2002 Special issue

Dopamine: generalization and bonuses

Sham Kakade[*], Peter Dayan

Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WCIN 3AR, UK

Abstract

In the temporal difference model of primate dopamine neurons, their phasic activity reports a prediction error for future reward. This model is supported by a wealth of experimental data. However, in certain circumstances, the activity of the dopamine cells seems anomalous under the model, as they respond in particular ways to stimuli that are not obviously related to predictions of reward. In this paper, we address two important sets of anomalies, those having to do with generalization and novelty. Generalization responses are treated as the natural consequence of partial information; novelty responses are treated by the suggestion that dopamine cells multiplex information about reward bonuses, including exploration bonuses and shaping bonuses. We interpret this additional role for dopamine in terms of the mechanistic attentional and psychomotor effects of dopamine, having the computational role of guiding exploration. © 2002 Published by Elsevier Science Ltd.

Keywords: Dopamine; Reinforcement learning; Exploration; Temporal difference; Generalization

1. Introduction

Much evidence, reviewed by Schultz (1998), suggests that dopamine (DA) cells in the primate midbrain play an important role in reward and action learning. Electrophysiological studies in both instrumental (Schultz, 1992, 1998) and classical (Waelti, Dickinson, & Schultz, 2001) conditioning tasks support a theory that DA cells signal a global prediction error for summed future reward in appetitive conditioning tasks (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997), in the form of a temporal difference (TD) prediction error term. One use of this term is training the predictions themselves, a standard interpretation for the preparatory aspects of classical conditioning; another is finding the actions that maximize reward, as in a two-factor learning theory for the interaction of classical and instrumental conditioning. Storage of the predictions involves at least the basolateral nuclei of the amygdala (Hatfield, Han, Conley, Gallagher, & Holland, 1996; Holland & Gallagher, 1999; Whitelaw, Markou, Robbins, & Everitt, 1996) and the orbitofrontal cortex (Gallagher, McMahan, & Schoenbaum, 1999; O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001; Rolls, 2000;

Schoenbaum, Chiba, & Gallagher, 1998, 1999; Schultz, Tremblay, & Hollerman, 2000; Tremblay & Schultz, 2000a, b). The neural substrate for the dopaminergic control over action is rather less clear (Dayan, 2002; Dickinson & Balleine, 2001; Houk, Adams, & Barto, 1995; Montague et al., 1996).

The computational role of dopamine in reward learning is controversial for various reasons (Gray, Young, & Joseph, 1997; Ikemoto & Panksepp, 1999; Redgrave, Prescott, & Gurney, 1999). First, stimuli that are not associated with reward prediction are known to activate the dopamine system in a non-trivial manner, including stimuli that are novel and salient, or that physically resemble other stimuli that do predict reward (Schultz, 1998). In both cases, an important aspect of the dopamine response is that it sometimes consists of a short-term increase above baseline followed by a short-term decrease below baseline. Second, dopamine release is associated with a set of motor effects, such as species- and stimulus-specific approach behaviors, that seem either irrelevant or detrimental to the delivery of reward. We call these motor effects mechanistic because of their apparent independence from prediction or action.

In this paper (see also Suri & Schultz, 1999; Suri, 2002), we study various of these apparently anomalous activations of dopamine cells. We interpret the short term increase and decrease in the light of generalization as an example of partial information—the response is exactly what would be

* Corresponding author. Tel.: +44-20-7679-1175; fax: +44-20-7679-1173.

E-mail addresses: sham@gatsby.ucl.ac.uk (S. Kakade), dayan@gatsby.ucl.ac.uk (P. Dayan).
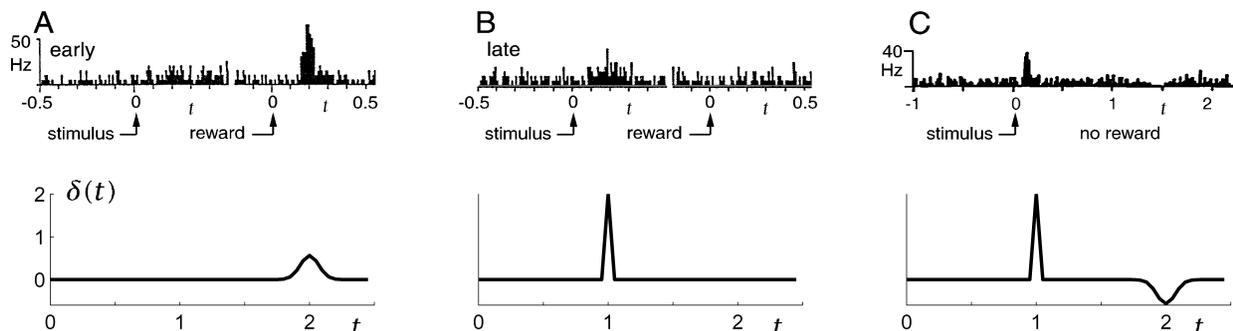
Fig. 1. Activity of DA cells (upper), and temporal difference prediction error model thereof (lower). The maximal firing rate is a function of the bin size used in the original data. (A) In early learning trials, a single dopamine cell responds to the delivery of the reward, but is barely excited by the delivery of the predictive stimulus. This is matched by the temporal difference prediction error $\delta(t)$, which follows the reward signal $r(t)$. (B) In later learning trials, a dopamine cell responds to the delivery of the stimulus, but not the reward. This is again matched by $\delta(t)$—there is no response at the time of the reward because the animal can predict the occurrence of the reward based on the stimulus. In the lower traces of (A) and (B), it is assumed that there is a fixed interval between stimulus and reward, so, unlike the upper traces, the model dopamine activity is not separately triggered on these two events. (C) In a different experiment, with otherwise similar activity, a dopamine cell is again activated by the predictive stimulus, but when the reward is unexpectedly *not* delivered, activity dips below baseline around the time the reward was expected. In this case, since $r(t) = 0$, the temporal difference error signal $\delta(t)$ follows $\Delta v = v(t + 1) - v(t)$, which has the same characteristic. (A) and (B) adapted from Mirenowicz and Schultz (1994); (C) adapted from Schultz, Dayan, and Montague (1997)—note the difference in the timescale of C. The absolute magnitudes of the firing rates in this plot and the other plots of neural activity are rendered a little arbitrary by the different resolutions with which time is binned in the different plots, but the patterns of responding are consistent.

expected where the animal to be initially incompletely certain as to whether or not the presented stimulus was the one associated with food. We interpret the short-term effects after new stimuli as suggesting that the DA system *multiplexes* information about bonuses on top of information about rewards. Bonuses are fictitious quantities added to rewards (Dayan & Sejnowski, 1996; Sutton, 1990) or values (Ng, Harada, & Russell, 1999) to ensure appropriate exploration in new or changing environments.

In Section 2, we describe the TD model of dopamine activity. In Section 3 we discuss generalization; in Section 4 we discuss novelty responses and bonuses.

2. Temporal difference and dopamine activity

Fig. 1 shows three aspects of the activity of dopamine cells, together with the associated TD model. The electrophysiological data in Fig. 1(A) and (B) are based on a set of reaction-time operant conditioning trials, in which monkeys are learning the relationship between an auditory conditioned stimulus (the CS) and the delivery of a juice reward (the unconditioned stimulus or US). The monkeys had to keep their hands on a resting key until the sound was played, and then they had to depress a lever in order to get juice. Fig. 1(A) shows the activity of a single dopamine cell in early trials, before the monkey has learned the relationship between the CS and the reward. At this point, dopamine cells fire substantially over baseline to the reward, but only weakly, if at all, to the CS. Once the relationship is established, the cells fire to the CS (Fig. 1(B)), but not to the (now expected) reward. Fig. 1(C) shows the result from a different experiment in which a monkey expected a reward on the basis of a CS, but the reward was

not delivered. In this case, activity of dopamine cells goes below baseline at the time that they were formerly activated by the presentation of reward. Note also the rather temporally precise response to the predictive stimulus. Although these data come from operant tasks, Waelti et al. (2001) showed that dopamine cells are activated similarly in Pavlovian circumstances. Thus, we consider how dopamine provides information about errors in prediction of reward, irrespective of actions. This information is also useful for the operant task to choose appropriate actions.

The pattern of neural responses in the upper plots of Fig. 1 is matched by the three lines in the lower part of the figure, which show the prediction error $\delta(t)$ in a TD model under equivalent circumstances. In the model, assuming that there is only one CS, predictions $v(t)$ are made at time t during a trial using information about the state of the trial (or more generally the state of the environment) at time t. A simple assumption is that

$$v(t) = \sum_{s \le t} w(t - s)u(s), \tag{1}$$

where $u(s) = 1$ if the conditioned stimulus is presented at time s in the trial, and 0 otherwise, and $w(t - s)$ is the weight associated with the stimulus having been presented $t - s$ timesteps before. In writing Eq. (1), we have made the tacit assumption that there is a different state of the trial for every timestep since the stimulus was presented. This assumption amounts to the use of a serial compound conditioned stimulus, i.e. a different stimulus for every timestep (Kehoe, 1977); more realistic models of interval timing (Church, 1984; Gibbon, Malapani, Dale, & Gallistel, 1997) have also been suggested (Grossberg & Schmajuk, 1989). If multiple stimuli are presented, so the state is defined by the simultaneous presence of multiple cues, then, as in Rescorla and Wagner (1972) rule (though not necessarily the data,
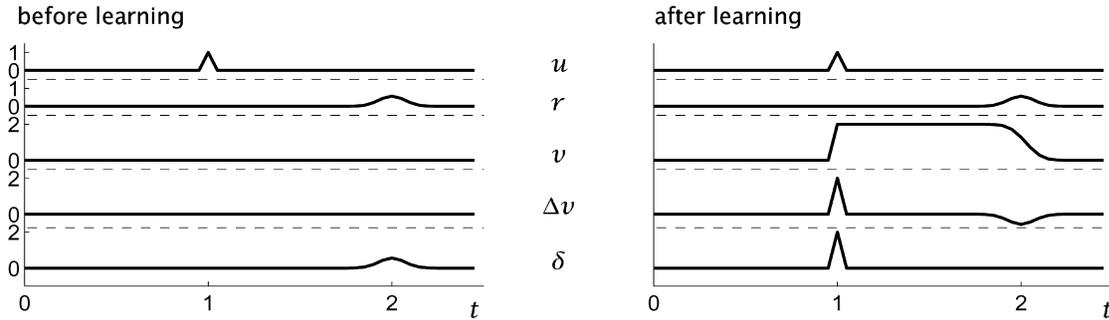
Fig. 2. Construction of the temporal difference prediction error signal δ in the model of Fig. 1. Left and right plots show the various key signals, including $u(t)$, marking the presentation of the stimulus, $r(t)$, the immediate reward, $v(t)$, the prediction of summed future reward, $\Delta v(t+1) = v(t+1) - v(t)$, the time difference of this prediction, and $\delta(t) = r(t) + \Delta v(t+1)$, the temporal difference error signal. Left plots show the traces before learning; right plots show the same signals after learning. The difference lies in the prediction signal $v(t)$, which, at the end of learning, rises to matches the integral of $r(t)$ when the stimulus is presented, only then to decline as the reward is being provided. Its temporal difference $\Delta v(t+1)$ exactly negates the activity $r(t)$ associated with the reward at the time of the reward. The lower plots are aligned with $t+1$ rather than t since $\delta(t)$ depends on $v(t+1)$. Adapted from Dayan and Abbott (2001).

see, Pearce, George, Redhead, Aydin, & Wynne, 1999), the net prediction $v(t)$ is the sum of the predictions made by each stimulus. In the temporal difference model, the value $v(t)$ is supposed to predict the sum future rewards within a trial

$$v(t) \sim \sum_{s \geq t} r(s). \qquad (2)$$

where $r(s)$ is the value of the reward immediately provided at timestep s. The use of a sum over discrete time steps rather than an integral over continuous time steps is mainly for convenience (see Doya, 1999 for a fully continuous version). Below, we divide time into fine steps, each nominally of 50 ms, which seems roughly the temporal resolution of the dopaminergic activity. As in one form of a Bellman equation (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998), Eq. (2) gives rise to the consistency condition

$$v(t) \sim r(t) + v(t+1), \qquad (3)$$

which states that the sum future reward from time t is the sum of the immediate reward and the future reward from time $t+1$. Thence the temporal difference prediction error, which is the inconsistency between left and right hand sides of Eq. (3), is

$$\delta(t) = r(t) + v(t+1) - v(t). \qquad (4)$$

This prediction error can be used to train the weights using the delta-like rule:

$$\Delta w(t-s) = \epsilon \delta(t) u(s), \qquad (5)$$

where the $u(s)$ picks out the timestep at which the stimulus was presented and ϵ is a learning rate parameter. If there is stochasticity either in the amount or timing of reward, then Eqs. (2) and (3) should involve averages. Strictly speaking, Eq. (5) is called the TD(0) learning rule; there are variants (Sutton, 1988) involving eligibility traces for stimuli (Hull, 1943) which lead to faster learning, at least under some circumstances. The substantial theory underlying temporal difference learning (Bertsekas & Tsitsiklis, 1996; Sutton &

Barto, 1998) indicates circumstances under which the learning rule in Eq. (5) leads to the predictions $v(t)$ becoming correct on average.

Fig. 2 shows the model's construction of the $\delta(t)$ shown in the lower traces of Fig. 1. The two sets of traces show successively the stimulus, u, the reward r, the prediction v, the temporal difference of the predictions $\Delta v = v(t+1) - v(t)$, and the temporal difference prediction error δ, which is the sum of r and Δv. Here, we consider the simple Pavlovian case that a stimulus is shown around time $t_0 = 1$ in a trial, the reward is provided during a small time interval around time 2, and the sum total amount of reward presented is about 2 units. Before learning (left plots), $w(s) = 0$ and so $v(t) = 0$ for all t, and so $\delta(t) = r(t)$. Once learning is complete (right plots), $v(t) = 0$ until the stimulus is shown. Then $v(t) = 2$, the total sum of the reward expected in the future of the trial. Finally, as the reward is delivered, around time $t = 2$, $v(t)$ slowly decreases to 0, since less reward is expected in the future once some has already been provided. Turning this back into a prediction error signal leads to $\delta(t)$ being 2 at $t = 1$, and then being 0 throughout the remainder of the trial. There is no prediction error at the time of the reward, since the reward is expected. If the reward is not delivered, then $\delta(t)$ follows the temporal difference signal Δv. This shows the same peak at $t = 1$, but then is negative at the time the reward should have been delivered, since the delivery of reward is no longer masking this negativity. This TD prediction error model of the activity of dopamine cells has been validated in quite a wide variety of circumstances (Schultz, 1998), including a recent study of Pavlovian blocking (Waelti et al., 2001). Note that these responses, as with those shown later, are those only of a fraction of the relevant cells, albeit often most of the fraction that respond at all to any of the cues. In general, models have yet to account for the full variability that exists in the data.

In an instrumental conditioning context such as an abstract maze task, animals are assumed to choose between different possible actions to maximize their sum future
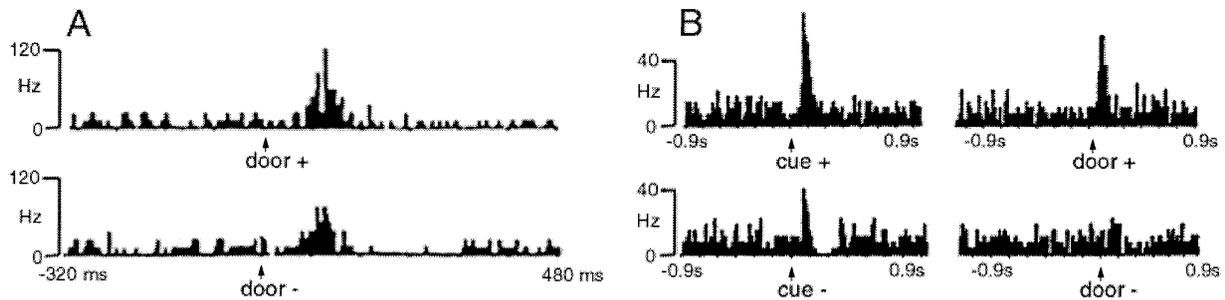
Fig. 3. Generalization responses. (A) Responses of a dopamine cell in an experiment in which there are two doors, one (door +), behind which there is always food, and the other (door −), behind which there is never food. These traces show just the response around the time of the door opening (signaled by a collection of auditory and visual cues). The response to door + is stronger than that to door − ; the latter is followed by a depression below baseline. (B) In this case, a light (cue + or cue −) signals which door is to open, and there is a random interval between 2 and 3 s before the associated door opens. Phasic activation and depression is associated with cue − , as for door − in (A), but there is activation to both cue + and door + . Adapted from Schultz and Romo (1990).

return. This is usually challenging, since rewards might be delayed for long periods. However, the temporal difference prediction error signal can also be used to control action choice. The basic idea, which is a form of a standard engineering algorithm called policy iteration (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998), starts from the fact that the learned values of states estimate the sum of all the delayed rewards starting from those states. Thus, states with high value are good destinations, even if the act of getting to them does not itself lead to substantial reward.

More concretely, a policy is a systematic, though possibly stochastic, way of choosing actions. Consider a case in which the policy is indeed stochastic, but the other aspects of the problem, notably the delivery of reward, are deterministic. Then, applying the TD rule will result in learning the average values of the states in the maze, until $\langle \delta(t) \rangle = 0$, averaging over randomness in the policy. If an action is executed at a state and $\delta(t) > 0$, then this implies that the action may be better than average, since the value of taking the action (the sum of the reward $r(t)$ associated with the action and the value, $v(t + 1)$, of the next state achieved by the action) is bigger than the average value $v(t)$ of the current state. Conversely, if $\delta(t) < 0$ then the action may be worse than average. This temporal difference prediction error signal thus provides immediate criticism for actions (such as turning left or right in the maze) even if the rewards will only be provided at much later times (such as at the goal). Of course, as the animal changes its policy, the values change too. For instance, the goal in a maze might only be 10 steps away from the start given the correct strategy, but on average 200 steps from the start gives a uniform probability of choosing any action at any location.

One action control strategy (Montague, Dayan, Person, & Sejnowski, 1995) involves choosing a random action (i.e. a random movement direction), which is performed continually until $\delta(t)$ is negative. Then a different, random, action is selected and is itself performed continually until $\delta(t)$ is negative, when the process repeats. Montague et al. called this learned klinokinesis after the chemotactic strategy of bacteria. A more general strategy, called the

actor-critic (Barto, Sutton, & Anderson, 1983) uses $\delta(t)$ to train a system for selecting actions, favoring those that lead to positive values over those that lead to negative values. This step is analogous to one part of policy improvement; it must be followed by a step of policy evaluation, which is the application of Eq. (5) to learn the values $v(t)$ associated with the new choice of actions. In practice, values and policies are usually updated in tandem. In learning systems such as the actor-critic, there is an inevitable trade-off between exploitation of existing knowledge about how to get rewards, and exploration for new and good actions that lead to even greater rewards. In studying novelty responses, we consider models of bonuses, which are fictitious rewards that are designed to encourage appropriate exploration.

3. Generalization and uncertainty

Fig. 3 shows two aspects of the behavior of dopamine cells that are not obviously in accord with the temporal difference model. These come from two related tasks (Schultz & Romo, 1990) in which there are two boxes in front of a monkey, one of which always contains food (door +) and one of which never contains food (door −). On a trial, the monkey keeps its hand on a resting key until one of the doors opens (usually accompanied by both visual and auditory cues). If door + opens, the monkey has to move its hand into the associated box to get a food reward. If door − opens, then the monkey has to keep its hand on the resting key until the next trial. Fig. 3(A) shows the response of a single dopamine neuron in just this task. Fig. 3(B) shows the responses in an augmented version of the task in which there are cue lights (called instruction lights) near to the doors, indicating which door will open. Here, cue + is associated with door + and cue − with door − . To complicate matters, in this case there is also a randomly variable interval (2−3 s) between the illumination of the cue light and the opening of the door.

Comparing Figs. 1(B) and 3(A), we see that the response to door + is as we might expect, showing a phasic
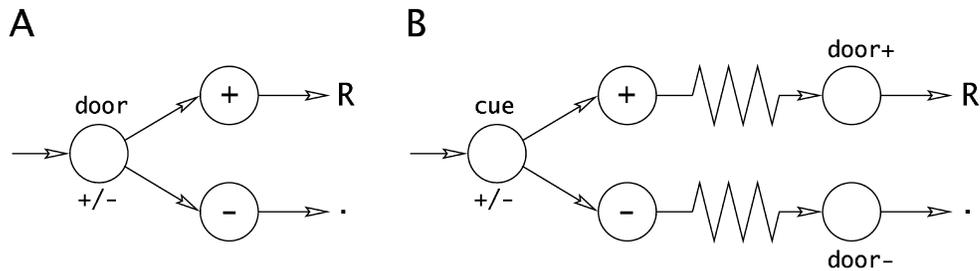
Fig. 4. Model for generalization responses. (A) In the model of Fig. 3(A), the initial information about door + or door − is ambiguous (leading to the state labeled door ± . The ambiguity is resolved in favor of one or other door, leading either to reward (R) or nothing (·). (B) The same ambiguity now applies to the initial cue (giving rise to the ambiguous state cue ± . Now the uncertainty at the time of the door opening is resolved; however, there is a variable delay (squiggly line) between the cue and the door opening.

activation to the delivery of a stimulus that predicts a forthcoming reward. However, the response to door − is not expected under the TD model. Why should dopamine cells be activated by a cue that is known not to be followed by a reward? Something similar is evident in the bottom row of Fig. 3(B), only now to cue − , the reliable predictor that there is to be no reward. Schultz and his collaborators have called these *generalization* responses (Schultz, 1998).

Note first that this activity is probably not associated with an expectation of future reward. For one thing, there is no depression at the time the reward would normally be delivered (Fig. 1). For another, in the blocking study of Waelti et al. (2001), similar responses were observed to stimuli not predictive of reward, and an explicit condition was run in which a reward was unexpectedly delivered. This reward led to the same sort of activity that is evident before learning in Fig. 1, suggesting that the reward was indeed not expected.

Closer study of the bottom rows of Fig. 3(A) and (B) shows a feature that Schultz and Romo (1990), and also subsequent studies, have frequently noted for generalization responses. The phasic activations to door − and cue − appear to last for a shorter time and be lower than to door + and cue + and are accompanied by a phasic *depression* below baseline activity levels, which lasts for something around 100 ms. Different cells show this to different degrees but, for instance, Schultz and Romo (1990) report that it is between two and three times more prevalent for the door − than the door + responses.

Fig. 4 shows our models of these responses. These are based on the idea of partial observability, which will be later interpreted as a refinement of generalization. We suggest that the initial information from the world is ambiguous as to whether the stimulus is actually positive (i.e. door + or cue +) or negative (i.e. door − or cue −). This puts the animal briefly into a state of uncertainty (labeled door ± and cue ±) that is resolved based on further processing of input information that is already available, or the collection of further information, for instance from an attentional shift. Note that physical saccades are usually too slow to account for these timescales. Following resolution of the uncertainty, there are two possible states, one of which leads to

the reward (called +), and other which does not (called −). TD treats the intermediate and resolved states just like all other states, and so learns a value for them which is the average of the values following those states. Consider the example above Fig. 1 in which a sum total reward of 2 was provided around $t = 2$. If the uncertainty when the stimulus is first provided is total (so the probabilities of + and − are equal). Then, the value of the intermediate state is 1, and the values of + and − are 2 and 0, respectively.

Fig. 5(A) shows the output of the TD model of Fig. 4(A) in the equivalents of the door + and door − condition. As in the data of Fig. 3, the model response to door − shows phasic activation followed by phasic depression. The depression comes because the prediction goes from 1 to 0 when the uncertainty is resolved, and so the temporal difference in the predictions Δv becomes negative. In the absence of rectification of below-baseline activity in the model, the depression is more evident in the model than in the actual data. Compared with Fig. 2, the response to door + in Fig. 5(A) is lower and longer. This comes from the time taken for the resolution of the uncertainty. The timescale of the activation in the door + condition and the activation–depression in the door − condition is set by the timescale of the model and the time for the resolution of the ambiguity. Here we made the simplest assumption that the resolution happens completely within one (extra) timestep.

Comparing the responses in Fig. 3(A) and (B), we see that those at the time of cue + and cue − are quite similar to those associated with door + and door − . Further, in the case of cue − , there is almost no change in the activity when door − actually opens. In fact, the existence of the tiny dip in the response at the time of door − , which is also just about apparent in the data, disappears after more extensive training. However, the second main difference from the straightforward temporal difference model that is shown by this experiment is that there is a significant increase in the activity when door + opens. Under conventional TD, this activity might be expected to be predicted away, since the monkey can predict that it is going to get a reward based just on seeing cue + . Contrary to this expectation, Fig. 5(B) shows the output of the temporal
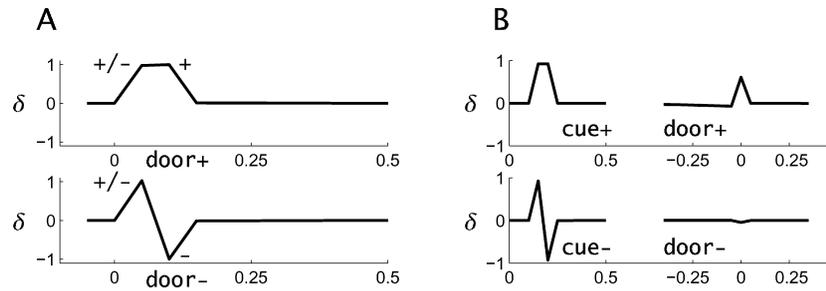
Fig. 5. Dopamine responses for the temporal difference model of Fig. 4, to be compared with Fig. 3. (A) $\delta(t)$ for the case without a prior cue. Here the states $\pm$ and $+$ and $-$ are explicitly labeled. (B) $\delta(t)$ for the case with a prior cue. Here, the responses at the time of **door** $+$ and **door** $-$ are aligned with the opening of the door before averaging (as was also done in the data of Fig. 3(B)).

difference model of the activity in this experiment. Just as in the data, there is extra activity at the time of **door** $+$. The reason for the activity at this time is that there is a random interval between **cue** $+$ and **door** $+$ (see also Suri & Schultz, 1999). In contrast, in the experiment in Fig. 1, by the time the monkey's behavior is automatic, there exists a relatively constant time between stimulus and reward. In the case of Fig. 3, based on **cue** $+$, the monkey can predict *that* it is to receive a reward, but it cannot predict *when* that reward will arrive. That is, there does not exist a setting of weights in experiment 1 that makes a exactly correct prediction of the time of the delivery of the reward, since that time is varying over the trials. In contrast, **door** $+$ is a temporally reliable predictor of the delivery of reward. The consequence of this is that the prediction is ultimately associated with both **cue** $+$ and **door** $+$.

Consistent with this explanation, Schultz, Apicella, and Ljungberg (1993) performed an experiment related to the one in Fig. 3(B). In the condition they called the 'instructed spatial task', there was always a fixed interval between a cue and a trigger instruction that indicated that a movement was now required. In this case, the only activity in the dopamine cells was at the time of the cue, a fact replicated by the temporal difference model since the cue is a temporally reliable predictor of the delivery of reward. They had another condition (called the 'spatial delayed response task') more like the one in Fig. 3(B), and saw a similar pattern of response to that shown there.
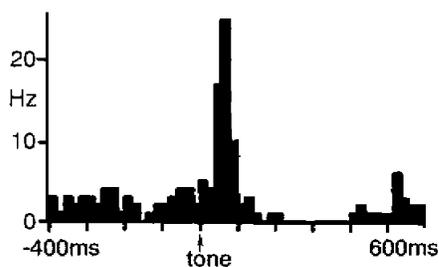

Fig. 6. Novelty response with phasic activation and depression. This shows a histogram of the activity of a single dopamine cell in cat VTA in response to repetitions of an initially novel tone. This neuron shows a clear pattern of activation and depression in response to the stimulus. Adapted from Horvitz, Steward, and Jacobs (1997).

4. Novelty responses

Another main difference between the temporal difference model of the activity of dopamine cells and their actual behavior has to do with novelty. Salient, novel, stimuli are reported to activate dopamine cells for between a few and many trials. One example of this may be the small response at the time of the stimulus in the top line of Fig. 1(A). Here, there is a slight increase in the response locked to the stimulus, with no subsequent decrement below baseline. In this case, the activity could just be the early stages of learning the prediction associated with the stimulus, as subsequently seen more fully in Fig. 1(B) and (C). However, such novelty responses are seen more generally to stimuli that are not predictive of reward. In this case, they decrease over trials, but quite slowly for very salient stimuli (Schultz, 1998). Fig. 6(A) shows a more dramatic example from an experiment in which novel auditory stimuli were played to a cat while the activity of dopamine cells was recorded (Horvitz, Stewart, & Jacobs, 1997). Here, just as in the case of generalization for **door** $-$ and **cue** $-$, the activation in response to the stimulus is rapidly followed by depression below baseline. This response, as with others to very salient stimuli, is quite persistent, lasting for many trials.

4.1. Novelty bonuses

In the theoretical reinforcement learning literature, there are two main theoretical approaches to novelty responses. In one set of theories, novelty acts like a surrogate reward, i.e. something that is itself sought out. This surrogate reward distorts the landscape of predictions and actions, as states predictive of future novelty come to be treated as if they are rewarding. There is some evidence that animals do indeed treat novelty as rewarding (Reed, Mitchell, & Nokes, 1996). It is also computationally reasonable in an instrumental conditioning context, at least in moderation, since it allows animals to plan to visit novel states a number of times so that they can explore the consequences of different actions at those states.

In temporal difference terms, this sort of novelty response, which we call a *novelty bonus*, comes from
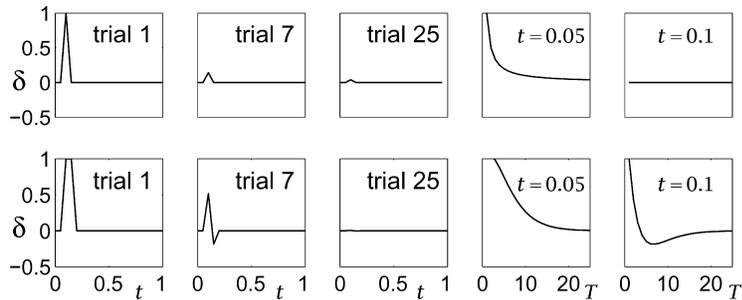
Fig. 7. Activity of model dopamine cells given novelty bonuses. The plots show different aspects of the TD error δ as a function of time t within a trial (first three plots in each row) or as a function of number t of trials (last two). (Upper) A novelty signal was applied for just the first timesteps of the stimulus and decayed hyperbolically with trial number as $1/T$. (Lower) A novelty signal was applied for the first two timesteps of the stimulus and now decayed exponentially as $e^{-0.3T}$ to demonstrate that the precise form of decay is irrelevant. Trial numbers and times are shown in the plots. The learning rate was $\epsilon = 0.3$.

replacing the true environmental reward $r(t)$ at time t with

$$r(t) \rightarrow r(t) + n(u(t), T),$$

where $u(t)$ is the state at time t and $n(u(t), T)$ is the novelty of this state in trial t. Here we imagine that the mechanism that provides $n(u(t), T)$ uses information about the novelty of the stimuli associated with state $u(t)$, and makes the novelty signal decrease over trials t as the stimuli associated with the state become familiar. The effect of the novelty bonus on the temporal difference prediction error is then

$$\delta(t) = r(t) + n(u(t), T) + v(t + 1) - v(t). \tag{6}$$

The upper plots in Fig. 7 show the effect of including such a novelty bonus, in a case in which just the first timestep of a new stimulus in any given trial is awarded a novelty signal which decays hyperbolically to 0 as the stimulus becomes more familiar. Here, a novel stimulus is presented for 25 trials without there being any reward consequences. The effect is just a positive transient which decreases over time, a putative model of the effect shown in the top row of Fig. 1(A). Learning has no effect on this, since the stimulus cannot predict away a novelty signal that lasts only a single timestep. The lower plots in Fig. 7 show that it is possible to get a small phasic depression through learning, (though less dramatic than that in Fig. 6), if the novelty signal is applied for the first two timesteps of a stimulus (for instance if the novelty signal is calculated relatively slowly). In this case, the initial effect is just a positive signal (leftmost graph), the effect of TD learning gives it a negative transient after a few trials (second plot), and then, as the novelty signal decays to 0, the effect goes away (third plot). The righthand plots show how $\delta(t)$ behaves across trials for the first two timesteps. If there were no learning, then there would be no negative transient. The growth and decay of the phasic depression of the dopamine signal is determined by both the speed at which the novelty signal decays and the dynamics of learning.

4.2. Shaping bonuses

Since novelty bonuses distort the reward function, they

can have a deleterious effect on instrumental behavior if they do not decrease to 0 adequately quickly. Ng et al. (1999) suggested a second theory for a form of novelty responses that they called *shaping bonuses*. Shaping bonuses are guaranteed *not* to distort optimal policies, although they can still change the exploratory behavior of agents. Shaping bonuses are derived from a potential function $\phi(u)$ of the state u, so that the estimated value $v(t)$ at time t is replaced by

$$v(t) = v(t) + \phi(t). \tag{7}$$

Here, $\phi(t) = \phi(u(t))$ is the value of the potential function associated with the state at time t and is assumed to be set high for states associated with novel stimuli and that therefore deserve exploration. Also, $v(t)$ is the conventional plastic estimate of the prediction associated with the same state (as in Eq. (1)). If we substitute this into the temporal difference prediction error of Eq. (4), we get

$$\delta(t) = r(t) + \phi(u(t + 1)) - \phi(u(t)) + v(t + 1) - v(t). \tag{8}$$

The difference from the novelty bonus of Eq. (6) is that the shaping bonus enters into $\delta(t)$ via the *difference* between the potential functions for one state and the previous state. If the shaping bonuses are fixed, they can also be seen as coming from the initializing values $v(u)$ given to the states (see also Suri & Schultz, 1999, for application of this to dopamine responses). In fact, it is a standard practice in reinforcement learning, for which there exists a formal basis (Brafman & Tennenholtz, 2001), to use optimistic initial values for states in order to encourage exploration.

Ng et al. (1999) provide a formal proof of this non-distorting property of shaping bonuses. However, an intuition for this result comes from considering the sum of the un-shaped prediction errors over a whole trial

$$\sum_{t \geq 0} \delta(t) = v(t_{end}) - v(0) + \sum_{t \geq 0} r(t),$$

where t_{end} is the time at the end of the trial. Assuming that $v(t_{end}) = 0$ and $v(0) = 0$, i.e. that the monkey confines its reward predictions to within a trial, we can see that any additional influences on $\delta(t)$ whose sum effect over the
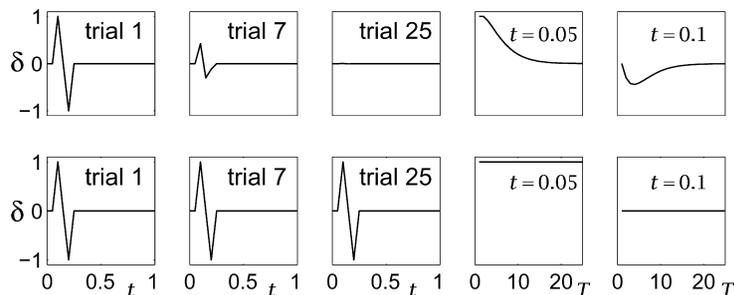
Fig. 8. Activity of the dopamine system given shaping bonuses (the figure has the same format as Fig. 7). (Upper) The plots show different aspects of the temporal difference prediction error δ as a function of time t within a trial (first three plots) or as a function of number T of trials (last two) for the first two significant timesteps during a trial. Here, the shaping bonus comes from a $\phi(t) = 0$ for the first two timesteps a stimulus is presented within a trial, and 0 thereafter, irrespective of trial number. The learning rate was $\epsilon = 0.3$. (Lower) The same plots for $\epsilon = 0$.

course of a whole trial is 0, preserve the sum of future rewards. This is the key quantity that optimal control methods seek to maximize. Responses such as that in Fig. 6 with activation and depression may well indeed have no net effect on the sum of the prediction errors.

The upper plots in Fig. 8 show the effect of shaping bonuses on the temporal difference prediction error. Here, the potential function is set to the value 1 for the first two timesteps of a stimulus in a trial, and 0 otherwise. The most significant difference between this sort of shaping bonus and the novelty bonus of Eq. (6) is that the former exhibits a negative transient even in the very first trial, whereas, for the latter, it is a learned effect. Although the data in Fig. 6 show the transient depression, the development over early trials is not clear. If the learning rate is non-zero, then shaping bonuses are exactly predicted away over the course of normal learning, since $v(t)$ comes exactly to compensate for $\phi(t)$. Thus, even though the same bonus is provided on trial 25 as trial 1, the temporal difference prediction error becomes 0 since the shaping bonus is predicted away. The dynamics of the decay shown in the last two plots is controlled by the learning rate for temporal difference learning. The lower plots show what happens if learning is switched off at the time the shaping bonus is provided—this would be the case if the system responsible for computing the bonus takes its effect before the inputs associated with the stimulus are plastic. In this case, the shaping bonus is preserved.

It may be that as stimuli become less novel, the shaping bonuses associated with them are reduced to 0. In this case the components of the plastic values $v(t)$ that compensate for the shaping bonuses will track $\phi(u(t))$ as it decreases. However, the theoretical guarantee offered by Ng et al. (1999) that shaping bonuses will not distort action learning may not survive such decreases.

5. Discussion

We have suggested a set of interpretations for the activity of the DA system to complement that of reporting prediction error for reward. First, we considered activating

and depressing generalization responses, arguing that they come from short-term ambiguity about the predictive stimuli presented. Second, we considered novelty responses, showing that they are exactly what would be expected where the dopamine cells to be reporting a prediction error for reward in a sophisticated reinforcement learning setting in which an explicit link is made to exploratory behavior.

We accounted for activating and depressing generalization responses by suggesting that the initial information available about a predictive cue is ambiguous, but that this ambiguity is resolved by extra information, that could come as a result of an act of the monkey (such as a shift of attention), or from the result of on-going neural processing. The latter is somewhat analogous to the finding that uncertainty about the direction of motion coming from the aperture problem is resolved over the course of tens of milliseconds in the activity of MT cells (Pack & Born, 2001). Ambiguity is a form of generalization, in that the aspects of the stimulus that are distinctive between door + and door − or cue + and cue − are initially ignored. We have considered the case in which all the dopamine cells receive the same sensory information. Since there is some substantial variation in the behavior of the dopamine cells, it would be interesting to consider a population of cells that receive predictions based on different sensory inputs, some more or less ambiguous about the cues and even about time. It is possible that some cells in the population would be activated only by door + and cue + , i.e. would be instantly unambiguous, whereas others would never tell the difference between the + and − conditions. This might require some extra assumptions about the learning rates associated with different aspects of the representation of the stimuli. We certainly cannot yet fully account for all the multifarious dopamine cell responses.

Generalization responses emerge naturally from a conventional temporal difference framework, provided that ambiguity is taken into account. Novelty responses require extending the framework. We showed that two aspects of the data could be accounted for by two different extensions, namely novelty and shaping bonuses. Novelty

bonuses distort the structure of the rewards, and so can distort things like the policies that optimize long-term rewards. In contrast, shaping bonuses do not distort the optimal policies and they can be exactly learned away. This means that the dopamine system can *multiplex* information about novelty and prediction errors for future reward without damaging interference. This extra information is directly available for target structures, such as the prefrontal cortex and the striatum.

Other evidence confirms the role of dopamine in processing novelty (for a review, see Ikemoto & Panksepp, 1999). For instance, dopamine is implicated in specifically novelty-induced motor activity (Hooks & Kalivas, 1994); there are suggestions that individual differences in the dopamine system of humans links novelty seeking to susceptibility to drug addiction (Bardo, Donohew, & Harrington, 1996); and there is an active, if currently somewhat inconclusive, debate about the role of a specific (D_4) dopamine receptor gene and novelty seeking in humans (Ekelund, Lichtermann, Jaervelin, & Peltonen, 1999; Paterson, Sunohara, & Kennedy, 1999). Concretely, the release of dopamine in the striatum of rats is associated with at least some aspects of attentional orienting to stimuli (Han, McMahan, Holland, & Gallagher, 1997; Ward & Brown, 1996). Thus, the release of dopamine from the phasic activation above baseline might be associated with behaviors that allow novel stimuli or states to be approached and explored. These behaviors might continue despite the subsequent depression below baseline of the dopamine system shown in Fig. 6, for instance. This continuation would be important to organize a temporally extended set of exploratory actions—orienting is one of the shortest of these. The attentional aspect of the theory is broadly in line with Redgrave et al.'s (1999) proposal for the role of phasic dopamine responses, although these authors consider this component as an alternative rather than an addition to a prediction error signal.

We regard these effects associated with dopamine as being *mechanistic*, in the sense that they are hard-wired habits and not learned. In the case of partial observability, as for generalization responses, dopamine release due to the uncertain prediction of reward could directly cause further investigation, and therefore resolution of the uncertainty. When mechanistic and conditioned behaviors conflict, the former seem to dominate, as in the well-known descriptions of the failures of instrumental conditioning (Breland & Breland, 1961).

Apart from exploratory behaviors, the phasic activation of the dopamine system might have other neural effects. For instance, following studies by Braver, Barch, and Cohen (1999), Cohen, Braver, and O'Reilly (1998) and Williams and Goldman-Rakic (1995) suggested that the release of dopamine might gate stimulus information into prefrontal working memory. The novelty response of the dopamine system could thus allow a new stimulus to be stored in working memory until its consequences for future reward have been evaluated.

We considered novelty and shaping bonuses. There is another type of bonus that is often considered in theoretical reinforcement learning, namely an ongoing *exploration* bonus (Dayan & Sejnowski, 1996; Sutton, 1990) which is used to ensure continued exploration. Sutton (1990) suggested adding to the estimated value of each state, a number proportional to the length of time since it was last visited. This ultimately makes it irresistible to go and visit states that have not been visited for a long time. Sutton (1990) actually showed that, in his task, the best behavior came from using exploration bonuses in a $\mathcal{Q}$-learning system. $\mathcal{Q}$-learning (Watkins, 1989) is a theoretically somewhat better motivated version of the actor-critic model; however, its use of state-action values rather than state values as in the critic, makes it harder to map onto the neural substrate. Dayan and Sejnowski (1996) derived a Sutton-like exploration bonus as an adaptive consequence of a model of environmental change. There is no evidence for a continuing exploration bonus in the dopamine data, perhaps not surprisingly, since the tasks undertaken by the monkey offer little possibility for any persistent trade-off between exploration and exploitation.

The most major remaining conflict between the data on the response of the dopamine system and the temporal difference model has to do with the involvement of dopamine in reporting punishment or aversion. Neurophysiological recording studies are somewhat equivocal on this (Guarraci & Kapp, 1999; Mirenowicz & Schultz, 1996), although there is certainly no evidence for the sort of phasic responses to unexpected rewards or reward-predicting stimuli such as those shown in Fig. 1(A) and (B). However, a number of experiments using techniques *other* than neurophysiological recordings of the dopamine cells has suggested that dopamine is released under aversive as well as appetitive conditions (Salamone, 1994). We (Daw, Kakade, & Dayan, 2002) are exploring opponent process models (Grossberg & Schmajuk, 1987; Solomon & Corbit, 1974) involving an putative opponent interaction between dopamine and the dorsal raphe serotonin system, in the computational context of average-case reinforcement learning (Daw & Touretzky, 2000).

Acknowledgments

References

Bardo, M. T., Donohew, R. L., & Harrington, N. G. (1996). Psychobiology

of novelty seeking and drug seeking behavior. *Behavioural Brain Research*, *77*, 23–43.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transaction on Systems, Man and Cybernetics*, *SMC-13*, 834–846.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Cambridge, MA: Athena Scientific.

Brafman, R. I., & Tennenholtz, M. (2001). *R-MAX—A general polynomial time algorithm for near-optimal reinforcement learning*, IJCAI, 953–958.

Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, *46*, 312–328.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*, 681–684.

Church, R. M. (1984). Properties of the internal clock. *Annals of the New York Academy of Sciences*, *423*, 566–582.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1998). In A. C. Roberts, & T. W. Robbins (Eds.), *The prefrontal cortex, executive and cognitive functions*. Oxford: OUP.

Daw, N. D., & Touretzky, D. S. (2000). Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing*, *32*, 679–684.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4–5).

Dayan, P. (2000). Motivated reinforcement learning. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Neural information processing systems, 14*, Cambridge, MA: MIT Press.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, *25*, 5–22.

Dickinson, A., & Balleine, B. (2001). The role of learning in motivation. In C. R. Gallistel (Ed.), *Learning, motivation and emotion* (3rd ed) (*Vol. 3*). *Steven's handbook of experimental psychology*, New York: Wiley.

Doya, K. (1999). Reinforcement learning in continuous time and space. *Neural Computation*, *12*, 243–269.

Ekelund, J., Lichtermann, D., Jaervelin, M.-R., & Peltonen, L. (1999). Association between novelty seeking and type 4 dopamine receptor gene in a large Finnish cohort sample. *American Journal of Psychiatry*, *156*, 1453–1455.

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

Gibbon, J., Malapani, C., Dale, C. L., & Gallistel, C. R. (1997). Toward a neurobiology of temporal cognition: Advances and challenges. *Current Opinion in Neurobiology*, *7*, 170–184.

Gray, J. A., Young, A. M., & Joseph, M. H. (1997). Dopamine's role. *Science*, *278*, 1548–1549.

Grossberg, S., & Schmajuk, N. A. (1987). Neural dynamics of attentionally modulated pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology*, *15*, 195–240.

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

Guarraci, F. A., & Kapp, B. S. (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, R. W., Holland, P., & Gallagher, M. (1997). The role of an amygdalo-nigrostriatal pathway in associative learning. *Journal of Neuroscience*, *17*, 3913–3919.

Hatfield, T., Han, J.-S., Conley, M., Gallagher, M., & Holland, P. (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.

Holland, P. C., & Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences*, *3*, 65–73.

Hooks, M. S., & Kalivas, P. W. (1994). Involvement of dopamine and excitatory amino acid transmission in novelty-induced motor activity. *Journal of Pharmacology, Experimental Therapeutics*, *269*, 976–988.

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

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton (Century).

Ikemoto, S., & Panksepp, J. (1999). *Brain Research Reviews*, *31*, 6–41.

Kakade, S., & Dayan, P. (2000). In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Dopamine bonuses*. NIPS.

Kehoe, E.J (1977). *Effects of serial compound stimuli on stimulus selection in classical conditioning of the rabbit nictitating membrane response*. PhD Thesis, University of Iowa.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, *72*, 1024–1027.

Mirenowicz, J., & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, *379*, 449–451.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*, 725–728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Ng, A. Y., Harada, D., & Russell, S (1999). Policy invariance under reward transformations: Theory and application to reward shaping. *Proceedings of the 16th International Conference on Machine Learning*.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, *4*, 95–102.

Pack, C. C., & Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, *409*, 1040–1042.

Paterson, A. D., Sunohara, G. A., & Kennedy, J. L. (1999). Dopamine D4 receptor gene: Novelty or nonsense? *Neuropsychopharmacology*, *21*, 3–16.

Pearce, J. M., George, D. N., Redhead, E. S., Aydin, A., & Wynne, C. (1999). The influence of background stimuli on summation in autoshaping. *Quarterly Journal of Experimental Psychology, Comparative, Physiology Psychology*, *52*, 53–74.

Redgrave, P., Prescott, T., & 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, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II, current research and theory* (pp. 64–69). New York: Aleton (Century/Crofts).

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, *10*, 284–294.

Salamone, J. D. (1994). The involvement of nucleus accumbens dopamine in appetitive and aversive motivation. *Behavioural Brain Research*, *61*, 117–133.

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

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

Schultz, W. (1992). Activity of dopamine neurons in the behaving primate. *Seminars in the Neurosciences*, *4*, 129–138.

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

Schultz, W., & Romo, R. (1990). Dopamine neurons of the monkey midbrain, contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neuroscience*, *63*, 607–624.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913.

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

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*, 272–283.

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation. I. Temporal dynamics of affect. *Psychological Review*, *81*, 119–145.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks*, *15*(4–6), 523–533.

Suri, R. E. & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*, 871–890.

Sutton, R. S. (1988). Learning to predict by the methods of temporal difference. *Machine Learning*, *3*, 9–44.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Machine Learning, Proceedings of the Seventh International Conference*, 216–224.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tremblay, L., & Schultz, W. (2000a). Reward-related neuronal activity during go–nogo task performance in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*, 1864–1876.

Tremblay, L., & Schultz, W. (2000b). Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*, 1877–1885.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43–48.

Ward, N. M., & Brown, V. J. (1996). Covert orienting of attention in the rat and the role of striatal dopamine. *Journal of Neuroscience*, *16*, 3082–3088.

Watkins, C. J. C. H (1989). *Learning from delayed rewards*. PhD dissertation, University of Cambridge.

Whitelaw, R. B., Markou, A., Robbins, T. W., & Everitt, B. J. (1996). Excitotoxic lesions of the basolateral amygdala impair the acquisition of cocaine-seeking behaviour under a second-order schedule of reinforcement. *Psychopharmacology*, *127*, 213–224.

Williams, G. V., & Goldman-Rakic, P. S. (1995). Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature*, *376*, 572–575.