Attention and learning to discriminate compound stimuli

Olof Leimar

What should be the main points of a manuscript on this topic?

There is now a very large literature on the effects of attention on learning about complex (compound) stimuli. The papers on the general topic are from experimental psychology, neuroscience, and the computer-science oriented study of reinforcement learning, including neural networks. I think a manuscript on the topic needs to make a clear point in relation to all this previous literature.

My suggestion is to focus on the learning situation for an individual that encounters increasingly complex environments and needs to discriminate between complex (compound) stimuli. Developing cleaner fish would be one example of such a situation, but there are of course many others. The overall question would be which kinds of learning rules could be adaptive 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).

I suggest that a paper could illustrate how a few different learning algorithms that have been proposed perform in a few complex environments. Below is a summary of how this could work.

Learning environments

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. My suggestion is that, 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 presence/absence of a particular feature, as well as qualitative aspects like the shape or colour of compound stimuli.

Inspired by the cleaner fish, I suggest that we use one quantitative stimulus dimension (e.g., client size), together with a number of presence/absence dimensions (much experimental psychology deals with presence/absence of component stimuli). The expected reward should depend linearly on the stimuli for the dimensions. There should also be 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 reward' from compound stimulus k, with 'features' x_{km}, $m = 1, \ldots, M$, is given by

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

where W_m is the 'true' expected value per x (e.g., for size), or the 'true' expected value for a feature (for a 0/1 stimulus dimension), and z_R is normally distributed with zero mean and standard deviation σ_R. We might, for instance, have $\sigma_R = 0.5$.

As a possible example, I suggest the following four stimulus dimensions, together with tentative '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.

From combinations of these I suggest the following four types of compound stimuli (e.g., corresponding to four client species).

1. The first type has small size, e.g. $x_1 = \bar{x}_{\text{small}} + z_{x_1}$, with z_{x_1} normally distributed with mean zero and standard deviation σ_{x_1}, and absence of features in the other dimensions. This could be a species of small clients; perhaps with $\bar{x}_{\text{small}} = 1$ and $\sigma_{x_1} = 0.5$.

2. The second type has large size, e.g. $x_1 = \bar{x}_{\text{large}} + z_{x_1}$, with z_{x_1} normally distributed with mean zero and standard deviation σ_{x_1}, 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.

Learning trials

A possible sequence of learning trials is for an agent to first go through T trials where the compound stimuli in each trial are randomly drawn from the first two types described above (we might have $T = 2000$). Then, in the following T trials the environment becomes more complex, such that the two compound stimuli in each trial are randomly drawn from all four types.

Learning algorithms

I think we should illustrate a small number of qualitatively different and well-recognised learning algorithms. The first and perhaps most important is Rescorla-Wagner learning (Rescorla and Wagner 1972). There are many 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), and we could pick one, or possibly two, of these. I suggest we pick the 'unified model' in Pearce and Mackintosh (2010). Finally there are Kalman filter inspired approaches, as outlined by Dayan et al. (2000), for instance some of the algorithms described by Sutton (1992b). I suggest we pick the 'K1 algorithm' and, possibly, an additional one like the 'IBDB algorithm' from Sutton (1992b).

Description of algorithms for discrimination learning

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. For simplicity, we can focus on situations where the agent is facing two compound stimuli, each randomly selected from a set of compound stimulus types. In each round t the agent uses estimated values Q_{kt} of each present compound stimulus k to make a choice. Let us assume that the agent makes estimates as follows:

$$Q_{kt} = \sum_{m=1}^{M} w_{mt} x_{kmt}, \tag{2}$$

where the w_{mt} are the learned weights (expected reward values) of the feature components. The estimated value Q_{kt} 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:

$$p_{kt} = \frac{\exp\big(\omega Q_{kt}\big)}{\sum_{l=1}^{K} \exp\big(\omega Q_{lt}\big)}, \tag{3}$$

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.

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

$$\delta_t = R_t - Q_{kt}, \tag{4}$$

where R_t the reward experienced from selecting compound stimulus k and Q_{kt} is from equation (2). 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

$$w_{m,t+1} = w_{mt} + \kappa_{kmt}\delta_t \text{ if the choice is } k, \tag{5}$$

where the κ_{kmt} is a learning rate, which can differ between stimulus dimensions and can change with time. There are different assumptions about the learning rate; typically κ_{kmt} is proportional to x_{kmt} (e.g., $\kappa_{mkt} = \alpha x_{kmt}$ for Rescorla-Wagner). We can, however, note that neither the original formulation by Mackintosh (1975) nor that by Pearce and Hall (1980) precisely followed the formula in equation (5), whereas the 'unified model' by Pearce and Mackintosh (2010) did.

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

$$\kappa_{kmt} = \alpha x_{kmt}, \tag{6}$$

where α 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 (2) 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 (2) 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_{kmt} is 0 or 1) and use the notation V_{mt} for the reward value weights w_{mt} in equation (2).

The Pearce and Mackintosh 2010 'unified model'

As mentioned, there is the issue of a trade-off between exploration and exploitation in learning. Sometimes discussions about this refer to Mackintosh (1975) and Pearce and Hall (1980) as stand-ins for exploitation and exploration. A synthesis of the two approaches appears in Pearce and Mackintosh (2010), which has been elaborated and commented in other papers. The 'unified model' expresses learning rates

$$\kappa_{kmt} = \alpha_{mt}\sigma_{mt}\beta x_{kmt}, \tag{7}$$

where α_{mt} and σ_{mt} represent 'attentional associability', corresponding to the ideas in Mackintosh (1975), and 'salience associability', according to the ideas in Pearce and Hall (1980), respectively, of the feature state (terminology from Le Pelley (2004)), and β is a constant associability factor. Equation (7) corresponds to equation (2.6) in Pearce and Mackintosh (2010). We should also note that Pearce and Mackintosh (2010) only dealt with presence/absence component stimuli, whereas we extend their approach to also apply to a continuous stimulus dimension (like client size).

According to Pearce and Mackintosh (2010), the learning rate α_{mt} should increase, i.e. $\alpha_{m,t+1} > \alpha_{mt}$, if $\left|R_t - w_{mt}\right| < \left|R_t - \sum_{n\neq m} w_{nt}x_{knt}\right|$, and decrease if the inequality is reversed, assuming that the feature m is present in the selected compound stimulus k (note that there seems to be a misprint in their equations 2.2a, 2.2b). For the 'salience associability' they suggested that $\sigma_{m,t+1} = \left|R_t - w_{mt}\right|$ (their equation 2.5). They added, referring to work by Le Pelley, that the variation in learning rates should be restricted to $0.05 \leq \alpha_{mt} \leq 1.0$ and $0.5 \leq \sigma_{mt} \leq 1.0$.

We need a specific implementation of the dynamics of the α_{mt}. One possibility, suggested by Le Pelley (2010), is

$$\alpha_{m,t+1} = \alpha_{mt} + \gamma_\alpha\Big(\left|R_t - \sum_{n\neq m} w_{nt}x_{knt}\right| - \left|R_t - w_{mt}\right|\Big), \tag{8}$$

where γ_α is a 'meta learning rate' for the α_{mt}. It is worth noting that this 'meta learning dynamics' does not have a stable equilibrium in simple situations, such as when several feature dimensions indicate reward. Instead, a typical outcome of equation (8) seems to be that one learning rate keeps increasing and the others keep decreasing. It thus seems that the limits suggested by Pearce and Mackintosh (2010), for instance that $0.05 \leq \alpha_{mt} \leq 1.0$, are really needed, and that an equilibrium will be where the learning rate α_{mt} is 1.0 for one stimulus dimension and is 0.05 for the other dimensions.

A similar approach for σ_{mt}, also suggested by Le Pelley (2010), is

$$\sigma_{m,t+1} = (1 - \gamma_\sigma)\sigma_{mt} + \gamma_\sigma \big| R_t - w_{mt} \big|, \tag{9}$$

where γ_σ is a 'meta learning rate' for the σ_{mt} (for $\gamma_\sigma = 1$ we get the Pearce and Mackintosh (2010) suggestion). Note that this learning rate enters in a different way from in equation (8). The dynamics in equation (9) can in principle have a stable equilibrium, with $\sigma_{mt} \approx \big| R_t - w_{mt} \big|$, although the limits $0.5 \leq \sigma_{mt} \leq 1.0$ are likely to sometimes come into action.

Concerning the order in time of the learning weight update in equation (5) and the learning rate updates in equations (8, 9), Pearce and Mackintosh (2010) assume that the weight update comes first, followed by the learning rate updates, although the learning rate updates use the estimated weights before they were updated. This time order is perhaps a bit strange.

Generally speaking, these Pearce and Mackintosh (2010) approaches, as interpreted by Le Pelley (2010), seem a bit unsatisfactory in their details. A possible reason for this is that the experimental psychologist involved are not so well trained in mathematics, resulting in a certain amateurism in their formal equations.

Kalman-filter inspired learning, according to Dayan et al. 2000

Dayan et al. (2000) propose a learning approach based on the so-called Kalman filter. The Kalman filter is a classical application of optimal control theory. Under certain specific assumptions, it solves the problem of which 'controls' an agent should use, based on imperfect inputs from the environment, in order to optimally control a system. As a part of the approach, the Kalman filter delivers a prediction of an output variable, and for the application to learning, this output is the estimated reward. In the learning approaches, the 'control' part of the filter is left out, with reward prediction being the part that is used. This then means that Dayan et al. (2000) study classical conditioning, but we can extend it to action value learning.

The particular Kalman-filter assumptions about the environment need to hold for learning in nature, but the theory might still serve as a useful approximation. An interesting aspect of the theory (described in Box 1 in Dayan et al. (2000)) is that learning rates, like the κ_{kmt} in equation (5), should be given by an estimate of the uncertainty in how much a feature in dimension m contributes to the expected reward value in equation (2). At least in an intuitive sense, this shows some similarity to the Mackintosh (1975) and Pearce and Hall (1980) ideas. Pearce and Mackintosh (2010) in fact cite Dayan et al. (2000), stating that the ideas described in Box 1

correspond to Pearce and Hall (1980) and that the ideas described in Box 2 of Dayan et al. (2000), which we have not dealt with here, correspond to Mackintosh (1975). These assessment need however, not be correct, and it seems that no one has published on the issue.

At any rate, it can be useful to include some form of Kalman-filter approach to compare with the Rescorla-Wagner and the Pearce and Mackintosh (2010) learning approaches. One possibility is then to follow Box 1 in Dayan et al. (2000), which cites some approach in Sutton (1992b). The final equation in Box 1 of Dayan et al. (2000) corresponds to equation (10) in Sutton (1992b), as can be seen from his equation (8). There are several different learning procedures proposed in Sutton (1992b), which are different approximations of Kalman filter learning, but perhaps the so-called K1 algorithm is the one that Dayan et al. (2000) had in mind. The K1 algorithm performed rather well in the learning simulations in Sutton (1992b). See also Sutton (1992a) and Gershman (2015) for more explanation about learning and the Kalman filter.

The K1 algorithm

The K1 learning algorithm is given by equations (10), (11), (13), and (15) in Sutton (1992b). Assuming that compound stimulus k is selected in trial t, the learning rates are given by

$$\kappa_{kmt} = \frac{s_{mt}^2 x_{kmt}}{\sum_{n=1}^{M} s_{nt}^2 x_{knt}^2 + \sigma_R^2}, \tag{10}$$

corresponding to equation (10) in Sutton (1992b) (and to the final equation in Box 1 of Dayan et al. (2000), if one notes that they only look at 0/1 features, for which $x_{knt} = x_{knt}^2$). Also, σ_R^2 in equation (10) is the observation error in the reward. This is the same as the variation in reward in equation (1), and represents the stochasticity, but in reality it will not be known by the agent. It might be possible to include learning of the stochasticity into a learning algorithm, but so far this seems only to have been attempted for a single stimulus dimension (Piray and Daw 2021), and the approach in that paper seems rather complicated.

The K1 iterative procedure for updating the learning rates κ_{kmt} is as follows. First, we write the s_{mt}^2, which are 'uncertainties in the predictions' w_{mt} of the 'true' W_m, as

$$s_{mt}^2 = \exp\bigl(\beta_{m,t+1}\bigr), \tag{11}$$

corresponding to equation (11) in Sutton (1992b). These β_{mt} are updated through

$$\beta_{m,t+1} = \beta_{mt} + \mu\delta_t x_{kmt} h_{mt}, \tag{12}$$

corresponding to equation (13) in Sutton (1992b), where μ is a 'meta learning rate', δ_t is from equation (4), and h_{mt} is an additional quantity introduced by Sutton (1992b). It is determined by the following iterative procedure

$$h_{m,t+1} = \big[h_{mt} + \kappa_{kmt}\delta_t\big]\big[1 - \kappa_{kmt}x_{kmt}\big]^+, \tag{13}$$

corresponding to equation (15) in Sutton (1992b). The notation $\big[X\big]^+$ means equal to X for positive X and zero otherwise. We need some assumptions about how to start off the iterations, and also about the values of σ_R^2 in equation (10) and μ in equation (12). We might try $\sigma_R^2 = 0.25$, $\beta_{m0} = \log(\sigma_R^2)$, and $h_{m0} = 0$; this follows Sutton (1992b). From this, and with suitable assumptions about the learning environment (see above), we get successive updates of κ_{kmt} and w_{mt}, with the latter from equation (5). For the meta learning rate, we could try $\mu = 0.05$.

The IBDB algorithm

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

$$\kappa_{kmt} = \exp\big(\beta_{m,t+1}\big)x_{kmt}, \tag{14}$$

corresponding to equation (17) in Sutton (1992b). The β_{mt} are updated as in equation (12) above. A somewhat different quantity h_{mt} in that equation is determined by the following iterative procedure

$$h_{m,t+1} = h_{mt}\big[1 - \kappa_{kmt}x_{kmt}\big]^+ + \kappa_{kmt}\delta_t, \tag{15}$$

corresponding to equation (20) in Sutton (1992b). Also, according to Sutton (1992b) we should have $\beta_{m0} = \log(1/M)$.

Results

Figure 1 illustrates the performance of four different learning algorithms. The parameters for the different approaches have been chosen to achieve approximately the best performance for that approach, in terms of the expected reward loss for the agent, compared to an ideal case where the agent would know the true expected reward for each presented compound stimulus.

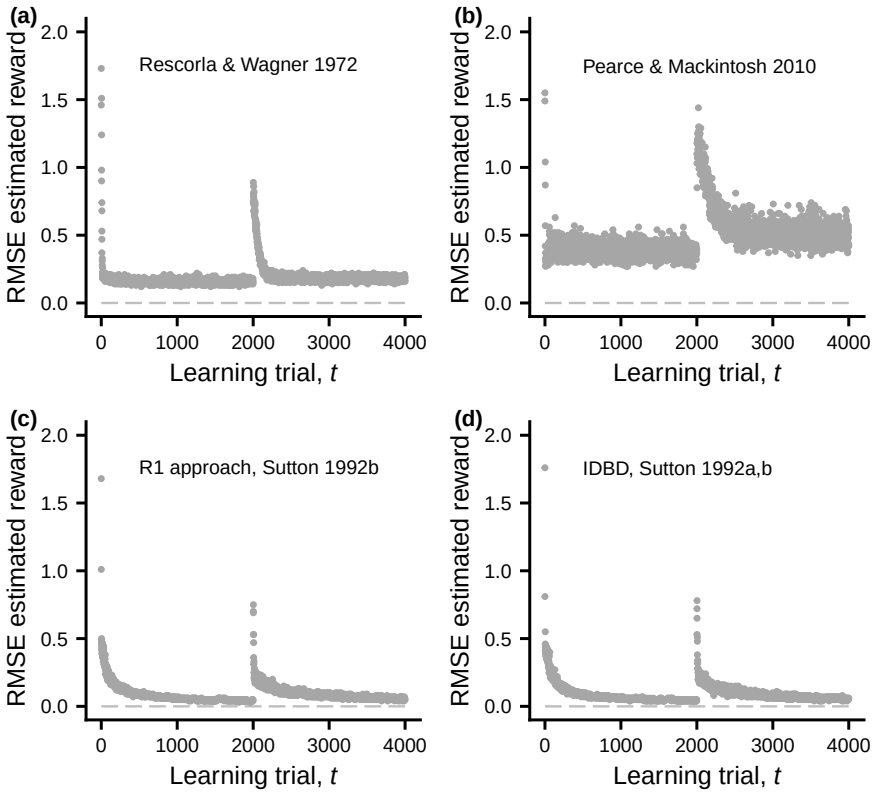

Figure 1: Illustration of the root-mean-square-error for the agent's estimate (Q_{kt}) of the reward from the selected compound stimulus, plotted against the trial number, t. (RMSE is similar to a standard deviation but instead measuring the deviation of an estimate from the true value) **a** Rescorla-Wagner learning, with $\alpha = 0.04$. **b** Pearce and Mackintosh (2010) learning, with $\alpha_{m0} = \sigma_{m0} = 0.5$, $\beta = 0.25$, $\gamma_\alpha = 0.01$, $\gamma_\sigma = 1.0$. **c** K1 approach from Sutton (1992b), with $\mu = 1.0$. **d** IDBD approach from Sutton (1992a,b), with $\mu = 1.0$. Other parameters in the learning simulations were $T = 2000$, $\bar{x}_{\text{small}} = 1.0$, $\bar{x}_{\text{large}} = 2.0$, $\sigma_{x_1} = 0.5$, $W_1 = 1.0$, $W_2 = 0.0$, $W_3 = 1.0$, $W_4 = -1.0$, $\sigma_R = 0.5$.

The deviation of an agent's estimate from the true value, as illustrated in the figure, 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 between two compound

stimuli (because the deviations might be similar for the two stimuli). Table 1 shows the expected reward loss per trial for the different learning approaches, split into four equal parts of the learning sequence. The expected reward per trial for an ideal case is $\bar{R} = 2.10$, so the losses in the table are around one percent.

Table 1: Expected reward loss per trial, compared to an ideal case where an agent knows the true expected reward for each presented compound stimulus. The loss is presented for different parts of the sequence of trials.

Trial interval	RW	PM	R1	IDBD
1 to 1000	0.024	0.028	0.024	0.024
1001 to 2000	0.024	0.026	0.023	0.023
2001 to 3000	0.027	0.047	0.020	0.021
3001 to 4000	0.019	0.023	0.018	0.019

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 (Figure 1). In particular, the Pearce and Mackintosh (2010) unified model did poorly, mainly because the learning rates were at the different limits for most of the time. The Sutton (1992b) R1 approach seems to win the competition, but the difference from Rescorla-Wagner was not very big.

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.

Gershman, S. J. 2015. A unifying probabilistic view of associative learning. *PLOS Computational Biology*, 11(11):e1004567.

Le Pelley, M. E. 2004. The role of associative history in models of associative learning: a selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology Section B*, 57(3b):193–243.

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.

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.

Piray, P. and Daw, N. D. 2021. A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, 12(1):6587.

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