Flexibility of learning in complex worlds

Olof Leimar

Introduction

There is large literature on variation in the rate of learning about compound stimuli, including effects of attention on learning. The papers on the general topic are from experimental psychology, neuroscience, and the computer-science oriented study of reinforcement learning, including neural networks.

The focus of this work is on the learning situation for an individual that encounters increasingly complex environments and needs to discriminate between compound (complex) stimuli. Developing cleaner fish would be one example of such a situation, but there are of course many others. The overall aim is to investigate the possible adaptive value of flexibility of learning, in the sense of variation in the rates of learning, in such situations.

One general idea about learning in complex environments is the much studied issue of exploration vs. exploitation (see, e.g., Sutton and Barto 2018). Another general and possibly related idea, studied by neuroscientists, but apparently not by experimental psychologist, is that the volatility vs. stochasticity of rewards in the environment can influence learning rates (the latter idea perhaps originated from using Kalman filters as models of learning).

Model description

The learning environments could be loosely based on the situation for developing cleaner fish (Redouan sent out some slides dated 1 June 2022). Still, the environments should be thought of as fairly general situations encountered by learning animals during their life. They could in principle include different numbers of stimulus dimensions, M, and different reward structures. For simplicity, we limit ourselves to additive reward structures (although there are other structures that can occur in nature). There are several types of stimulus dimensions, for instance quantitative measures, the absence/presence of a particular feature (a 0/1 dimension, as well as qualitative aspects like the shape or colour of compound stimuli.

Inspired by the cleaner fish, we use one quantitative stimulus dimension (e.g., client size), together with a number of absence/presence dimensions. Experimental psychology typically deals with absence/presence of stimulus components. The

expected reward is assumed to depend linearly on the stimulus components for the different dimensions. There is also random variation in rewards (e.g., for client fish, there is work by Alexa Grutter finding that the number of parasites is correlated with size, but the correlations are not very close to one). All in all, this means that the 'true expected reward' from compound stimulus k, with stimulus components or 'features' x_{km}, $m = 1, \ldots, M$, is given by

$$\bar{R} = \sum_{m=1}^{M} W_m x_{km}, \tag{1}$$

where W_m is the true expected value per x (e.g., for size), or the true expected value of a feature (for a 0/1 stimulus dimension). One possibility is to assume an additive random error in the actual rewards, but it might be more biologically realistic to assume that the actual reward R has a log-normal distribution, so that

$$R = \bar{R} \exp(z_R), \tag{2}$$

where z_R is normally distributed with zero mean and standard deviation σ_R. We might, for instance, have $\sigma_R = 0.1$, which is used in Figure 1a below.

Stimulus dimensions and compound stimuli

In order to characterise many (up to 10) different compound stimuli, there are 10 stimulus dimensions. The first four dimensions are as follows, together with their true values.

1. The first dimension, x_1, is quantitative, like client size, and has a positive true value, e.g., $W_1 = 1.0$.

2. The second dimension is 0/1, and has a zero true value, $W_2 = 0$, so it is an irrelevant dimension.

3. The third dimension is 0/1 and has a positive true value, e.g., $W_3 = 1.0$, so it is a relevant dimension.

4. The fourth dimension is 0/1 and has a negative true value, e.g., $W_4 = -1.0$, so it is also a relevant dimension.

An additional six 0/1 dimensions are described in Table 1 below. From combinations of the four first dimension we have four types of compound stimuli (e.g., corresponding to four client species).

1. The first type has small size, e.g. $x_1 = \bar{x}_{\text{small}} \exp(z_x)$, with z_x normally distributed with mean zero and standard deviation σ_x, and absence of features in the other dimensions. This could be a species of small clients; perhaps with $\bar{x}_{\text{small}} = 1$ and $\sigma_x = 0.25$.

2. The second type has large size, e.g. $x_1 = \bar{x}_{\text{large}} \exp(z_x)$, again with z_x normally distributed with mean zero and standard deviation σ_x, and presence of a feature in the second dimension, and absence of features in the other dimensions. This could be a species of large clients ($\bar{x}_{\text{large}} = 2$) that are characterised by a feature x_2 that is irrelevant for reward (size is sufficient to predict reward).

3. The third type of compound stimulus is the same as the second for the first two dimensions, but it has a feature present in the third dimension and no feature in the fourth. This is then species of more valuable large clients. Perhaps parrotfish could be an example

4. The fourth type of compound stimulus is the same as the second for the first two dimensions, and it has no feature in the third dimension but a feature present in the fourth. This is then species of less valuable large clients. Perhaps damselfish could be an example.

These compound stimuli, together with six additional compound stimuli, are described in Table 1. The distribution of rewards from the first two compound stimuli is illustrated in Figure 1a (see below).

Learning trials

We consider two cases of sequences of learning trials. In both cases, there is first a phase of T trials of learning ($T = 1000$) with only the first two compound stimuli (e.g., one species of small clients and one species of large clients). This is followed by a phase of an additional T trials of learning in a more complex world. In case 1, the agent learns to discriminate between the first four compound stimuli in Table 1. In case 2, the world is even more complex, such that the agent learns to discriminate all 10 compound stimuli in Table 1. In both cases, the agent can choose between two compound stimuli in each trial, and these are randomly drawn from all types that occur in that phase of learning of that case.

Table 1: Characteristics of stimulus dimensions and compound stimuli used. There are 10 compound stimuli that can be distinguished using 10 stimulus dimensions. The first dimension represents size, with values small (1.0) and large (2.0), and the others are absence/presence (0/1) dimensions. The reward values per feature (W_m) are given in the second column and the features of the different compound stimuli are in the following columns.

Dim	Value	CS1	CS2	CS3	CS4	CS5	CS6	CS7	CS8	CS9	CS10
1	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0
2	0.0	0	1	1	1	0	0	1	0	0	0
3	1.0	0	0	1	0	0	0	0	0	0	0
4	−1.0	0	0	0	1	0	0	0	0	0	0
5	2.0	0	0	0	0	1	0	0	0	0	0
6	1.0	0	0	0	0	0	1	0	0	0	0
7	−1.0	0	0	0	0	0	0	1	0	0	0
8	2.0	0	0	0	0	0	0	0	1	0	0
9	1.0	0	0	0	0	0	0	0	0	1	0
10	−1.0	0	0	0	0	0	0	0	0	0	1

Learning algorithms

There are many learning algorithms that have been proposed in the literature. Perhaps most important is Rescorla-Wagner learning (Rescorla and Wagner 1972), and this will be used for the results below. There are number of additional variants of algorithms suggested by experimental psychologist (e.g., Mackintosh 1975; Pearce and Hall 1980; Le Pelley 2010; Pearce and Mackintosh 2010; Esber and Haselgrove 2011). These all have problems when applied to the situations we want to study, so they are not so suitable to illustrate the possible adaptive value of flexible learning in complex worlds. We should definitely discuss this in our manuscript, and perhaps give some results in the supplements, to illustrate the problems these algorithms encounter.

There are also Kalman filter inspired approaches, as outlined by Dayan et al. (2000), for instance some of the algorithms described by Sutton (1992b). An issue with the Kalman filter is that it assumes that the agent knows the relative volatility/stochasticity in rewards, which is unrealistic. The 'IDBD algorithm' from Sutton (1992a,b) does not make such an assumption, but makes use of a 'meta learning rate' that needs to have a suitable value for learning to work. The so-called Autostep algorithm (Mahmood et al. 2012) is an elaboration of the IDBD algorithm that is more robust, so Autostep appears like a good choice for us to compare the adaptive

value of flexible learning rates with the fixed rates of Rescorla-Wagner. Thus, we will compare Rescorla-Wagner with Autostep.

Although the Autostep algorithm is a bit complicated (it is given in Table 1 of Mahmood et al. (2012)), the basic idea behind IDBD is intuitive, and is explained by Sutton (1992a). We might also cite the short review by Sutton (2022).

Estimated values and choices

As above, we use a feature vector $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kM})$, where x_{km} is the state of component m of the compound stimulus k. Formally then, the situation faced by an agent in a learning round or trial is the collection of feature vectors $\boldsymbol{x}_k$ of the compound stimuli that are present in the trial. We focus on situations where the agent is facing two compound stimuli, each randomly selected from a set of compound stimulus types. In each round the agent uses estimated values Q_k of each present compound stimulus k to make a choice. Let us assume that the agent makes estimates as follows:

$$Q_k = \sum_{m=1}^{M} w_m x_{km}, \tag{3}$$

where the w_m are the learned weights (expected reward values) of the feature components. The estimated value Q_k is a learned estimate of the reward from the action of selecting that compound stimulus.

With K actions ($K = 2$ choices), an individual uses the estimated values to determine its probability p_{kt} of performing an action, $a = k$, using a soft-max function to convert values to choice probabilities, as follows:

$$p_k = \frac{\exp\big(\omega Q_k\big)}{\sum_{l=1}^{K} \exp\big(\omega Q_l\big)}, \tag{4}$$

where ω is a parameter (we might have $\omega = 5.0$). For two actions ($K = 2$), p is a logistic function of the difference between expected reward values, which is illustrated in Figure 1b (see below).

This is then an example of action-value learning, which can be viewed as a modification of classical conditioning to make it applicable to instrumental conditioning (see sections 2.2 and 2.5 in Sutton and Barto (2018) for discussion of this learning approach). Note also that action-value learning can be regarded as a simplified version of Sarsa, where individuals do not use any sophisticated states and where each

learning trial is a separate episode (terminology from Sutton and Barto (2018)). For the learning algorithms we study here, we get a connection to the presentation in Sutton and Barto (2018) by assuming that the state in a trial is just the compound stimuli or 'objects' that are present in that trial and that an individual can choose between.

Learning updates

Supposing that compound stimulus k was selected in trial t, we have the so-called prediction error

$$\delta_t = R_t - Q_{kt}, \tag{5}$$

where R_t the reward experienced from selecting compound stimulus k and Q_{kt} is from equation (3). The prediction error can be used to update the learning rates (see below), and it is also used to update the learned weights w_{mt}. The learning algorithms we study all assume that

$$w_{m,t+1} = w_{mt} + \alpha_{mt} x_{kmt} \delta_t \text{ if the choice is } k, \tag{6}$$

where the α_{mt} is a learning rate, which can differ between stimulus dimensions and can change with time.

Rescorla-Wagner learning updates

As baseline learning updates we can take the formulation by Rescorla and Wagner (1972). This amounts to assuming 'constant' learning rates, in the sense that

$$\alpha_{mt} = \alpha_{\mathrm{RW}}, \tag{7}$$

where α_{RW} is a constant, independent of the stimulus dimension m and time (trial) t.

This agrees with a formulation by Sutton and Barto (2018): Q_{kt} in equation (3) corresponds to Sutton & Barto's $Q(S_t, A_t)$ in the Sarsa formulation in their equation (6.7). Taking into account their linear function approximation approach in chapter 9, and their episodic semi-gradient Sarsa in the Box on page 244 in chapter 10, Q_{kt} in equation (3) corresponds to their $\hat{q}(S, A, \boldsymbol{w})$, if one takes the action A to be the choice of compound stimulus k (which then ought to be present in state S).

Also, we get the closest correspondence to the original Rescorla-Wagner formulation if we have 0/1 features (x_{km} is 0 or 1) and use the notation V_m for the reward value weights w_m in equation (3).

The IDBD and Autostep algorithms

The IDBD learning algorithm, derived in Sutton (1992a), is given by equations (13), (17), and (20) in Sutton (1992b). Assuming that compound stimulus k is selected in trial t, the learning rates are given by

$$\alpha_{mt} = \exp\bigl(\beta_{m,t+1}\bigr), \tag{8}$$

corresponding to equation (17) in Sutton (1992b). The β_{mt} are updated through

$$\beta_{m,t+1} = \beta_{mt} + \mu\delta_t x_{kmt} h_{mt}, \tag{9}$$

corresponding to equation (13) in Sutton (1992b), where μ is a 'meta learning rate', δ_t is from equation (5), and h_{mt} is an additional quantity introduced by Sutton (1992b). The quantity h_{mt} in equation (9) is determined by the following iterative procedure

$$h_{m,t+1} = h_{mt}\bigl[1 - \alpha_{mt} x_{kmt}^2\bigr]^+ + \alpha_{mt} x_{kmt}\delta_t, \tag{10}$$

corresponding to equation (20) in Sutton (1992b). The notation $\bigl[X\bigr]^+$ means equal to X for positive X and zero otherwise.

A problem with the IDBD algorithm, pointed out by Mahmood et al. (2012), is that it is very sensitive to the exact value of the meta-learning rate μ. To avoid this, they introduced a number of 'tricks', resulting in a much more robust algorithm, which the called Autostep (the 'step size', or meta-learning rate, is automatically set to good value). The precise method of achieving this is given in Table 1 of Mahmood et al. (2012).

Results

Some modelling details and results from the first phase of learning (with only two types of compound stimuli) appear in Figure 1. The reward distribution from small and large clients is shown in panel a, the sigmoid choice curve from equation (4) in panel b, and the learning rates α_{mt} and estimated values w_{mt} are shown in panels

c and d, averaged over 100 replicates and blocks of 10 learning trials. As seen in Fig. 1d, Rescorla-Wagner and Autostep have similar performance in the first phase of learning.

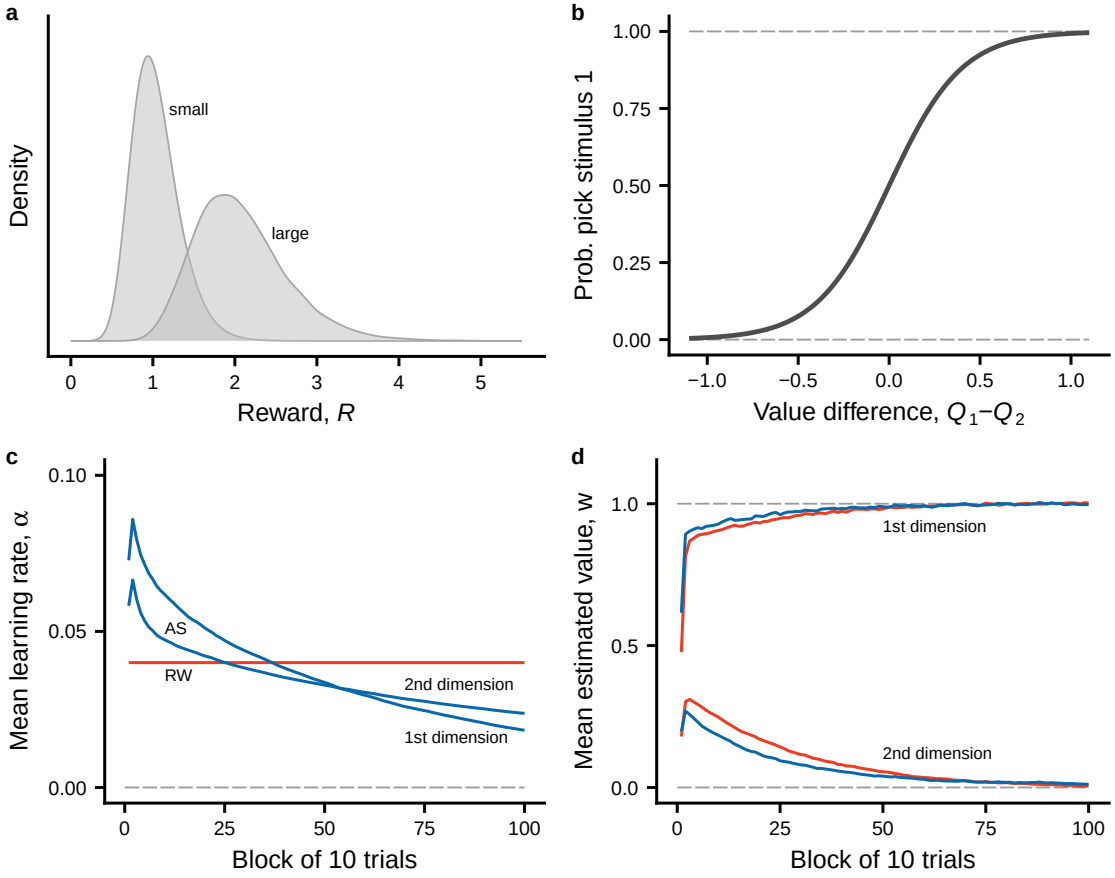

Figure 1: Overview of the first phase, where agents learn to discriminate between two types of compound stimuli ('small' and 'large' clients). **a** Distribution of rewards from the two types of compound stimuli. **b** The function relating the probability of choice to the difference in estimated values of the two compound stimuli presented in a trial. **c** Learning rates for Rescorla-Wagner (RW) and Autostep (AS) for the two stimulus dimensions (first dimension is relevant and second is irrelevant). **d** Estimated values for Rescorla-Wagner and Autostep for the two stimulus dimensions (first dimension has true value 1.0 and second has true value 0). There are 10 trials in a block and data are averages over 100 replicate learning simulations.

There are two cases for the second phase of learning, and the outcome of these learning simulations are illustrated in Figure 2.

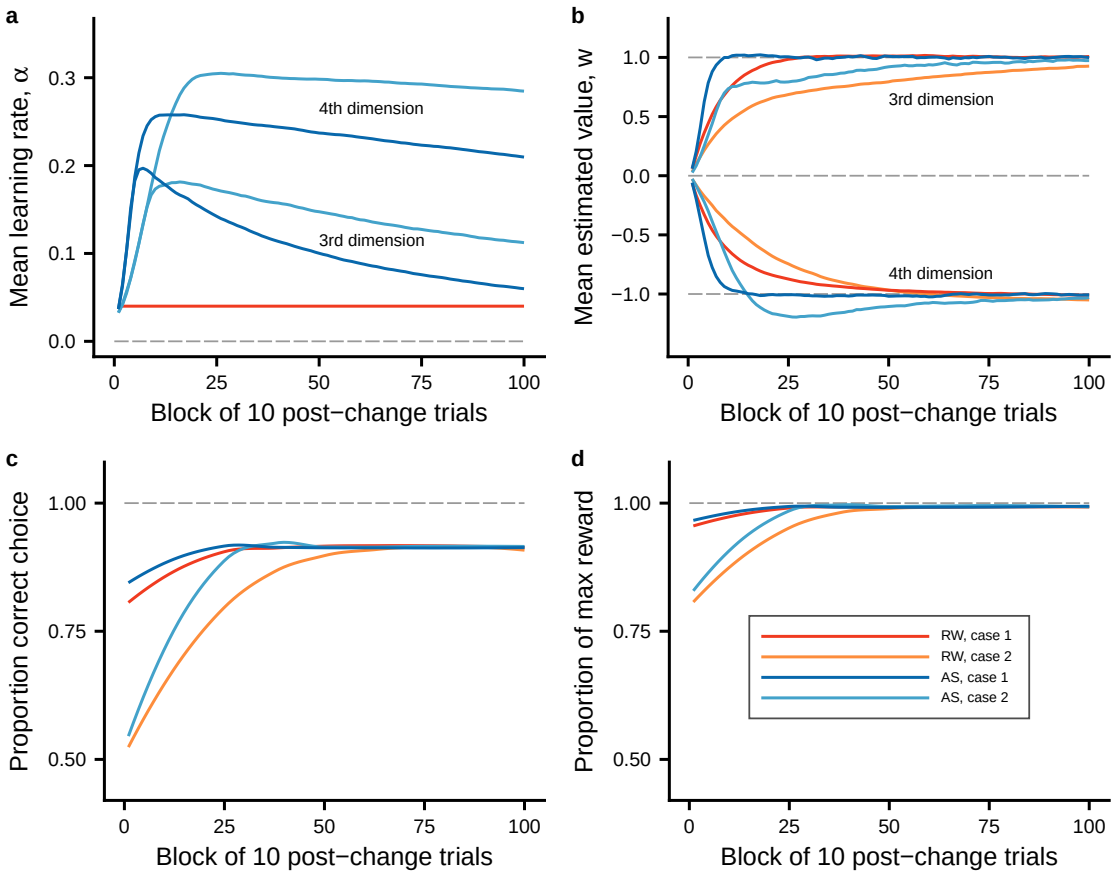

Figure 2: Comparisons of the second phase of learning, when the world becomes more complex, between Rescorla-Wagner and Autostep and for the three cases studied. Colour coding in panel **d** of the curves applies to all panels. **a** Learning rates for the different learning algorithms and cases (note that the learning rate for Rescorla-Wagner is constant). As an illustration, the third and fourth stimulus dimensions are shown. Note that the features in these dimensions were not present in the first phase. The results are similar for the other new dimensions in Table 1 (dimensions 5-10). **b** Estimated values for the different learning algorithms and cases, for stimulus dimensions 3 and 4. **c** Proportion of choices that are correct, in the sense of the agent choosing the compound stimulus with higher true value. **d** Proportion of reward gained out of the maximum true expected reward available in a trial.

Figure 3 gives another illustration of the performance of Rescorla-Wagner and Autostep learning algorithms for the two simulated cases. The performance is measured in terms of the deviation of an agent's estimate from the true value; this is often implemented as the root mean square error (RMSE). As seen from the the previous figures, RMSE is not the only thing that matters. Thus, even if an agent deviates in its estimates, it can still be the case that it makes a correct choice be-

tween two compound stimuli (because the deviations might be similar for the two stimuli).

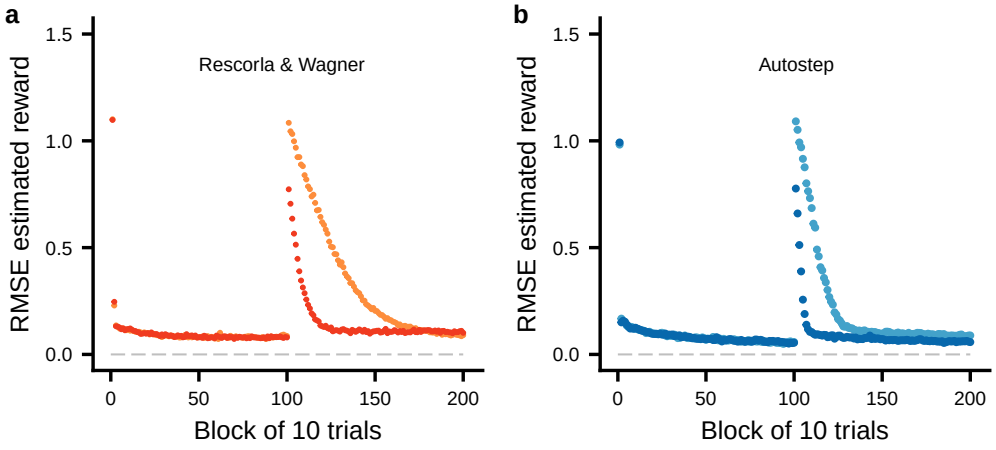

Figure 3: Illustration of the root mean square error (RMSE) of the agent's estimate (Q) of the reward from the selected compound stimulus, plotted against the trial block, over both phases of learning. There are 10 trials in a block and data are averages over 100 replicate learning simulations. (RMSE is similar to a standard deviation but instead measuring the deviation of an estimate from the true value.) **a** Rescorla-Wagner learning, with $\alpha_{\mathrm{RW}} = 0.04$. **b** Autostep learning, following Mahmood et al. (2012). The colour coding is as in Fig. 2d.

Discussion

There is of course a lot to say, but overall the losses from the imperfections of learning were not very big. There are clear differences between the approaches in how good they are at estimating the true reward value of compound stimuli in the second phase, when the world became more complex (Figures 2 and 3), and there were also differences in the rewards obtained (Fig. 2d). Perhaps these differences in achieved rewards are enough to explain why one finds flexible learning rates in experiments.

Literature Cited

Dayan, P., Kakade, S., and Montague, P. R. 2000. Learning and selective attention. *Nature Neuroscience*, 3(11):1218–1223.

Esber, G. R. and Haselgrove, M. 2011. Reconciling the influence of predictiveness

and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences*, 278(1718):2553–2561.

Le Pelley, M. E. 2010. The hybrid modeling approach to conditioning. In Schmajuk, N., editor, *Computational Models of Conditioning*, pages 71–107. Cambridge University Press, Cambridge, UK.

Mackintosh, N. J. 1975. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4):276–298.

Mahmood, A. R., Sutton, R. S., Degris, T., and Pilarski, P. M. 2012. Tuning-free step-size adaptation. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2121–2124.

Pearce, J. M. and Hall, G. 1980. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6):532–552.

Pearce, J. M. and Mackintosh, N. J. 2010. Two theories of attention: a review and possible integration. In Mitchell, C. J. and Le Pelley, M. E., editors, *Attention and Associative Learning: From Brain to Behaviour*, pages 11–40. Oxford University Press, Oxford, UK.

Rescorla, R. A. and Wagner, A. R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: current research and theory*, pages 64–99. Appleton-Century-Crofts, New York.

Sutton, R. S. 1992a. Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 171–176. MIT Press, Cambridge, MA.

Sutton, R. S. 1992b. Gain adaptation beats least squares? In *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, pages 161–166. Yale University, New Haven, CT.

Sutton, R. S. 2022. A history of meta-gradient: Gradient methods for meta-learning. *arXiv preprint, 2202.09701*, pages 1–8.

Sutton, R. S. and Barto, A. G. 2018. *Reinforcement learning: An introduction second edition*. MIT Press, Cambridge, MA.