Reinforcement Learning

Peter Dayan
Gatsby Computational Neuroscience Unit
UCL
dayan@gatsby.ucl.ac.uk

Abstract

Reinforcement learning studies the prediction and control of events of affective importance in terms of psychological and neural rules for adaptation. Reinforcement learning began as a marriage between ideas in mathematical behavioral psychology and artificial intelligence, and now has links all the way from neuromodulatory systems in vertebrates to engineering and statistical theories of adaptive optimizing control. In this chapter, we describe the basic theory underlying reinforcement learning, and its links with neuroscience, psychology, statistics and engineering.

1 Introduction

Figure 1 shows a very simple maze problem that might be posed to a hungry rat, or even a robot, involving the choice of direction to run at three choice points A, B and C. The rat repeatedly starts at location A, runs forward through the maze, and ends at one of the four shaded boxes. When it arrives at a termination position, it is awarded a random number of food pellets whose average is written in the box. Then it is then is replaced at A, only to start again.

This deceptively straightforward maze poses a sequential decision-making task in microcosm. We use it to discuss the standard techniques of reinforcement learning, and their links with classical and instrumental conditioning. We seek to use formal mathematical and computational descriptions to model the nature and properties of behavior in such tasks.

First, consider the decision the rat has to make when it arrives at B or C. In either case, it faces a standard T-maze task with appetitive outcomes. Such tasks are also known as stochastic two-armed bandit problems, where the equivalent of pulling one of the arms is choosing one of the directions, and the stochasticity comes from the randomness in the number of pellets actually delivered in each shaded box. The rat has to learn from its random experience that it gets more food on average if it runs left at B and right at C. Furthermore, it may need to consider the possibility that the experimenter might change the relative worths of the termination boxes, in which case it would have continually to explore the different possibilities.

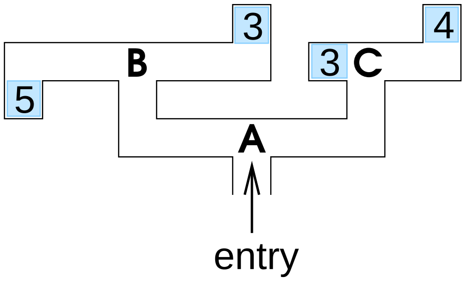

Figure 1: The maze task. The figure shows a simple maze task. The rat (or robot) enters from the bottom, and is only allowed to run forwards. The shaded boxes are end points where the rat receieves food pellets. The average number of pellets is written in each box (with Poisson variability in our experiments). The potential choice points are at A, B and C. After Dayan & Abbott (2001).

Second, consider the more difficult case of the choice at location A. Whichever direction the rat chooses, it receives no direct reward; the only consequence is that it gets to location B or C. Information that location B is a better choice than C has somehow to filter back to the choice of direction at A, in order that the rat can learn to go left. This is sometimes known as the *temporal credit assignment problem,* since, if the rat goes left at A and right at B, how does it know that the first direction was good and the second bad, rather than *vice-versa.* Credit (or, in this case, blame) has somehow to be assigned to the second direction. In more challenging sequential decision tasks, there may be many steps before an actual reward is delivered.

Both these problems are familiar in a variety of fields, including psychology, statistics and control theory. Reinforcement learning (see Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996) has blended and adapted solutions that have been suggested in each. In summary, the essential idea for solving the first problem is to consider the model rat as adopting a parameterized stochastic *policy* at B and C, which specifies the probabilities with which it turns left or right at each location. The parameters of the policy are adjusted in the light of the rewards received by the rat, so that actions leading to larger average rewards come to be favored. This is essentially a formalization of Thorndike's (1898; 1911) famous law of effect. An alternative is to sample both directions in order to learn the average reward associated with each, and then choose whichever seems more lucrative. These methods are close to those suggested by Bush & Mosteller (1955; and, more distantly, Rescorla & Wagner, 1972), formalized in the engineering discipline of stochastic learning automata (Narendra & Thatachar, 1989), and analyzed in a reinforcement learning context by Williams (1992).

The essential idea for solving the second problem is to learn the attractiveness of locations B and C, and choose an action at A that leads to the more attractive outcome. Here, the natural measure of attractiveness of the location, which is usually called its *value*, is the number of pellets that is expected to be delivered in the future path through the maze, starting from that location. Thus, going left at A, which leads to a location (B) from which a large number of pellets is expected in the future, should be favored over going right, which leads to a location (C) from which only a small number of pellets is expected. Of course, the values of B and C depend on the policies adopted there; as these policies change, the action that is preferred at A can change too. However, provided that the polices are always getting better, so the values of B and C only increase, then the policy at A will change for the better too. In turn, the value of A, the expected number of pellets associated with starting from A, can also be learned, using information encountered about the expected number starting from B and C.

There has historically been great interest in psychology in the way that chains of behavior, like the successive choices of direction in the maze, are established and maintained by reinforcement, and this method for solving the second problem therefore has a large number of forebears. One is secondary conditioning – since the value of location A is acquired via a secondary association with the rewards that follow B and C. Values acquired during secondary conditioning are used to control instrumental choices of action (resulting in a form of two-factor conditioning theory, *eg* Mowrer, 1956; Mackintosh, 1974). A second link is to a method suggested by Deutsch (1960) as to how information about possible future rewards in the maze could come to influence present actions. There is a slightly more remote relationship with Hull's suggestion of goal gradients (Hull, 1943; 1952). Another link is to Samuel's (1959) checker playing program and Michie & Chambers' (1968) BOXES control system, since the expectation of future reward starting from A is determined by making it consistent with the expectations of future reward starting at the locations accessible from A, namely B and C. A further link, which we explain in more detail below, is to the engineering method of dynamic programming via policy iteration (Bellman, 1957). Finally, albeit in a slightly different setting, ethologists use dynamic programming to formalize behaviors such as these in the context of optimal foraging decisions (Mangel & Clark, 1988), and this provides a link between psychological and ethological ideas about learning. In reinforcement learning, the method was suggested and analyzed by Sutton (1988); Barto, Sutton & Anderson (1983); Sutton & Barto (1990), and linked to dynamic programming by Watkins (1989) and Werbos (1990).

Many drugs of addiction hijack the mechanisms normally employed by animals to process appetitive information. This has led to the collection of a substantial body of experimental data in rats, monkeys and humans on the neural underpinning of adaptive action choice, which bears directly on the neural basis of reinforcement learning. In particular, the neuromodulator dopamine is thought to play a critical role in the appetitive aspects of addictive drugs, and so is a natural candidate for the neural substrate for aspects of reinforcement learning. Consistent with this, as we will see, measurements of the activity of dopaminergic neurons during conditioning show that they exhibit some of the characteristic properties of the reinforcement learning signal involved in propagating information on prediction errors about future rewards to train present predictions and control present actions. Altogether, reinforcement learning melds neural, psychological, ethological and computational constraints on a single problem.

In this chapter, we first consider the two control problems described above in the context of the maze, and show the reinforcement learning solution and its links with more standard accounts of classical and instrumental conditioning. In section 4, we discuss the putative neural basis, providing a model of the activity of dopamine cells. Finally, in section 5, we consider some current efforts to model attentional aspects of conditioning. Whereas the standard reinforcement learning models are strongly influenced by engineering ideas about optimal control; the attentional models are influenced by statistical ideas about learning.

2 Action Choice

Choosing a direction at B or C is more straightforward than at A, because immediate information is provided about the quality of the action in terms of the delivery of pellets of food. The main difficulty comes if the pellets are delivered in a stochastic manner at each termination point, so averages have to be taken over many trials, or if the experimenter actually changes the average values over time. These lead to a critical choice for the rat between *exploration* and *exploitation,* that is, between choosing the action which is currently believed to be worse in order to make sure that it really is, and choosing the action that is currently believed to be best in order to take advantage of its apparently superior characteristics. A fair wealth of theoretical work has been devoted to optimal tradeoffs between exploration and exploitation (*eg* Berry & Fristedt, 1985), and there is even some evidence of this sort of 'meta-optimization' in animals (Krebs, Kacelnik & Taylor, 1978). However, we shall confine ourselves to the simpler problem of learning out which direction is best when the averages are fixed.

We consider two simple methods for choosing directions. One is based on parameterized stochastic policies, whose parameters are changed to increase the average number of pellets. This is sometimes called a *direct* method, since the policy is directly specified and improved. The second is based on learning the average number of pellets provided by each direction, and then preferring the direction associated with the greatest mean reward. This is called an *indirect* method, since the policy is an indirect function of the estimated mean rewards. For convenience, we start by considering just location B, and so omit an argument or index indicating the location at which the choice is being made.

The Direct Actor

In the direct method, the choice between the directions is specified by a stochastic policy. Let π_{L} be the probability of turning left at B, and $\pi_{\mathrm{R}} = 1 - \pi_{\mathrm{L}}$ be the probability of turning right there. A natural way to parameterize these is to use the sigmoid rule

$$\pi_{\mathrm{L}} = \sigma(\beta m) \; , \tag{1}$$

based on an action value m, where $\sigma(x) = 1/(1 + \exp(-x))$ is the standard logistic sigmoid function. This is actually a special form of the softmax or Luce choice rule

(Luce, 1959). If m is positive, then the larger its magnitude, the more likely the rat is to go left at B. If m is negative, then the larger its magnitude, the more likely the rat is to go right at B. Here, the value of β sets the scale for the action value, and, by controlling the frequency with which the rat samples the alternative believed to be worse, controls the trade-off between exploration and exploitation.

Say the rat chooses to go left on a particular trial, and receives a reinforcement of r_L pellets of food. How should it use this information to change the parameter m to increase the expected number of pellets it will receive? Intuitively, if r_L is large, then m should be increased, since that makes π_L larger; if r_L is small, then m should be decreased, since that makes π_L smaller. One way to achieve this (Williams, 1992) is to change the parameter according to

$$m \to m + \epsilon(r_L - \tilde{r})(1 - \pi_L) \ , \tag{2}$$

together with the equivalent expression if the rat chooses to go right instead

$$m \to m - \epsilon(r_R - \tilde{r})(1 - \pi_R) \tag{3}$$

on receiving a reward of r_R), then the number of pellets the rat can expect to receive will increase, at least on average. Here ϵ is a (small) *learning rate,* and $\tilde{r}$ is called a *reinforcement comparison* term which sets a standard against which the actual number of reward is compared.

We should step back at this point to consider the implications of equation 2. This is an algorithmic learning or adaptation rule for the parameter m, indicating how it should change in the light of experience. It was derived from the computationally sound intent of increasing the average reward. As we see in the rest of this chapter, in reinforcement learning, many such algorithmic and computational proposals about the maximization of reward in more complicated circumstances are studied.

Rule 2 makes perfect sense – if the actual number of pellets delivered is greater than the comparison term, then m should be increased; if it is less than the comparison term, then m should be decreased. Oddly, provided the reinforcement comparison term $\tilde{r}$ does not depend on whether the rat chose to go left or right, then the fact that the average number of pellets increases does *not* depend on its actual value. For an intuition as to why this might be so, consider the case that $\tilde{r}$ is very large and positive. Then, if the rat goes left on a trial, then m tends to decrease according to equation 2, even if going left is better than going right. However, if the rat goes right on a trial, then, according to equation 3, m will increase by an even larger amount, and so m will still tend to grow on average. The value of $\tilde{r}$ does not affect the average outcome of the learning rule. Nevertheless, the value of $\tilde{r}$ does, however, affect the variance of this learning rule, and so can determine subsidiary factors such as the speed of learning. A natural value of $\tilde{r}$ is just the average reward across all directions.

Figure 2 shows the course of learning to go left at B using the direct actor. Here, the rewards have Poisson distributions, with means 5 and 3 for going left and right respectively. Figure 2A shows π_L for one particular session; the randomness is clear.

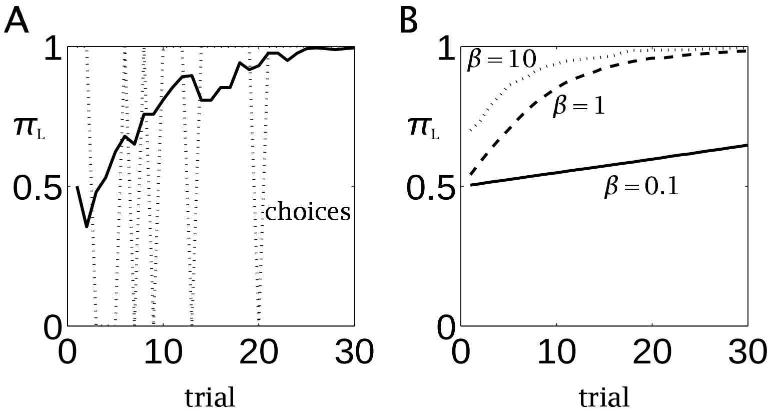

Figure 2: Simulation of the direct actor. A) The probability π_L of going left over 30 trials at location B (solid line) and the actual choices of direction (dotted line in a single session; 1 means left; 0 means right). Here $\beta = 1, \epsilon = 0.4, \tilde{r} = 4$ and the number of pellets has a Poisson distribution. B) The average probabilities (over 500 runs) of going left at B for $\beta = 0.1, 1, 10$.

Figure 2B shows the average behavior for various values of β. For this simple problem, larger values of β lead to faster learning. However, if β (or indeed ϵ) is too large, then the rat can be overly tempted by the first reward it experiences, and might never explore to find a better direction. The extent to which this is true depends on the actual mean and variance of the reward delivery, something about which the rat is likely to have only vague expectations at the start of learning.

Altogether, the direct actor is a very straightforward learning controller for which there is a relatively easy way of controlling the trade-off between exploration and exploitation. We use it later for our full model for the maze, in which case we need different action values for each location. We will call these $m(\mathsf{A})$, $m(\mathsf{B})$ and $m(\mathsf{C})$.

The Indirect Actor

The indirect actor uses the same action choice rule as the direct actor

$$\pi_L = \sigma(\beta(q_L - q_R)) \ , \tag{4}$$

but with two parameters q_L and q_R that have a completely different semantics from m in equation 1. The idea is that q_L should come to be the average number of pellets provided for going left, and q_R the average number for going right. In this case, equation 4 automatically favors the direction leading to the greater reward, where the size of the difference is judged according to β.

Say, as above, that the rat chooses from this distribution to go left, and receives a reinforcement of r_L pellets of food. How should the estimate of the average return of going left be altered? An obvious suggestion is to use

$$q_{\mathrm{L}} \rightarrow q_{\mathrm{L}} + \eta(r_{\mathrm{L}} - q_{\mathrm{L}}) \tag{5}$$

where η is again a learning rate. This tends to increase q_{L} if the estimate is too low, and increase it if the estimate is too high. Alternatively, it moves q_{L} towards a target value r_{L}, by altering q_{L} according to the *prediction error* $\delta = r_{\mathrm{L}} - q_{\mathrm{L}}$. This learning rule is a simple form of the Rescorla-Wagner rule (Rescorla & Wagner, 1972). The only difference from standard applications of the Rescorla-Wagner rule is that the target value r_{L} here is stochastic, whereas in most standard cases it is deterministic (and is often written as λ, the asymptotic value to which the association tends over trials). In this case, the equivalent asymptotic value is the mean reward for going left, which we write as $\langle r_{\mathrm{L}} \rangle$.

Sutton & Barto (1981) pointed out that this Rescorla-Wagner rule is a simple form of the least mean squares or delta rule in engineering (Widrow & Hoff, 1960). There is a substantial body of theory (see Widrow & Stearns, 1985) about how the delta rule will lead q_{L} to approximate the average value of the reward associated with going left, and which can be used for such tasks as understanding the consequences of different settings of η. An equivalent rule can be used to train q_{R} to estimate the average reward associated with going right. As for the direct actor, we later need different values q_{L} and q_{R} for each locations. Where necessary, we write these *state-action* values as $q_{\mathrm{L}}(\mathsf{X})$ and $q_{\mathrm{R}}(\mathsf{X})$, where $\mathsf{X} = \mathsf{A}, \mathsf{B}, \mathsf{C}$ labels the location.

We will later consider methods that use the average number of pellets $v(\mathsf{B})$ associated with starting from B, averaging over the randomness of the choice of direction as well as the outcome at each end-point. This can also be learnt using exactly the same learning rule

$$v(\mathsf{B}) \rightarrow v(\mathsf{B}) + \eta(r - v(\mathsf{B})) \tag{6}$$

where r is the number of pellets received, whichever direction the rat chooses. Sometimes, $v(\mathsf{B})$ is called the *value* of B.

Although the direct and indirect actors look very similar, they actually have quite different interpretations and end points. If, as in figure 1 at B, going left is more lucrative on average than going right, then m for the direct actor should tend to increase without bound towards ∞, albeit extremely slowly because the increments are scaled by the ever-decreasing probability that the rat chooses to go right. However, $q_{\mathrm{L}} - q_{\mathrm{R}}$ for the indirect actor should come to take on the value of the average difference in worth between the two directions, even if this difference is very small. This makes the role for β very different in the two algorithms. For the direct actor, if β is fixed and $\tilde{r}$ is constant, then going left will generally be favored over going right to an arbitrary degree, even if left is only a little better than right. However, the indirect actor will tend to match its behavior more closely to the average outcomes. The preferred direction will only come to be chosen on every trial if β is set to be very large. Albeit in a rather more complicated reward context, Egelman, Person & Montague (1998) showed that the propensity of this rule towards matching made it a good model of human choice behavior.

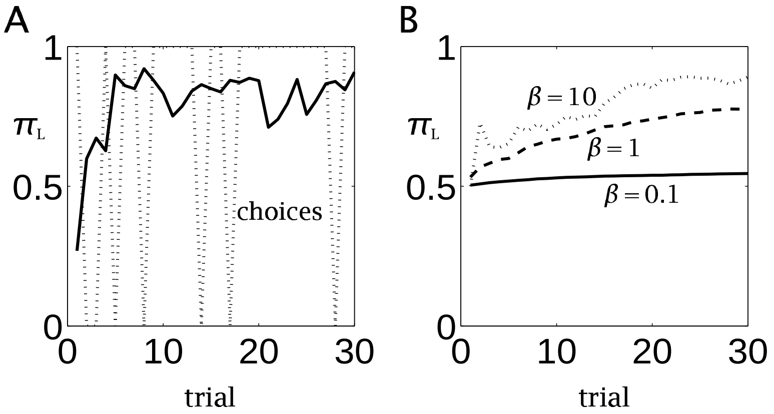

Figure 3: Simulation of the indirect actor. A) The probability π_L of going left over 30 trials at location B (solid line) and the actual choices of direction (dotted line; 1 means left; 0 means right). Here $\beta = 1, \eta = 0.2$ and the number of pellets has a Poisson distribution. B) The average probabilities (over 500 runs) of going left at B for $\beta = 0.1, 1, 10$.

For the same task as figure 2, figure 3 shows the course of learning to go left at B using the indirect actor. As there, figure 3A shows a sample of the randomness in a single run and figure 2B shows the mean effect over many runs for a few different values of β. The effect described above is at work in the persistently poor performance for small $\beta = 0.1$. Whereas for the direct actor, in figure 2B, the probability of turning left continually increases, for the indirect actor, it has an asymptotic value of $\pi_L = 0.55$. If β is too large, then, like the direct actor, the balance between exploration and exploitation can be disturbed.

3 Sequential Action Choice

It is clear that neither direct nor indirect actors can directly be employed to choose which direction to take at A, since the equivalent of the reward terms r_L or r_R are always 0. The task here is closely analogous to the intensively studied case of sequential chains of behavior, reinforced by a final goal. The investigations of this led to sophisticated, complicated, and contentious learning theories, such as those of Hull (see Hull, 1952), and Deutsch (1960). At the heart of the reinforcement learning solution to the temporal credit assignment problem is the very simple and old idea of using secondary reinforcement to learn values $v(\mathsf{B})$ and $v(\mathsf{C})$ which reflect the attractiveness of locations B and C, and then choosing the direction at A that leads to the more attractive location. Here, $v(\mathsf{B})$ acts like a surrogate reward for the choice of going left at A, and $v(\mathsf{C})$ acts like a surrogate reward for the choice of going right at A.

The main novelty (Sutton, 1988; Sutton & Barto 1990) in this approach lies in the way that secondary reinforcement is formalized and thus in the definition of the values. The suggestion is that the value of a location should come to be the total reward expected starting from that location, adding up over all future steps. The expectations of total future reward should be mutually consistent at successive steps, so, for instance, if the rat always chooses to go left at A, then the value of A should be the same as the value of B. The temporal difference learning rule is based on the discrepancy in the value estimates at successive steps.

Given these values, the actual reward term r_{L} in in equations 2 and 4 might simply be replaced by the expected future reward $v(\mathsf{B})$, and learning a policy at A using either direct or indirect actors could proceed. Of course, $v(\mathsf{B})$ and $v(\mathsf{C})$ will change as the policy at B and C improves, but this should only improve the choice of direction at A.

We first consider a more formal treatment of the problem of learning values, including what to do if pellets can be provided at any step. We then discuss how the information about values can be used to learn a good policy at A.

The Critic

The values of locations are the expected total number of pellets that will be received, starting from each location, and following a fixed, though possibly stochastic, policy. Once the policy is fixed, the problem of learning the values can be seen as one for classical conditioning, *ie* predicting the rewarding consequences of locations without reference to the actions actually adopted. This makes the link to secondary conditioning clearer – B is directly associated with the delivery of pellets; A is directly associated with getting to B, and so A should be associated with the delivery of pellets too. Indeed, the reinforcement learning method, called *temporal difference* learning (Sutton, 1988) that we are about to describe for learning the values, was explicitly motivated by the failure of secondary conditioning to come under the umbrella of the Rescorla-Wagner rule. This failure comes about because the Rescorla-Wagner rule does not take account of the sequence of presentation of the stimuli (here, the sequence of locations in the maze).

Unfortunately, though the idea underlying temporal difference learning is simple, the notation is rather ugly. Starting from any location X, there are only two ways to get pellets. One is that some might be immediately provided for choosing a direction (as at B and C). These are called immediate rewards, and, as above, we write $\langle r_{\mathrm{L}}(\mathsf{X})\rangle$ and $\langle r_{\mathrm{R}}(\mathsf{X})\rangle$ for the average numbers of pellets provided for going left and right at location X. More generally, we write $\langle r_{\mathrm{D}}(\mathsf{X})\rangle$ for the average number of pellets provided for choosing a direction D, and this allows for the possibility that pellets might be provided at intermediate locations in the maze on the way to the end points. The other way to get pellets is that they might be expected to be provided in the farther future, based on the location Y the rat gets to directly from X on account of its action. For instance, if the rat gets from A to B, then it can expect to get pellets in the future, following the action it takes at B. Once learning is complete, $v(\mathsf{Y})$ is exactly the number it can expect to get from Y. In fact, the expected future reward from X is simply the sum of these two contributions, averaged over the choice of action at X. Sutton & Barto's (1990) key insight was to introduce the idea of predicting the sum of all future rewards, rather than just the current reward, and this led directly to temporal difference learning.

To put this more concretely, we can write a formula for the value of A as

$$\begin{aligned} v(\mathsf{A}) &= \pi_{\mathrm{L}}(\mathsf{A})\,(\langle r_{\mathrm{L}}(\mathsf{A})\rangle + v(\mathsf{B})) + \pi_{\mathrm{R}}(\mathsf{A})\,(\langle r_{\mathrm{R}}(\mathsf{A})\rangle + v(\mathsf{C})) && (7)\\ &= \pi_{\mathrm{L}}(\mathsf{A})v(\mathsf{B}) \qquad\qquad\quad + \pi_{\mathrm{R}}(\mathsf{A})v(\mathsf{C}) && (8)\end{aligned}$$

In equation 7, we have written out explicitly the two contributions mentioned above and averaged them over the probabilities of going left and right at A. Equation 8 follows, since no pellets are immediately provided for the action at A. Similarly, we can write the value of B as

$$v(\mathsf{B}) = \pi_{\mathrm{L}}(\mathsf{B})\langle r_{\mathrm{L}}(\mathsf{B})\rangle + \pi_{\mathrm{R}}(\mathsf{B})\langle r_{\mathrm{R}}(\mathsf{B})\rangle = \pi_{\mathrm{L}}\langle r_{\mathrm{L}}\rangle + \pi_{\mathrm{R}}\langle r_{\mathrm{R}}\rangle = \pi_{\mathrm{L}}5 + \pi_{\mathrm{R}}3 \qquad (9)$$

where, in this case, the contributions to the number of pellets from the future steps in the maze (the equivalents of $v(\mathsf{B})$ and $v(\mathsf{C})$ in equation 7) are zero, since the trial ends in one of the shaded boxes. We have also included the reduced notation of the previous section.

If we were to follow the Rescorla-Wagner rule of equation 6 to learn $v(\mathsf{A})$, we might use the right hand side of equation 7 as the target for $v(\mathsf{A})$ (replacing the r), and use the difference between the target value and the current value $v(\mathsf{A})$ to drive learning

$$v(\mathsf{A}) \to v(\mathsf{A}) + \eta\Big(\Big[\pi_{\mathrm{L}}(\mathsf{A})\,(\langle r_{\mathrm{L}}(\mathsf{A})\rangle + v(\mathsf{B})) + \pi_{\mathrm{R}}(\mathsf{A})\,(\langle r_{\mathrm{R}}(\mathsf{A})\rangle + v(\mathsf{C}))\Big] - v(\mathsf{A})\Big) \qquad (10)$$

There are three problems with equation 10, all of which temporal difference learning addresses.

The first problem with equation 10 is that the target value for $v(\mathsf{A})$ involves knowing the true values $v(\mathsf{B})$ or $v(\mathsf{C})$. Of course, at the beginning of learning, these are no better known than $v(\mathsf{A})$. Temporal difference learning ducks this concern, by *bootstrapping* its estimates from each other. That is, learning proceeds for all locations simultaneously, and as $v(\mathsf{B})$ and $v(\mathsf{C})$ become more accurate, then $v(\mathsf{A})$ becomes more accurate too.

The second problem is that the learning rule involves an average over both the possible directions the rat could choose at A, rather than just the single direction the rat can actually choose on a single trial. Gallistel (personal communication, 2001) calls this the problem of the 'road not taken'. Fortunately, just as we use the random number of pellets r_{L} received for going left on a single trial in equation 6 and still make q_{L} be the average number of pellets consequent on turning right $\langle r_{\mathrm{L}}\rangle$, we can use the actual value $v(\mathsf{B})$ or $v(\mathsf{C})$ encountered by the rat on a single trial, and still learn the average of equation 8.

In order to have a slightly more general description of this, we need to use the more general way of describing a single step in the maze that we mentioned above. Consider the case that the rat starts at a location X, picks a random direction D according to $\pi_{\mathrm{L}}(\mathsf{X})$ and $\pi_{\mathrm{R}}(\mathsf{X})$, receives an immediate reward of $r_{\mathrm{D}}(\mathsf{X})$ pellets and moves to location Y. Here, $r_{\mathrm{D}}(\mathsf{X}) + v(\mathsf{Y})$ is a quantity that is random, since the choice of action D and the delivery of $r_{\mathrm{D}}(\mathsf{X})$ pellets are both random. Starting from $\mathsf{X} = \mathsf{A}$, if direction D is left, then Y is B; if direction D is right, then Y is C. Averaging over these sources of stochasticity, the mean value of $r_{\mathrm{D}}(\mathsf{A}) + v(\mathsf{Y})$ is

$$\pi_{\mathrm{L}}(\mathsf{A})\left(\langle r_{\mathrm{L}}(\mathsf{A})\rangle + v(\mathsf{B})\right) + \pi_{\mathrm{R}}(\mathsf{A})\left(\langle r_{\mathrm{R}}(\mathsf{A})\rangle + v(\mathsf{C})\right) \tag{11}$$

which is just the right hand side of equation 7. This is also true for general moves in the maze. Thus, $r_{\mathrm{D}}(\mathsf{X}) + v(\mathsf{Y})$ can be thought of as a random sample of the value of $v(\mathsf{X})$ in the same way that r_{L} in equation 5 is a random sample of $\langle r_{\mathrm{L}}\rangle$. The overall learning rule uses this sample to replace the expression in the square brackets of equation 10. That is it uses the difference

$$\delta = r_{\mathrm{D}}(\mathsf{X}) + v(\mathsf{Y}) - v(\mathsf{X}) \tag{12}$$

between the sampled target $(r_{\mathrm{D}}(\mathsf{X}) + v(\mathsf{Y}))$ and current $(v(\mathsf{X}))$ values of X to drive learning, just as the Rescorla-Wagner rule of equation 5 uses the difference between the sampled target (r_{L}) and current (q_{L}) state-action values. The final learning rule is

$$v(\mathsf{X}) \to v(\mathsf{X}) + \eta\delta \tag{13}$$

The difference δ of equation 12 is called the *temporal difference* and plays a central role in both biological and engineering aspects of reinforcement learning. Its name derives from the difference $v(\mathsf{Y}) - v(\mathsf{X})$ between the estimates of the values of two successive states.

The third concern is that a straightforward implementation of equation 13 seems to require that the rat have a very simple representation of each location, with a single parameter $v(\mathsf{X})$ for each location. As is completely standard in theories of conditioning, we might more reasonably expect X to be represented by the activity of a set of atomistic stimulus representation or cue units $(u_1, u_2, \ldots, u_n) = \mathbf{u}$, whose activity at X, called $\mathbf{u}(\mathsf{X})$ might be determined by the unique cues at each location in the maze, or perhaps the activity of place cells associated with the locations. Associated with each cue unit u_i is a weight or parameter w_i that determines its contribution to the values of locations at which the cue is active. As in the Rescorla-Wagner rule in such circumstances, the value of location X, $v(\mathsf{X})$, is a sum over the weights of just the cues that are active at X, and only the weights associated with active cues are changed in the face of prediction errors.

An easy way to put this version of the Rescorla-Wagner rule more formally is to assume that a stimulus representation unit $u_i(\mathsf{X})$ takes the value 0, if the cue is not present at X, or 1 if the cue is present. Then, because of this simple binary form, the Rescorla-Wagner rule's assumption that the value of X is determined as a sum over the cues that are active there, can be written

$$v(\mathsf{X}) = \sum_i w_i u_i(\mathsf{X}) = \mathbf{w} \cdot \mathbf{u}(\mathsf{X}) \ . \tag{14}$$

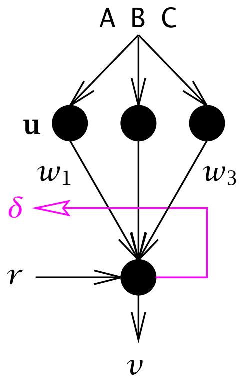

Figure 4: The architecture of the critic. The stimulus representation **u**, which here has one element for each location in the maze, maps to the value estimate v through a set of modifiable weights **w**. The prediction unit incorporates information about the delivered rewards r and calculates the temporal difference error δ, which is used to change the weights.

Further, the extension to the temporal difference learning rule of equation 13 specifies changes to the weights for just the cues that are active at X

$$w_i \to w_i + \eta \delta u_i(\mathsf{X}) \tag{15}$$

This rule shares most of the properties of the Rescorla-Wagner rule to do with stimulus competition (such as blocking), and also shares many of its faults. Figure 4 shows the learning architecture in a different way, where, in this case, there are just three cues and three weights ($n = 3$).

The extensive theory of temporal difference learning (see Sutton & Barto, 1988; Bertsekas & Tsitsiklis, 1996) indicates circumstances under which this learning rule will make the values $v(\mathsf{X})$ come to satisfy equation 8. More concretely, figure 5 shows the course of learning of the values of A, B and C (solid lines), together with their target values (dashed line) in a single session of 100 trials in the maze of figure 1 for the case that the rat chooses left and right equally often. It is apparent that the values are learned quickly and accurately.

The Sequential Actor

We have so far shown how to learn the values of the locations as the expected future rewards starting from those locations. How can we use this information to learn a good policy at location X? The idea, just as in equations 12 and 13, is to use $r_{\mathrm{D}}(\mathsf{X}) + v(\mathsf{Y})$ as a sampled estimate of the value of performing action D. This sampled estimate can just replace r_{L} in the learning rules for the direct and indirect actor, as we discuss in the next two sections.

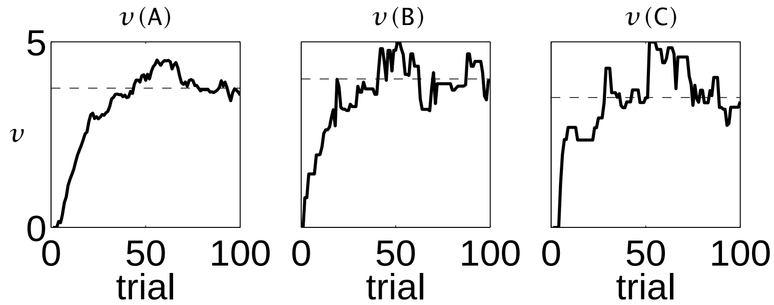

Figure 5: Learning the values. The solid lines show the course of learning of $v(\mathsf{A}), v(\mathsf{B})$ and $v(\mathsf{C})$ over a single session of 100 simulated runs through the maze. The dashed lines show the true values from equation 8. Here $\epsilon = 0.2$.

The Actor-Critic

In discussing the direct actor, we introduced the idea of the reinforcement comparison term $\tilde{r}$, and suggested that it might take on the value of the average reward. In this case, an estimate of the current average reward is just $v(\mathsf{X})$, and so, for instance, learning rule 2 becomes

$$m(\mathsf{X}) \rightarrow m(\mathsf{X}) + \epsilon(r_{\mathrm{L}}(\mathsf{X}) + v(\mathsf{Y}) - v(\mathsf{X}))(1 - \pi_{\mathrm{L}}) = m(\mathsf{X}) + \epsilon\delta(1 - \pi_{\mathrm{L}}) \qquad (16)$$

and depends on the values only through the temporal difference error δ of equation 12. Therefore, the same temporal difference error term is used to train both the value of X *and* the policy at X. To put it another way, $\delta > 0$ means either that the estimate of the value $v(\mathsf{X})$ is too pessimistic, or that the estimate is right on average, but the action D is better than average (or both). In the first case, the estimate should ideally be changed; in the second, the policy should be changed. Figure 6 shows the course of learning of actions at A, B and C in the maze. The rat rapidly learns to turn left at A and B, and also to turn right at C. This combination of temporal difference learning and the direct actor was first suggested by Barto, Sutton & Anderson (1983) under the name of the actor-critic architecture.

In the case that stimulus units $u_i(\mathsf{X})$ represent the locations in the maze, then, as in equation 14, a set of action weights m_i can be used to parameterize the action choice at X

$$m(\mathsf{X}) = \sum_i m_i u_i(\mathsf{X}) \qquad (17)$$

The equivalent of the actor learning rule 16 is

$$m_i \rightarrow m_i + \epsilon\delta(1 - \pi_{\mathrm{L}})u_i(\mathsf{X}) \; . \qquad (18)$$

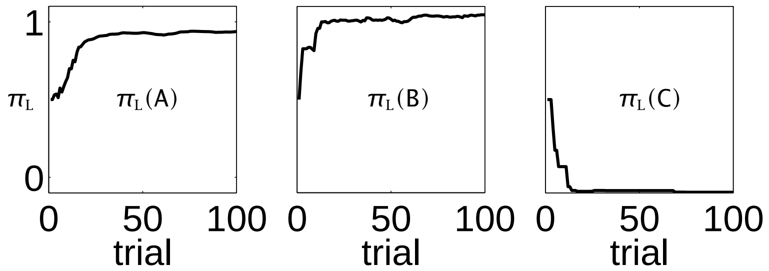

Figure 6: Learning the policies. The lines show the course of learning $\pi_L(\mathsf{A}), \pi_L(\mathsf{B})$ and $\pi_L(\mathsf{C})$ over a single session of 100 simulated runs through the maze using temporal difference learning and the direct actor. Here $\epsilon = 0.4$.

Q-learning

It would be possible to use exactly the indirect actor as described in section 1 in conjunction with the estimate $r_D(\mathsf{X}) + v(\mathsf{Y})$ of the future rewards from a location. However, this is not done, because of the redundancy between learning the average value of each action at each location $q_D(\mathsf{Y})$ and, *separately,* the average value of the location $v(\mathsf{Y})$ itself. Rather, the conventional alternative is to define $v(\mathsf{Y})$ in terms of the state-action values $q_D(\mathsf{Y})$. There are two ways to do this. Since the state-action values define an explicit policy, $\pi(\mathsf{Y})$, one way is to define $v(\mathsf{Y})$ as a weighted average of the state-action values, where the weights are determined by the policy. However, the object of learning is to find the actions that maximize the expected reward. Since the rat is free to choose which direction to run at Y, it is appropriate to evaluate Y not according to the estimated average worth of all the possible directions at Y, but rather by the estimated worth of the *best* direction at Y. The second way to determine $v(\mathsf{Y})$ is thus

$$v(\mathsf{Y}) = \max_D q_D(\mathsf{Y}) \tag{19}$$

Whichever form for the value of Y is used, given a choice of action D, and a transition from location X to Y, the Q values should be updated as

$$q_D(\mathsf{X}) \to q_D(\mathsf{X}) + \eta\left(r_D(\mathsf{X}) + v(\mathsf{Y}) - q_D(\mathsf{X})\right) \tag{20}$$

Equation 20 is almost the same as the standard temporal difference rule of equation 13. This form of reinforcement learning is called Q-learning, and was invented by Watkins (1989).

Dynamic Programming

Markov decision problems (Puterman, 1994) provide a rich theoretical context for understanding sequential action choice tasks such as the maze. Methods in engineering for solving Markov decision problems are generally forms of a technique called dynamic programming (Bellman, 1957), to which reinforcement learning algorithms are closely related. For instance, the actor-critic implements a dynamic programming technique called *policy iteration*, involving policy evaluation (finding the values $v(\mathsf{X})$ that satisfy equation 8 and policy improvement (using the values to find actions that are better). Q-learning implements a dynamic programming technique called *value iteration.*

The link with dynamic programming has led to comprehensive theory (see Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996) as to circumstances under which Q-learning will find optimal policies in tasks like the maze. It has also opened many lines of investigation extending and improving the basic methods outlined here.

4 Dopamine and Reinforcement Learning

Unfortunately, there has been no complete investigation of the neural basis of the learning behavior of rats or other animals in a maze such as that in figure 1. However, as is evident in many other chapters in this book, a substantial body of neurobiological data, based on drug addiction, self-stimulation, lesion studies and neuropharmacology bears on how animals actually learn about rewards. A venerable organizing focus of this work has been that the neuromodulator dopamine plays a central role (see, for instance, Wise, this volume; Wise, 1982; Beninger, 1983; Wise & Bozarth, 1984; Simon & le Moal, 1988; Wise & Rompre, 1989; Robbins & Everitt, 1992; Koob, 1992). Note, however, that this is an active and controversial field of study, and there are also many recent challenges (see Berridge, & Robinson, 1998; Spanagel & Weiss, 1999; Ikemoto & Panksepp, 1999). Temporal difference learning has been used to model the activity of dopamine cells during reward learning, a link we describe in this section. The relevant dopamine cells are both in the ventral tegmental area, which project to areas associated with the limbic system, including the nucleus accumbens or ventral striatum, and the those in the substantia nigra pars compacta, which project to the dorsal striatum.

Figure 7 shows a histogram of the activity of dopamine cells of a thirsty macaque monkey in conditions before and after it learns a predictive association between a stimulus (which cues a motor response) and a reward (which is a drop of juice). These actually come from different experiments, but the qualitative pattern of results is robust. In figure 7A, left and right plots show activity stimulus-locked to the cue and the reward respectively. The cue is irrelevant before learning, and there is little response to it, whereas delivery of the reward induces a large and well-timed response from the cells. Figure 7B shows that after learning, the contingencies reverse – now the activity is consequent on the early reliable predictor and not the reward itself.

Figure 8 describes the account of this activity that comes from temporal difference learning (Montague, Dayan & Sejnowski, 1996; Houk, Adams & Barto, 1995; Friston *et al,* 1994; Schultz, Dayan & Montague, 1997; Wickens, 1993). The idea is that the difference between the activity and the baseline activity reports the temporal difference prediction error δ (equation 12). In early trials, the reward is unexpected, and so the prediction

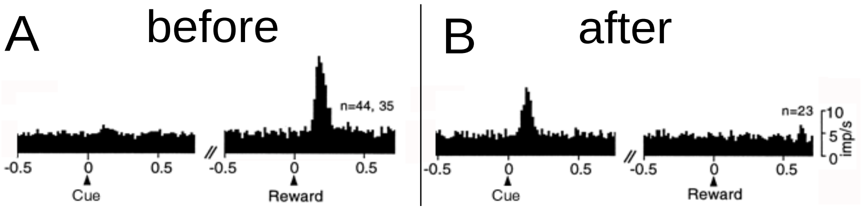

Figure 7: Histograms of population activity of dopamine cells stimulus-locked to a cue and to the reward during a conditioning task. A) Dopamine activity before learning shows response to the reward not the cue. B) Dopamine activity after learning shows response to the cue and not the reward. Adapted from Schultz (1998).

error, and so the model dopamine activity, follows the reward signal r. In later trials, the activity associated with the reward is expected, and so induces no prediction error, and so no non-baseline dopamine activity. However, the stimulus is not expected. It provides a signal that reward will be delivered in the future, and so is associated with a large and positive prediction error, and so a large and positive deflection above baseline of the dopamine activity. The way that this model accounts for secondary conditioning is clear. If another stimulus (say a tone) is presented before the light, then the activity of the dopamine system consequent on the light (seen in ‘after’) acts for the tone just as the activity consequent on the reward (seen in ‘before’) originally acted for the light. Thus, learning about the relationship between tone and reward will proceed. In the figure, the signal Δv is the difference in successive predictions of the reward. This is 0 whenever the predictions are not changing, *ie* before the cue, and between the cue and reward.

One facet of the temporal difference signal is shown by the plot of Δv in figure 8 after learning is complete. If the monkey is not provided with a reward when it is expecting it, then the temporal difference signal $\delta = \Delta v$ will be exactly this. The below-baseline activity around the time of the reward is a mark of the reward prediction. The rightmost plot in figure 9 shows dopaminergic activity in the same circumstances. The well-timed, below-baseline, activity is clear, just as in the model.

Figure 8 embodies a crucial assumption that the monkey has a way of keeping time precisely between the stimulus and the reward. Each time step is individuated by a different stimulus (called a serial compound conditioned stimulus, Gormezano & Kehoe, 1989 or a spectral representation of time, Grossberg & Schmajuk, 1989) in just the same way that each location in the maze might be identified by a single stimulus. Since figure 9 shows that the inhibition is well timed in the case that the reward is not provided, we know that this time must be represented somehow; however the precise neural substrate is unclear. O’Reilly, Braver & Cohen (1999); Braver, Barch & Cohen (1999) have suggested that one result of the dopamine response to the predictive stimulus is to gate the stimulus into working memory in prefrontal cortex. This would mean that information that the stimulus had recently been presented is available for prediction and control, and the complex developing pattern of population activity of neurons in the

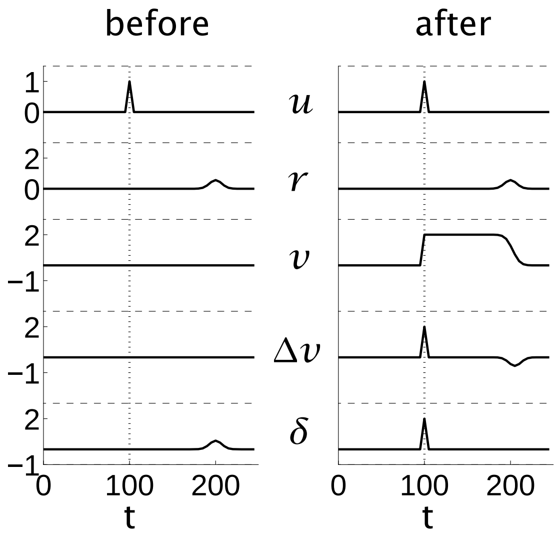

Figure 8: Model of the data of figure 7. The series of graphs show the cue (u), the activity consequent on the reward (r), the learned prediction (v), the temporal difference in the learned prediction (Δv), and the temporal difference error signal ($\delta = r + \Delta v$). Left plots show 'before' learning; right plots show 'after' learning. δ should be matched to the data in figure 7. Adapted from Montague, Dayan & Sejnowski (1996); Dayan & Abbott (2001).

prefrontal working memory may individuate the time steps between the stimulus and reward, as in proposals for how timing information is represented in the cerebellum, Buonomano & Mauk, 1994; Medina & Mauk, 2000). Suggestions for the involvement of the hippocampus (Grossberg & Merrill, 1996) and intracellular processes in striatal cells (Brown, Bullock & Grossberg, 1999) have also been made. In fact, the cerebellum has been suggested as another neural substrate for conditioning, particularly eyeblink conditioning, and Rescorla-Wagner (Gluck, Reifsnider & Thompson, 1990) and temporal difference (Moore, Berthier & Blazis, 1990) models of its involvement have been built. These lie outside the scope of this chapter. The subtleties of the timing behavior of animals in conditioning experiments are discussed in other chapters in this book (Church, this volume). The substantial timing noise that is evident in many experiments and is the subject of substantial theory itself (*eg* Gibbon, 1997) may have a significant impact on the pattern of activity of the dopamine cells.

The original suggestion for a temporal difference account of conditioning (Sutton & Barto, 1990) did not employ a serial compound conditioned stimulus, but instead relied on the earlier psychological notion of stimulus traces (Hull, 1943). The idea is that if u_i representing the cue is active at one time ($t = 100$ in figure 8), then the associated weight w_i is *eligible* to be changed according to temporal difference prediction errors δ that occur at later times ($t = 200$ in the figure, around the time of the reward). This allows the model to account for phenomena such as secondary conditioning. However,

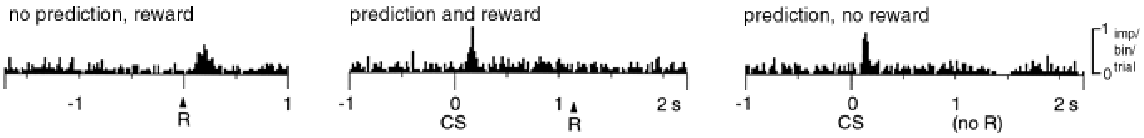

Figure 9: Activity of single dopamine cells to reward that is unexpected (left), predicted and delivered (center), or predicted and not delivered. The left and center plots match to figures 7 and the lowest lines in figure 8, the rightmost plot matches Δv 'after' learning in figure 8. The scale is in spikes per bin per trial. Adapted from Schultz, Dayan & Montague, 1997.

short of a way of keeping time between the stimulus and the reward, it does not make correct predictions about the activity of the dopamine cells. Stimulus traces actually play an interesting computational role in the temporal difference model, as explored by Watkins (1989).

As mentioned, there is an active debate on the full link between dopamine and appetitive conditioning. For instance, we have focused on the phasic behavior of dopamine cells in response to surprising rewards; there is also evidence from dialysis that dopamine is released, and from neurophysiology that the dopamine cells are activated in a more persistent manner, by aversive contingencies (Herman *et al,* 1982; Claustre *et al*, 1986; Abercrombie *et al,* 1989; Guarraci & Kapp, 1999). They are also activated by novel stimuli (Ljungberg, Apicella & Schultz, 1992; Horvitz, Stewart & Jacobs, 1996; although note that novelty can sometimes act as if it is itself rewarding, Reed, Mitchell & Nokes, 1996) and by stimuli that resemble other stimuli that are associated with rewards (Schultz, 1998). Although extensions of the temporal difference model have been suggested to account for all these behaviors (*eg* Contreras-Vidal & Schultz, 1999; Daw & Touretzky, 2000; Kakade & Dayan, 2001), more experiments and more theory are still required. These, and other findings have also led to suggestions about the role of dopamine other than prediction and reward learning, particularly focusing on orienting and the allocation of attention (Ward & Brown, 1996; Han *et al*, 1997; Yamaguchi & Kobayashi, 1998; Redgrave, Prescott & Gurney, 1999). It is worth noting that it is not yet completely clear how activity of the dopamine cells translates into the release of dopamine at target sites.

It is also not clear exactly what is responsible for the activity of the dopamine cells, or which synapses change their values to reflect the predictions of reward associated with stimuli and to reflect the appropriate choice of actions. There is some evidence in both rats and monkeys that the basolateral nucleus of the amygdala and the orbitofrontal cortex play an important role in storing the values associated with stimuli (Hatfield *et al*, 1996; Schoenbaum, Chiba & Gallagher, 1998; 1999; Gallagher, McMahan & Schoenbaum, 1999; Robbins & Everitt, 1999; Rolls, 2000). However, exactly how this information translates into the activity of dopamine cells is open to speculation. Another critical piece for the model is the way that dopamine can control the selection, and particularly the learned selection of actions. Early suggestions that this is a role for plasticity into the cortico-striatal afferents to the dorsal striatum are now under some question (see Houk, Davis & Beiser, 1995; Kropotov & Etlinger, 1999).

5 Statistical Models of Attention

The temporal difference model for learning values inherits a number of properties from the Rescorla-Wagner rule as to how different possible cues interact to make predictions about expected future rewards and to have their predictive weights change. As in equation 14, the predictions made by all the stimuli that are present are just summed, and, as in equation 15, the weights of all stimuli present are changed by the same amount. The only competition between stimuli is as in blocking, *ie* if one stimulus already predicts an outcome perfectly, then other stimuli that are simultaneously present will attract no learning, because there is no prediction error.

The temporal difference model thus also inherits the problems that these assumptions make for the Rescorla-Wagner rule. What both leave out is *attention.* Attention is a complicated and multi-faceted concept (see, for example, Parasuraman, 1998), with many different implications for different neural processes. In conditioning, there is a long history (see Dickinson, 1980; Mackintosh, 1983 for reviews) of the study of selective attention, originally sparked by the idea that there might be a limited capacity learning processor responsible for relating stimuli to affective outcomes, and that stimuli might compete to gain access to this processor. Even though the idea of limited capacity explanations for attention now has rather less currency (Allport, 1993), substantial experimental results on selective attention in conditioning remain to be explained.

The Rescorla-Wagner and temporal difference models do not take account of the possibility that, based on the past history of interaction, animals might accord more weight to the predictions made by some stimuli than others, or that some stimuli might attract faster learning than others. Alternative models of conditioning, for which there is quite some evidence, place substantial emphasis on these stronger forms of competition between stimuli. One of the best studied alternatives is that due to Pearce & Hall (1980), who suggested that animals should learn more about stimuli whose consequences are more uncertain. In fact, in its original form, the Pearce-Hall model explains even basic blocking by this mechanism, rather than by the lack of prediction error. Pearce & Hall (1980) also point out that stimuli about which an animal is uncertain are exactly the ones that should not be believed in making predictions about rewards. However, they did not suggest a quantitative account of how their predictions could be ignored. Grossberg (see 1982) suggests that all the stimuli that are present should compete to make predictions, although this competition is not, at least in a straightforward manner, based on uncertainty.

Holland and his colleagues (see Holland, 1997; Holland & Gallagher, 1999) have performed a wealth of tests of the neural basis of the Pearce-Hall model in appetitive conditioning. Primarily using selective lesions, they have identified a pathway in rats from the central nucleus of the amygdala, through the cholinergic basal forebrain to the parietal cortex that is critically involved in the faster learning accorded to stimuli whose outcomes are made uncertain. They have also identified a different pathway involved in the slower learning accorded to stimuli whose outcomes are perfectly predictable (as in latent inhibition, Lubow, 1989; see Baxter, Holland & Gallagher, 1997). Baxter *et al,* (1999) suggest that eliminating both these pathways leaves rats with an underlying learning behavior more consistent with the Rescorla-Wagner rule, confirming that a simple phenomenon like blocking is likely to be multiply determined. As an aside, the involvement of cholinergic systems in processing uncertainty is striking because of evidence of the action of ACh in changing the balance in favor of stimulus-driven or bottom-up input against recurrent or top-down input in determining the activity of cells in both the hippocampus and the cortex (Hasselmo, Anderson & Bower, 1992; Hasselmo, 1999). In cases in which the animal knows itself to be uncertain, recurrent and top-down input is less likely to be correct.

From a computational point of view, these two rather different forms of attention should emerge naturally from statistical considerations. The models of this, such as Kruschke (1997; 2000); Gallistel & Gibbon (2000); Kakade & Dayan (2000); Dayan, Kakade & Montague (2000) are not yet completely integrated into the full reinforcement learning models for tasks such as the maze, so we will just describe them briefly.

Competitive Combination

One way to generate competition between cues for making predictions about outcomes is to think of combining the views of multiple experts about the same event. In this circumstance, it is natural to weight each expert's view according to how reliable a predictor that expert has been in the past. Clearly, adding together the predictions made by all the experts, which is what equation 14 suggests, would be rather strange. There are various ways to formalize competitive combination (Jacobs, 1995) from a statistical viewpoint. Understanding how these models fit conditioning data provides insight into the statistical assumptions embodied by animal learners.

Two statistical competitive models have been advocated for conditioning. Kruschke (1997; 2000) suggested using the standard mixture of experts architecture (Jacobs, Jordan, Nowlan & Hinton, 1991), motivated partly by Mackintosh's (1975) attentional model, and Dayan & Long (1997); Kakade & Dayan (2000) suggested the original mixture of experts model (Jacobs, Jordan & Barto, 1991) motivated by the experimental phenomenon of downwards unblocking (Mackintosh, Bygrave & Picton, 1977; Holland, 1984; 1988) and results on the apparent sloth of learning in autoshaping (Gallistel & Gibbon, 2000). Although there are substantial differences in the assumptions and behavior of these models, they share the characteristic of competitive combination.

Figure 10 shows an example of combination according to the Jacobs, Jordan & Barto (1991) model. Here, competition is derived from unreliability. Two cues, F and G (say a light and a tone) both make unreliable predictions about the reward. The unreliability

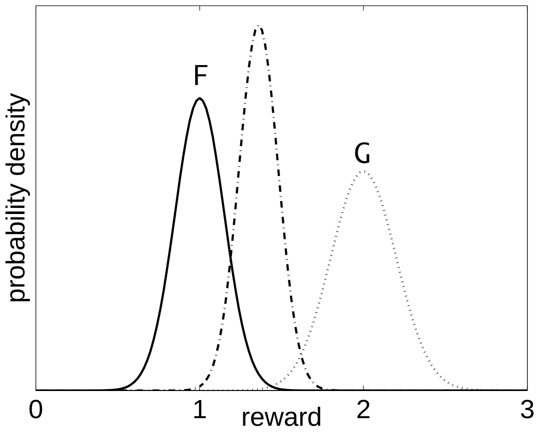

Figure 10: Unreliability model. Solid and dotted lines show the distributions of the predictions of reward made by two cues (F and G). These are Gaussian distributions with precisions (inverse variances) reflecting the reliabilities of the cues. The dot-dashed line shows the combined prediction, assuming that the individual predictions are statistically independent in an appropriate way. After Dayan, Kakade & Montague (2000).

is shown on the plot in the terms of the whole distribution over the actual reward given each cue by itself. Cue F (the solid line) predicts a single pellet of reward; cue G (the dotted line) two pellets. However, F is more reliable than G, as is evident from the fact that its distribution is sharper. That is, under cue F the reward is tightly constrained near 1 pellet; whereas under cue G, the reward is more loosely constrained near 2 pellets. Given both cues, how should a net prediction be constructed? Under a model that the unreliabilities of the cues are statistically independent, it turns out that these distributions should be multiplied together, and then reweighted. The net prediction (the dot-dashed line) is therefore of an intermediate number of pellets, nearer 1 than 2 (since F which predicts 1 pellet, is more reliable than G). The net prediction is more reliable than those of either F or G, since it integrates information from both. By contrast, the standard additive combination model for temporal difference learning or the Rescorla-Wagner rule, would predict 3 pellets of reward.

Purely additive models such as equation 14, and purely competitive models such as figure 10 both fail to account for some conditioning data. Competitive models have difficulty with circumstances such as overexpectation (see Lattal & Nakajima, 1998; Rescorla, 1999); additive models fail for phenomena such as downwards unblocking (see Dickinson, 1980). Providing a statistical basis for integrating these models is an important next step.

Competitive Adaptation

The opposite of a reliable predictor is a stimulus about whose consequences the animal is in substantial doubt. According to the Pearce-Hall model, this uncertainty should lead the animal to learn faster about the associated stimulus. In the context of the temporal difference model of equation 15, this can be implemented simply by providing each stimulus unit with its own learning rate η_i. However, what is a computational account of how these learning rates should be set? Once again, a psychological intuition has a computational basis, in this case in ideas about learning and adaptation in the face of uncertainty and change in the world. This basis was provided by Sutton (1992), who formalized an alternative way of thinking about the course of conditioning in terms of an engineering and statistical device called a Kalman filter (Anderson & Moore, 1979).

Sutton suggested that the conditioning behavior of animals is consistent with their trying to 'reverse-engineer' the relationship between stimuli and rewards that is established by the experimenter (or, in more natural circumstances, by nature). The process of reverse-engineering consists of combining together the information about the relationship that is provided by the sample stimuli and rewards provided on each trial. Initially, the animal will be quite uncertain about the relationship; given lots of trials, even noisy ones, it may become more certain. However, if the relationship continually changes (as it almost always does in both natural and experimental contexts), then very old trials may no longer be relevant to the current relationship, and therefore their information should be discarded. The Kalman filter is a statistically precise way of formulating the answer to this problem; Sutton (1990); Dayan, Kakade & Montague (2000) suggested an approximation to it as a way of modeling the phenomena that motivate the Pearce-Hall rule.

The approximate Kalman filter takes the prediction error δ of equation 12, and distributes it among all the stimuli that are present in a trial, in proportion to their uncertainties. The more certain the animal is about the predictions based on a stimulus, *ie* the more past trials have been used to establish this prediction, the less responsibility that stimulus takes for any new prediction error. Conversely, stimuli about which the animal has learned little, and about which it is therefore very uncertain, have their weights changed substantially more.

Figure 11 shows a simple version of this model at work in a case in which three different stimuli (labelled u_1, u_2 and u_3) are provided in partial combination with a number of pellets of reward over an extended period. Stimulus u_3 is only randomly related to the reward, and so attracts no learning; stimuli u_1 and u_2 are closely related to the reward, but in different ways at different times during the experiment. Figure 11B shows the predictive weights w_1 (solid) and w_2 (dashed) for u_1 and u_2 respectively. The learning of these weights is based on the uncertainties shown in figure 11C. Initially, the animal is highly uncertain about the associations of all the stimuli, and so when u_1 is presented in close combination with one pellet of reward, learning of w_1 is swift. The animal then becomes more certain about the relationship, so at t_*, when it changes, the anmal is rather slower to change its prediction. At time t_S, when u_2 is introduced

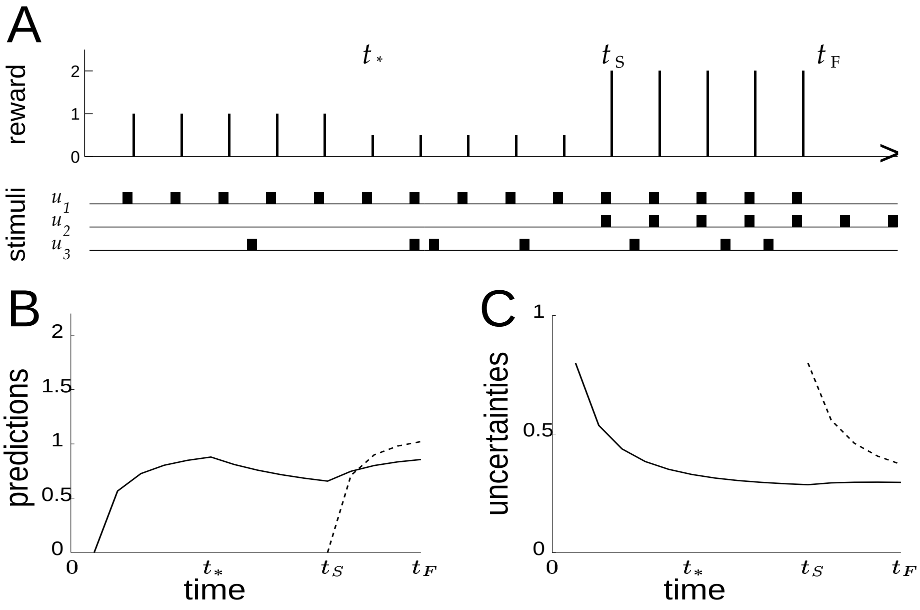

Figure 11: Uncertainty model. A) Conditioning experiment involving three stimuli u_1, u_2 and u_3, whose presence on a trial is indicated by a small black square, and a variable number of pellets of reward. t_*, t_S and t_F are particular times at which the stimulus or reward contingencies change. B) net predictions for u_1 (solid) and u_2 (dashed) over the course of the experiment. C) uncertainties associated with the same stimuli. The predictions change quickly (*ie* the weights change quickly) for stimuli whose predictions are highly uncertain. After Dayan, Kakade & Montague (2000).

in combination with a change in the number of pellets, the greater uncertainty of u_2 means that it learns much faster than u_1, and so establishes itself as a major predictor. This is exactly as expected from the Pearce-Hall model.

This model for the competitive allocation of learning is also rather preliminary, and has yet to be properly integrated with the model of competitive combination. In particular, it does not capture Pearce-Hall's observation that the uncertainties change adaptively in response to prediction errors, both in predictions of reward and also in predictions of other stimuli. For instance, the neural basis of attentional allocation has been studied using a task developed by Wilson, Boumphrey & Pearce (1992), in which uncertainties in stimulus-stimulus predictions are explicitly manipulated. In this task, the predictive consequences of one conditioned stimulus for another, change from one set of trials to the next. This change, which leads to an inevitable prediction error, makes for faster learning of the association between the first conditioned stimulus and a reward. Our simple account of the competitive allocation of learning cannot model this more sophisticated experiment.

In general, reinforcement learning methods can be used to learn stimulus-stimulus predictions as well as stimulus-outcome predictions, (Sutton & Pinette, 1985; Dayan, 1993; Suri & Schultz, 2001), and this capacity has even been used to explain phenomena such as sensory preconditioning. However, the range of properties of such models has yet to be fully explored. Indeed, there is debate in both the psychological literature (Gallistel, 1990; Gallistel & Gibbon, 2000) and the theoretical reinforcement learning literature (Kearns & Singh, 1999) about the relative merits of model-free methods such as the actor-critic or Q-learning and model-based methods that build and use models of the stimulus-stimulus and stimulus-reward contingencies of environments.

6 Discussion

Two areas of current work in reinforcement learning have a particular importance for psychological and neural modeling: eliminating the rather artificial definition of a trial, and performing planning over larger spatial and temporal horizons.

First, as has been forcefully pointed out by Gallistel & Gibbon (2000) and others, the definition of a trial in tasks such as the maze is problematical. Even in the maze case, animals might integrate reward information over multiple runs, if they are continually taken back to the start point when they reach one of the end boxes; further, many standard instrumental conditioning tasks involve extended series of repeated choices. Under such circumstances, the criterion of maximizing the sum total reward over a whole session is unreasonable, for instance because this sum may get very large. In dynamic programming, two ways of handling such cases have been suggested, and both of them have been translated into reinforcement learning. The first is to consider summed *discounted* rewards. That is, rewards that are expected to be received in the future are down-weighted by an amount that reflects how long it takes them to arrive. The second is to consider the long run average reward rate rather than the sum total reward.

The standard model of discounting in dynamic programming is exponential, that is a reward of r pellets received t timesteps from now is treated as being worth only $\gamma^t r$, where $0 \leq \gamma < 1$ is the discount factor. If γ is near 0, then only very proximal rewards have a significant impact on the sum discounted reward. Exponential discounting is standard in economic contexts over long times, in order to take account of inflation and interest rates. Exponential discounting is also easy to incorporate into the temporal difference model we described, for instance changing equation 12 to

$$\delta = r + \gamma v(\mathsf{Y}) - v(\mathsf{X})$$

However, in direct tests, humans and other animals do not appear to use exponential discounting. Rather, at least in many contexts, they discount rewards hyperbolically (Rachlin, 1970; 1974; Ainslie, 1975; 1992; Mazur, 1984), so the r pellets are worth $r/(\alpha + \beta t)$, where α and β are parameters. Hyperbolic discounting has a number of unexpected properties that have been experimentally confirmed. One example is preference reversals: offered a choice between \$10 now and \$20 next week, subjects might prefer to get \$10 now; whereas offered a choice between \$10 in six weeks and \$20 in seven weeks, subjects tend to be content to wait for the \$20. Exponential discounting cannot generate preference reversals like this. Hyperbolic discounting is much harder to accommodate within a temporal difference framework, since there is no equivalent to equation 8's recursive definition of value.

The alternative to the discounted model for dynamic programming is the average-case model. In this case, the goal is to maximize the reward rate over time rather than the cumulative reward. Kacelnik (1997) has argued that apparently hyperbolic discounting might really result from the average-case model, at least given the way that the preference experiments are run. There is an extensive body of dynamic programming (see Puterman, 1994) and reinforcement learning (Schwartz, 1993; Mahadevan, 1996; Tsitsiklis & Van Roy, 1999) theory on the average case, which turns out to require only a small modification to the temporal difference model we have discussed. The idea is to learn the overall average reward rate ρ, using a learning rule like the Rescorla-Wagner rule of equation 6, and then to modify the temporal difference prediction error of equation 12 by subtracting ρ from the expression there

$$\delta = r + v(\mathsf{Y}) - v(\mathsf{X}) - \rho \tag{21}$$

Daw & Touretzky (2000) suggested exactly this, and showed that it makes essentially the same predictions as the standard temporal difference model for the experiments that have been use to probe the dopamine system.

Equation 21 has an interesting interpretation as the difference between a phasic signal (the original δ of equation 12) and a tonic signal (representing ρ). One possibility is that these signals are actually represented by different neuromodulators – in the appetitive case, perhaps dopamine for the phasic signal and serotonin for the tonic signal. Concomitantly, a tonic dopamine signal might represent the equivalent of ρ in the aversive case, *ie* the mean punishment rate. This has been suggested as an interpretation of the data that dopamine is released in cases of aversive conditioning. This way of looking at equation 21 is a special case of an *opponent* interaction between appetitive and aversive systems, a long-standing psychological idea, which has been the subject of influential modeling by Solomon & Corbit (1974) and Grossberg (1984); Grossberg & Schmajuk (1987). Understanding the interactions between the various neuromodulators is a key current concern (Doya, 2000).

Another main area of investigation in reinforcement learning is the representation and manipulation of hierarchical structure. The idea is that many tasks are not best solved by considering only the simplest and shortest actions at every time (like a single step in the maze), but rather that planning can better proceed with longer spatial and temporal horizons. The expanded horizons might come from considering sets of actions, going by the names of 'chunks', 'options', 'macros' or 'subroutines' (see Sutton, Precup & Singh, 1999). These notions offer the prospect of going full circle back to some of the concerns that were central in the original study of sequential chains of behavior, such as the effects of multiple goals in an environment (stemming, for instance, from a sometimes hungry, sometimes thirsty, rat in a maze with both food and water rewards).

They also offer mechanisms for integrating model-free and model-based methods of learning and control. In fact, there are even suggestions (Frank, Loughry & O'Reilly, 2000) as to how the interaction between the prefrontal cortex and the basal ganglia might implement such subroutines.

There is substantial theory about how to use hierarchical structure such as options within a reinforcement learning context, and indeed proofs that appropriate structure can make for much faster learning. However, there are rather fewer proposals as to how such structure might be induced, as tasks are being solved. One idea is that unsupervised learning methods which are models for cortical plasticity (see Becker & Plumbley, 1996; Hinton & Sejnowski, 1999) and are designed to extract statistical structure from inputs, might be involved. Understanding this requires modeling the interactions between unsupervised and reinforcement learning, and between cortical and sub-cortical plasticity (Doya, 1999).

In conclusion, we have described the standard model of reinforcement learning for single and sequential action choice, and its main relationships with classical and instrumental conditioning, dynamic programming and the dopamine system. Reinforcement learning is thus one of the few areas in which constraints and ideas from many levels of computational analysis (computational, algorithmic and implementational, Marr, 1982) and many sorts of experimental data (ethological, psychological and neurobiological) can collectively be brought to bear.

Acknowledgements

Funding came from the Gatsby Charitable Foundation. I am most grateful to Randy Gallistel for his trenchant comments on an earlier draft.

References

Abercrombie, ED, Keefe, KA, DiFrischia, DS & Zigmond, MJ (1989) Differential effect of stress on in vivo dopamine release in striatum, nucleus accumbens, and medial frontal cortex. *Journal of Neurochemistry* **52**1655-1658.

Ainslie, G (1975) Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin* **82**:463-496.

Ainslie, G (1992) *Picoeconomics.* Cambridge: Cambridge University Press.

Allport, A (1993). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In DE Meyer & S Kornblum, editors, *Attention and Performance 14.* Cambridge, MA: MIT Press, 183-218.

Anderson, BDO & Moore, JB (1979) *Optimal Filtering.* Englewood Cliffs, NJ: Prentice-Hall.

Barto, AG, Sutton, RS & Anderson, CW (1983) Neuronlike elements that can solve difficult learning problems. *IEEE Transactions on Systems, Man, and Cybernetics* **13**:834-846.

Baxter, MG, Bucci, DJ, Holland, PC & Gallagher, M (1999) Impairments in conditioned stimulus processing and conditioned responding after combined selective removal of hippocampal and neocortical cholinergic input. *Behavioral Neuroscience* **113**:486-495.

Baxter, MG, Holland, PC & Gallagher, M (1997) Disruption of decrements in conditioned stimulus processing by selective removal of hippocampal cholinergic input. *Journal of Neuroscience* **17**:5230-5236.

Becker, S & Plumbley, M (1996) Unsupervised neural network learning procedures for feature extraction and classification. *International Journal of Applied Intelligence* **6**:185-203.

Bellman, RE (1957) *Dynamic Programming* Princeton, NJ: Princeton University Press.

Beninger, RJ (1983) The role of dopamine in locomotor activity and learning. *Brain Research Reviews* **6**:173-196.

Berridge, KC & Robinson, TE (1998) What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* **28**:309-369.

Berry, DA & Fristedt, B (1985) *Bandit Problems: Sequential Allocation of Experiments.* London, England: Chapman and Hall.

Bertsekas, DP & Tsitsiklis, JN (1996) *Neuro-Dynamic Programming.* Belmont, MA:Athena Scientific.

Braver, TS, Barch, DM & Cohen, JD (1999) Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry* **46**:312-328.

Brown, J, Bullock, D & Grossberg, S (1999) How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience* **19**:10502-10511.

Buonomano, DV & Mauk, M (1994) Neural network model of the cerebellum: Temporal discrimination and the timing of motor responses. *Neural Computation* **6**:38-55.

Bush, RR & Mosteller, F (1955) *Stochastic Models for Learning.* New York:Wiley.

Claustre, Y, Rivy, JP, Dennis, T & Scatton, B (1986) Pharmacological studies on stress-induced increase in frontal cortical dopamine metabolism in the rat. *Journal of Pharmacology and Experimental Therapeutics* **238**:693-700.

Contreras-Vidal, JL & Schultz, W (1999) A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience* **6**:191-214.

Daw, ND & Touretzky, DS (2000) Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing: An International Journal* **32**:679-684.

Dayan, P (1993). Improving generalisation for temporal difference learning: The successor representation. *Neural Computation,* **5**, 613-624.

Dayan, P & Abbott, LF (2001). *Theoretical Neuroscience.* Cambridge, MA: MIT Press.

Dayan, P, Kakade, S & Montague, PR (2000). Learning and selective attention. *Nature Neuroscience,* **3**, 1218-1223.

Dayan, P., & Long, T. (1997). Statistical Models of Conditioning. *Neural Information Processing Systems, 10*, 117-124.

Deutsch, JA (1960) *The Structural Basis of Behavior.* Cambridge, England: Cambridge University Press.

Dickinson, A (1980) *Contemporary Animal Learning Theory.* Cambridge: Cambridge University Press.

Doya, K (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks* **12**:961-974.

Doya, K (2000) Meta-learning, neuromodulation and emotion. In G Hatano, editor, *Proceedings of the 13th Toyota Conference on Affective Minds.* Holland: Elsevier Science.

Egelman, DM, Person, C, Montague, PR (1998) A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience* **10**:623-630.

Frank, MJ, Loughry, B & O'Reilly, RC (2000) *Interactions Between Frontal Cortex and Basal Ganglia in Working Memory: A Computational Model.* ICS Technical Report 00-01, Department of Psychology, University of Colorado at Boulder.

Friston, KJ, Tononi, G, Reeke, GN Jr, Sporns, O & Edelman, GM (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* **59**229-243.

Gallagher, M, McMahan, RW & Schoenbaum, G (1999) Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience* **19**:6610-6614.

Gallistel, CR & Gibbon, J (2000) Time, rate, and conditioning. *Psychological Review* **107**:289-344.

Gibbon, J (1977) Scalar expectancy theory and Weber's Law in animal timing. *Psychological Review* **84**:279-325.

Gluck, MA, Reifsnider, ES & Thompson, RF (1990) Adaptive signal processing and the cerebellum: Models of classical conditioning and VOR adaptation. In MA Gluck & DE Rumelhart, editors, *Neuroscience and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum Associates, 131-185.

Gormezano, I & Kehoe, EJ (1989) Classical conditioning with serial compound stimuli. In JB Sidowski, editor, *Conditioning, Cognition, and Methodology: Contemporary Issues in Experimental Psychology.* Lanham, MD: University Press of America, 31-61.

Grossberg, S (1982) Processing of expected and unexpected events during conditioning and attention: a psychophysiological theory. *Psychological Review* **89**:529-572.

Grossberg, S (1984) Some normal and abnormal behavioral syndromes due to transmitter gating of opponent processes. *Biological Psychiatry* **19**:1075-1118.

Grossberg, S & Merrill, JWL (1996) The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience* **8**:257-277.

Grossberg, S & Schmajuk, NA (1987) Neural dynamics of attentionally modulated Pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology* **15**:195-240.

Grossberg, S & Schmajuk, NA (1989) Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks* **2**:79-102.

Guarraci, FA & Kapp, BS (1999) An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential Pavlovian fear conditioning in the awake rabbit. *Behavioural Brain Research* **99**:169-179.

Han, J-S, McMahan, RW, Holland, PC & Gallagher, M (1997) The role of an amygdalonigrostriatal pathway in associative learning. *Journal of Neuroscience* **17**:3913-3919.

Hasselmo, ME, Anderson, BP & Bower, JM (1992) Cholinergic modulation of cortical associative memory function. *Journal of Neurophysiology* **67**:1230-1246.

Hasselmo, ME (1999) Neuromodulation: acetylcholine and memory consolidation. *Trends in Cognitive Science* **3**:351-359.

Hatfield, T, Han, J-S, Conley, M, Gallagher, M & Holland, PC (1996) Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. *Journal of Neuroscience* **16**:5256-5265.

Herman, JP, Guillonneau, D, Dantzer, R, Scatton, B, Semerdjian-Rouquier L, le Moal M (1982) Differential effects of inescapable footshocks and of stimuli previously paired with inescapable footshocks on dopamine turnover in cortical and limbic areas of the rat. *Life Sciences* **30**:2207-2214.

Hinton, GE & Sejnowski, TJ, editors (1999) *Unsupervised Learning: Foundations of Neural Computation.* Cambridge, MA: MIT Press.

Holland, PC (1984) Unblocking in Pavlovian appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes* **10**:476-497.

Holland, PC (1988) Excitation and inhibition in unblocking. *Journal of Experimental Psychology: Animal Behavior Processes* **14**:261-279.

Holland, PC (1997) Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. *Animal Learning & Behavior* **25**:373-399.

Holland, PC & Gallagher, M (1999) Amygdala circuitry in attentional and representational processes. *Trends In Cognitive Sciences* **3**:65-73.

Horvitz, JC, Stewart, T & Jacobs, BL (1997) Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research* **759**:251-258.

Houk, JC, Adams, JL& Barto, AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement.In JC Houk, JL Davis & DG Beiser, editors, *Models of Information Processing in the Basal Ganglia.* Cambridge, MA: MIT Press, 249-270.

Houk, JC, Davies, JL & Beiser, DG, editors (1995) *Models of Information Processing in the Basal Ganglia.* Cambridge, MA:MIT Press.

Hull, CL (1943) *Principles of Behavior.* New York, NY: Appleton-Century.

Hull, CL (1952). *A Behavior System.* New Haven, CT: Yale University Press.

Ikemoto, S & Panksepp, J (1999) The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews* **31**:6-41.

Jacobs, RA (1995) Methods for combining experts' probability assessments. *Neural Computation* **7**:867-888.

Jacobs, RA, Jordan, MI & Barto, AG (1991) Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. *Cognitive Science* **15**:219-250.

Jacobs, RA, Jordan, MI, Nowlan, SJ & Hinton, GE (1991) Adaptive mixtures of local experts. *Neural Computation* **3**:79-87.

Kacelnik, A (1997) Normative and descriptive models of decision making: Time discounting and risk sensitivity. In *Characterizing Human Psychological Adaptations.* Ciba Foundation. Chichester, England: John Wiley & Sons, 51-70.

Kakade, S & Dayan, P (2000). Acquisition in autoshaping. In SA Solla, TK Leen & K-R Muller, editors, *Advances in Neural Information Processing Systems, 12,* 24-30.

Kakade, S & Dayan, P (2001). Dopamine bonuses. In TK Leen, TG Dietterich & V Tresp, editors, *Advances in Neural Information Processing Systems, 13.* Cambridge, MA: MIT Press.

Kearns, M & Singh S. Finite-sample convergence rates for Q-learning and indirect algorithms. In In MS Kearns, SA Solla & DA Cohn, editors, *Advances in Neural Information Processing Systems, 11.* Cambridge, MA: MIT Press.

Koob, GF (1992) Drugs of abuse: anatomy, pharmacology and function of reward pathways. *Trends in Pharmacological Sciences* **13**:177-184.

Krebs, JR, Kacelnik, A & Taylor, P (1978) Test of optimal sampling by foraging great tits. *Nature* **275**:27-31.

Kropotov, JD & Etlinger, SC (1999) Selection of actions in the basal ganglia-thalamocortical circuits: Review and model. *International Journal of Psychophysiology* **31**:197-217.

Kruschke, JK (1997). Relating Mackintosh's (1975) theory to connectionist models and human categorization. Talk presented at the *Eighth Australasian Mathematical Psychology Conference.* Perth Australia.

Kruschke, JK (2001) Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology* In press.

Lattal, KM & Nakajima, S (1998) Overexpectation in appetitive Pavlovian and instrumental conditioning. *Animal Learning & Behavior* **26**:351-360.

Ljungberg, T, Apicella, P & Schultz, W (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology* **67**:145-163.

Lubow, RE (1989) *Latent Inhibition and Conditioned Attention Theory.* New York, NY: Cambridge University Press.

Mackintosh, NJ (1974) *The Psychology of Animal Learning.* London, England: Academic Press.

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

Mackintosh, NJ (1983) *Conditioning and Associative Learning.* Oxford:Oxford University Press.

Mackintosh, NJ, Bygrave, DJ & Picton, BM (1977) Locus of the effect of a surprising reinforcer in the attenuation of blocking. *Quarterly Journal of Experimental Psychology***29**:327-336.

Mahadevan, S (1996) Average reward reinforcement learning: foundations, algorithms, and empirical results. *Machine Learning* **22**:159-195.

Mangel, M & Clark, CW (1988) *Dynamic Modeling in Behavioral Ecology.* Princeton, NJ:Princeton University Press.

Marr, D (1982) *Vision.* New York:Freeman.

Mazur, JE (1984) Tests of an equivalence rule for fixed and variable reinforcer delays. *Journal of Experimental Psychology: Animal Behavior Processes* **10**:426-436.

Medina, JF & Mauk, MD (2000) Computer simulation of cerebellar information processing. *Nature Neuroscience* **3**:1205-1211.

Michie, D & Chambers, RA (1968) BOXES: An experiment in adaptive control. *Machine Intelligence* **2**:137-152.

Montague, PR, Dayan, P & Sejnowski, TK (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16**:1936-1947.

Moore, JW, Berthier, NE & Blazis, DEJ (1990) Classical eye-blink conditioning: Brain systems and implementation of a computational model. In M Gabriel & J Moore, editors, *Learning and Computational Neuroscience: Foundations of Adaptive Networks.* Cambridge, MA: MIT Press, 359-387.

Mowrer, OH (1956). Two-factor learning theory reconsidered, with special reference to secondary reinforcement and the concept of habit. *Psychological Review* **63**:114-128.

Narendra, KS & Thatachar, MAL (1989) *Learning Automata: An Introduction.* Englewood Cliffs, NJ:Prentice-Hall.

O'Reilly, RC, Braver, TS & Cohen, JD (1999) A biologically based computational model of working memory. In A Miyake & P Shah, editors, *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control.* New York, NY: Cambridge University Press, 375-411.

Parasuraman, R, editor (1998) *The Attentive Brain.* Cambridge, MA: MIT Press.

Pearce, JM & Hall, G (1980) A model for Pavlovian learning: Variation in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review* **87**:532-552.

Puterman, ML (1994) *Markov decision processes: discrete stochastic dynamic programming.* New York, NY: Wiley.

Rachlin, H (1970) *Introduction to Modern Behaviorism.* San Francisco, CA: WH Freeman.

Rachlin. H (1974) Self control. *Behaviorism* **2**:94-107.

Redgrave, P, Prescott, TJ & Gurney, K (1999) Is the short-latency dopamine response too short to signal reward error. *Trends in Neurosciences* **22**:146-151.

Reed, P, Mitchell, C & Nokes, T (1996) Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Animal Learning and Behavior* **24**:38-45.

Rescorla, RA (1999) Summation and overexpectation with qualitatively different outcomes. *Animal Learning & Behavior* **27**:50-62.

Rescorla, RA & Wagner, AR (1972) A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In AH Black & WF Prokasy, editors, *Classical Conditioning II: Current Research and Theory.* New York:Aleton-Century-Crofts, 64-69.

Robbins, RW & Everitt, BJ (1992) Functions of dopamine in the dorsal and ventral striatum. *Seminars in Neuroscience* **4**:119-127.

Robbins, TW & Everitt, BJ (1999) Interaction of the dopaminergic system with mechanisms of associative learning and cognition: Implications for drug abuse. *Psychological Science* **10**:199-202.

Rolls, ET (2000) Memory systems in the brain. *Annual Review of Psychology* **51**:599-630.

Samuel, AL (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**:211-229.

Schoenbaum, G, Chiba, AA & Gallagher, M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience* **1**:155-159.

Schoenbaum, G, Chiba, AA &; Gallagher, M (1999) Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *Journal of Neuroscience* **19**:1876-1884.

Schultz, W (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology* **80**:1-27.

Schultz, W, Dayan, P & Montague, PR (1997). A neural substrate of prediction and reward. *Science,* **275**, 1593-1599.

Schwartz, A (1993). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning.* San Mateo, CA: Morgan Kaufmann, 298-305.

Simon, H & le Moal, M (1998) Mesencephalic dopaminergic neurons: Role in the general economy of the brain. *Annals of the New York Academy of Sciences* **537**:235-253.

Solomon, RL & Corbit, JD (1974) An opponent-process theory of motivation. I. Temporal dynamics of affect. *Psychological Review* **81**:119-145.

Spanagel, R & Weiss, F (1999) The dopamine hypothesis of reward: Past and current status. *Trends in Neurosciences* **22**:521-527.

Suri, RE & Schultz, W (2001) *Internal Model Reproduces Anticipatory Neural Activity.* Submitted for publication.

Sutton, RS (1988) Learning to predict by the methods of temporal difference. *Machine Learning* **3**:9-44.

Sutton, R. (1992) Gain adaptation beats least squares? In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems,* 161-166.

Sutton, RS & Barto, AG (1981) Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review* **88**:135-170.

Sutton, RS & Barto, AG (1990) Time-derivative models of Pavlovian conditioning. In M

Gabriel & JW Moore, editors, *Learning and Computational Neuroscience.* Cambridge, MA:MIT Press.

Sutton, RS & Barto, AG (1998) *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Sutton, RS & Pinette, B (1985). The learning of world models by connectionist networks. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society,* Irvine, CA: Lawrence Erlbaum, 54-64.

Sutton, RS, Precup, D & Singh, S (1999) Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* **112**:181-211.

Thorndike, EL (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs* **2** (4, Whole Number 8).

Thorndike, EL (1911) *Animal Intelligence.* New York, NY: Macmillan.

Tsitsiklis, JN & Van Roy, B (1999) Average cost temporal-difference learning. *Automatica* **35**:1799-1808.

Ward, NM & Brown, VJ (1996) Covert orienting of attention in the rat and the role of striatal dopamine. *Journal of Neuroscience* **16**:3082-3088.

Watkins, CJCH (1989) *Learning from Delayed Rewards.* PhD Thesis, University of Cambridge, Cambridge, UK.

Werbos, PJ (1990) A menu of designs for reinforcement learning over time. In WT Miller, IIIrd & RS Sutton, editors, *Neural Networks for Control.* Cambridge, MA: MIT Press, 67-95.

Wickens, J (1993) *A Theory of the Striatum.* Oxford: Pergamon Press.

Cellular models of reinforcement. In JC Houk, JL Davis & ME Hoff, editors, *Models of Information Processing in the Basal Ganglia.* Cambridge, MA: MIT Press, 187-214.

Widrow, B & Hoff, ME (1960) Adaptive switching circuits. *WESCON Convention Report* **IV**:96-104.

Widrow, B & Stearns, SD (1985) *Adaptive Signal Processing.* Englewood Cliffs, NJ:Prentice-Hall.

Williams, RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* **8**:229-256.

Wilson, PN, Boumphrey, P & Pearce, JM (1992) Restoration of the orienting response to a light by a change in its predictive accuracy. *Quarterly Journal of Experimental Psychology* **44B**:17-36.

Wise, RA (1982) Neuroleptics and operant behavior: The anhedonia hypothesis. *Behavioral & Brain Sciences* **5**:39-87.

Wise, RA & Bozarth, MA (1984). Brain reward circuitry: Four circuit elements “wired” in apparent series. *Brain Research Bulletin* **12**:203-208.

Wise, RA & Rompre, P-P (1989) Brain dopamine and reward. *Annual Review of Psychology* **40**:191-225.

Yamaguchi, S & Kobayashi, S (1988) Contributions of the dopaminergic system to voluntary and automatic orienting of visuospatial attention. *Journal of Neuroscience* **18**:1869-1878.