cue i (α_i) will increase if its error term is smaller than the error term generated by all other cues (i.e. if cue i is the best available predictor of the outcome). Equation 1.3 ensures that attention to cue i (α_i) will decrease if its error term s, at best, no better than the error term generated by all other cues (i.e. if cue i is not the best available predictor of the outcome). In this way, Mackintosh's theory predicts that attention will be higher for predictive than non-predictive cues.

S2. The Pearce-Hall (1980) model

According to Pearce and Hall [5], the change in the association (ΔV) between a cue (i) and the outcome proceeds according to the equation:

$$\Delta V_i = S_i \ \alpha_i \ \lambda \qquad \text{Equation 2.1}$$

The terms within this equation that are common to Mackintosh's theory refer to the same quantities. The novel S_i is a parameter similar to Mackintosh's θ and determines the rate of learning, and depends upon the intensity of cue i. Pearce & Hall suggested that a cue can also be associated with a representation of the outcome's absence, a so-called no outcome ($\overline{V}$), and the change in the association ($\Delta\overline{V}$) between a cue (i) and the no outcome proceeds according to the equation:

$$\Delta\overline{V}_i = S_i \ \alpha_i \ \overline{\lambda} \qquad \text{Equation 2.2}$$

The magnitude of the absence of the outcome ($\overline{\lambda}$) is determined by the difference between the expectation of the outcome and λ itself (which on trials with the absence of the outcome will equal zero). Thus, this magnitude is equal to:

$$\overline{\lambda} = (\Sigma V_i - \Sigma\overline{V}_i) - \lambda \qquad \text{Equation 2.3}$$

Finally, the attention that is paid to cue *i* on the subsequent conditioning trial (α_i^{n+1}) is equal to the current absolute difference between the magnitude of the US, and the total associative strength of the cue:

$$\alpha_i^{n+1} = \left| \lambda - (\Sigma V_i - \Sigma \overline{V}_i) \right| \qquad \text{Equation 2.4}$$

Consequently, a cue that is paired, on some trials, with an outcome but on other trials with the absence of the outcome will enjoy a maintenance of its attention as the value of α will remain positive. However, a cue that is consistently followed by an outcome will suffer a loss of attention as the value of α will tend towards zero.

S3. The importance of distinguishing between salience and associability

The model we advocate envisages that reinforcers of opposite affective polarity should be capable of enhancing the salience of their predictors independently, each in proportion to its emotional significance. This implies that the acquired salience of a cue depends directly upon, but is detached from, its acquired affective value. While a common assumption in attentional theories of learning [2,4,5,7], this notion has been challenged by demonstrations that the associability of a cue—a widely used index of its salience—is partly determined by its affective value.

Using a human causal learning paradigm, Le Pelley, Oakshott, Wills, and McLaren [S1] required participants to play the part of a dietician and learn the effects that ingesting certain types of food (e.g. carrots, eggs) would have on fictitious patients. Participants were divided into two groups, one in which foods were followed by appetitive post-ingestive consequences (enjoyment reactions), and another in which they were followed by aversive post-ingestive consequences (allergic reactions). During Stage 1, some of the foods were highly predictive of their effects, whereas others had no predictive value. In Stage 2, half of the participants in each group were faced with a new problem. Novel combinations of previously predictive and nonpredictive foods now signalled, with identical correlation, a set of novel consequences drawn from the same affective category (e.g. novel types of allergy if allergies had been used in Stage 1). During a final test, participants were asked to rate the likelihood that each of these foods would by themselves cause the effects they were paired with in Stage 2. Replicating previous studies [16], Le Pelley et al. [S1] found that foods that had been highly

predictive in Stage 1 scored significantly higher than foods that had been nonpredictive, despite their being equally predictive of Stage-2 outcomes. These results thus confirm that a history of predictiveness can influence the associability of a cue, presumably through an enhancement of its salience.

The critical results came from the other half of the participants. They received similar training in Stage 2, except for the fact that post-ingestive consequences belonged to the opposite affective class (e.g. enjoyment reactions if allergies had been used in Stage 1). In the test, these participants did not rate foods that had been predictive during Stage 1 any different from foods that had been nonpredictive. Thus, it appears that for the acquired salience of a cue to translate into greater associability, the new learning experience must involve a reinforcer of the same affective class as the initial learning experience. Le Pelley et al. [S1] took their results to show that salience is not a general, value-detached property of the cue, but rather one that is modified separately within (and therefore specific to each motivational system. This interpretation is incompatible with the processes of salience modification we propose here.

How might the model be salvaged in the face of Le Pelley et al. [S1]'s results? In our view, the force of this challenge is undermined when we abandon the tacit assumption that the associability of a cue is entirely determined by its salience. In the associative learning literature, these terms are often used interchangeably, presumably because differences in associability remain one of the few tools at the investigator's disposal for inferring corresponding differences in salience. It is evident, however, that the influence that the salience of a cue exerts on its associability may be obscured by other factors, such as the cue's similarity to the outcome [S2]; its biological relevance or

preparedness [S3,S4], the degree of generalization from other cues [S5] and the inhibitory nature of appetitive-aversive interactions [33,S6].

The antagonistic relationship between appetitive and aversive motivational systems has been well documented [for a review: 36]. Most relevant here are classical studies showing that it is rather difficult to establish a predictor of shock as a signal for food (and vice versa), by comparison with a previously nonreinforced cue [S7,S8]. These results join Le Pelley et al.'s [S1] in demonstrating that the associability of a cue with a history of predictiveness may under some circumstances appear no different or even inferior to that of a relatively neutral stimulus. Learning models that assume hardwired inhibitory associations between appetitive and aversive affective representations, such as that proposed here, can readily accommodate this pattern of results. Briefly, these models predict that a cue previously associated with shock will have no difficulty in entering into an association with food. For a good part of this training, however, this new association will not be readily expressed in performance, because the cue will simultaneously activate the representation of shock, which will in turn suppress activation of the representation of food. As a consequence, acquisition of the cue-food association will appear retarded. Note that the influence of this process on associability is orthogonal to that of stimulus salience, which leaves the possibility intact that the acquired salience of a cue might be detached from its value[1].

It should now be possible to recognize that the experiments of Le Pelley et al. [S1] likely involved such appetitive-aversive interactions. In their procedure, an accurate predictor of, say, a certain allergy may not only have acquired substantial salience by the end of Stage 1, but should also have developed a strong association with that outcome.

When this cue is next paired with an enjoyment reaction during Stage 2, it is plausible that appetitive-aversive interactions might dampen the expression of this association, offsetting the advantage conferred by the cue's high salience. By contrast, appetitive-aversive interactions should have less of an influence on performance to a nonpredictive cue, for its association with allergy should be rather weak to begin with. This would explain why predictive and nonpredictive cues seemed to enter into the second association with equal readiness.

Thus, the distinction between salience and associability is crucial in the context of the current model, as it allows us to account for these and other findings that, on the surface, appear problematic (e.g. [S9]). Indeed, the main implication from the foregoing analysis is that some caution is advisable when making inferences about the salience of a cue on the basis of its associability.

S4. Parameters in the model

V and $\bar{V}$: In simulations of the model, an asymptote at 1 was imposed on the cue→reinforcer (V) and cue→no-reinforcer ($\bar{V}$) associations. This was done in order to prevent the run-away growth they would otherwise undergo in partial reinforcement schedules. Unbound increments in the strength of these associations is a consequence of the fact that they both contribute to the prediction errors used to adjust their values.

ε: Although we must remain agnostic as to the nature of the function *f* in Equations 1.2 and 1.3, for the purpose of simulating the model we took it to be the identity function.

This means that, in our simulations, ε is simply the sum of associative strengths of the cue. In partial reinforcement, therefore, ε equals the strength of the cue→reinforcer and cue→no-reinforcer associations ($V + \bar{V}$). Because we capped the asymptotes of V and $\bar{V}$ at 1, the maximum value that ε can take on in our simulations of partial reinforcement is 2.

α: allowing ε to take on values greater than 1 implies that α can in principle also be greater than 1 (see Equation 5). In order to keep the product of our learning rate parameters (α x β) between 0 and 1 in Equations 3.1 and 3.2, we stipulated for β to be in the range $0 < \beta < 0.1$. Alternatively, we could have normalized α by dividing its actual value by the maximum possible value that α could take in that specific training situation. Although this solution may appear more elegant, notice that it involves adjusting the normalization coefficient for each specific training situation, since the maximum possible value for ε in the denominator will vary depending on the number of reinforcer (and no-reinforcer) representations involved.

β: Because in partial reinforcement each of the cue→reinforcer and cue→no-reinforcer associations can either strengthen or weaken on a given trial, 4 types of changes in associative strength may take place as a result of this training: 1) an increment or 2) decrement in the cue→reinforcer association, and 3) an increment or 4) decrement in the cue→no-reinforcer association. As noted above, we acknowledge the possibility that these changes might occur at different rates, which prompts the question of how to go about selecting the 4 corresponding β values (see Equations 3.1 and 3.2). To begin to address this question, we set out to find the combinations of β values that satisfied what

we regard to be the central theoretical problem faced by the model. This problem, succinctly, is the fact that partial reinforcement, relative to continuous reinforcement, endows the cue with higher salience *if no better predictors of the trial outcomes are present*, but lower salience *if better predictors of the trial outcomes are present* [for a direct comparison, see 14]. For example, if a cue X is partially reinforced with food (X→food, X→nothing), then evidence suggests that as a result of this training the salience of X should be greater than that of a cue A that is continuously reinforced (A→food). In the simulations discussed below, we refer to this pattern of results as Condition 1. By contrast, if partial reinforcement with X is embedded in a true discrimination of the form AX→food, BX→nothing, where the outcome of each trial is unambiguously predicted by A and B, then evidence suggests that the salience of X should now be less than that of the continuously reinforced A. We shall refer to this pattern of results as Condition 2.

In order to explore the β-parameter space and identify the β combinations that satisfy Conditions 1 and 2, we simulated the above scenarios across multiple iterations that systematically varied the values of the 4 βs (β values were restricted to 0.01, 0.03, 0.05, and 0.07). Out of the 256 possible β combinations, 114 (≈ 45%) were found to satisfy Condition 1; i.e., these combinations allowed the model to predict that the salience of X following X→food, X→nothing training is greater than that of A following A→food training. A lawful relation across these 114 combinations could be identified: the product of the 2 βs for increments in the cue→food and cue→no-food associations was always greater than the product of the 2 βs for decrements in these associations[2]. This relation was further investigated with a wide range of β combinations ($0<\beta<0.1$),

which confirmed that only when this product rule is violated is the salience of A higher than that of X.

It is straightforward to see why this should be the case. According to the model, the acquired salience of X (i.e. ε) results from the sum of the X→food and X→no-food associations. If these associations are eroded during training through substantial extinction, then their combined values will not overcome that of the A→food association, and consequently the salience of X will be lower than that of A. In the opposite extreme, if little extinction of these associations is allowed to occur, then the model erroneously predicts that, following training of the kind AX→food, BX→nothing, the salience of X will be higher than that of A. Such was the case with 33 β combinations that violated Condition 2 (≈ 13% of the initial 256). In all of them, the sum of the X→food and X→no-food associations overcame the value of the A→food association. We proceeded therefore to eliminate these 33 cases in order to obtain the list of 81 β combinations that satisfied Conditions 1 and 2 (≈ 32% of the initial 256).

A final, refining step was taken, which consisted in testing the remaining 81 combinations across a number of simulations of salience-related problems investigated in the Pearce and Haselgrove labs [14, S12]. Fourteen combinations were then selected for providing the best fit for our data, and one of these was used in the simulations presented in Figure 1. The novel predictions generated from the model were also based on this particular set of parameters.

Supplementary notes

1. That the influence of salience on associability is orthogonal to that of other factors, such as appetitive-aversive interactions, is suggested by more recent experiments from the Le Pelley lab [S10]. They showed that cues that are highly predictive of an affectively neutral outcome subsequently enter into associations with either *positive or negative* outcomes more readily than previously nonpredictive cues.
2. It is notable that existing theories of associative learning (e.g. [32]) adopt the assumption that the acquisition of associative strength proceeds more rapidly than the loss of associative strength, and that Rescorla [S11] has provided evidence that the acquisition of conditioned responding proceeds more readily than the extinction of conditioned responding.

Supplementary references

S1 Le Pelley, M. E., Oakeshott, S. M., Wills, A. J. & McLaren, I. P. L. 2005 The outcome specificity of learned predictiveness effects: parallels between human casual learning and animal conditioning. *J. Exp. Psychol.: Anim. Beh. Proc.* **31**, 226–236.

S2 Rescorla, R. A. & Furrow, D. R. 1977 Stimulus similarity as a determinant of Pavlovian conditioning. *J. Exp. Psychol.: Anim. Beh. Proc.* **3**, 203–215.

S3 Garcia, J. & Koelling, R. A. 1966 Relation of cue to consequence in avoidance learning. *Psychon. Sci.* **4**, 123–124.

S4 Rozin, P. & Kalat, J. W. 1971 Specific hungers and poison avoidance as adaptive specializations of learning. *Psychol. Rev*. **78**, 459–486.

S5 Pearce, J. M. 1987 A model of stimulus generalization for Pavlovian conditioning. *Psychol. Rev*. **94**, 61–73.

S6 Dickinson, A. & Pearce, J.M. 1977 Inhibitory interactions between appetitive and aversive stimuli. *Psychol. Bull.* **84**, 690–711.

S7 Konorski, J. & Szwejkowska. G. 1956 Reciprocal transformations of heterogeneous conditioned reflexes. *Ada Biologiae Experimenialis* **17**, 141–165.

S8 Scavio, M. J. 1974 Classical-classical transfer: Effects of prior aversive conditioning upon appetitive conditioning in rabbits. *J. Comp. Physiol. Psychol.* **86**, 107–115.

S9 Dickinson, A. & Mackintosh, N. J. 1979 Reinforcer specificity in the enhancement of conditioning by posttrial surprise. *J. Exp. Psychol.: Anim. Beh. Proc.* **5**, 162–177.

S10 Le Pelley, M. E., Reimers, S. J., Beesley, T., Spears, R., Murphy, R. A. & Calvini, G. 2010 Stereotype formation: Biased by association. *J. Exp. Psychol.: Gen.* **139**, 138–161.

S11 Rescorla, R. A. 2002 Comparison of the rates of associative change during acquisition and extinction. *J. Exp. Psychol.: Anim. Beh. Proc.* **28**, 406–415.

S12 Pearce, J. M., Esber, G. R., George, D. N., & Haselgrove, M. 2008 The nature of discrimination learning in pigeons. *Learn. Behav.* **36**, 188–199.