Behavioral/Systems/Cognitive

Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network

Wei-Xing Pan,[1] Robert Schmidt,[2] Jeffery R. Wickens,[2] and Brian I. Hyland[1]
Departments of [1]Physiology and [2]Anatomy and Structural Biology, School of Medical Sciences, University of Otago, Dunedin 9001, New Zealand

Behavioral conditioning of cue–reward pairing results in a shift of midbrain dopamine (DA) cell activity from responding to the reward to responding to the predictive cue. However, the precise time course and mechanism underlying this shift remain unclear. Here, we report a combined single-unit recording and temporal difference (TD) modeling approach to this question. The data from recordings in conscious rats showed that DA cells retain responses to predicted reward after responses to conditioned cues have developed, at least early in training. This contrasts with previous TD models that predict a gradual stepwise shift in latency with responses to rewards lost before responses develop to the conditioned cue. By exploring the TD parameter space, we demonstrate that the persistent reward responses of DA cells during conditioning are only accurately replicated by a TD model with long-lasting eligibility traces (nonzero values for the parameter λ) and low learning rate (α). These physiological constraints for TD parameters suggest that eligibility traces and low per-trial rates of plastic modification may be essential features of neural circuits for reward learning in the brain. Such properties enable rapid but stable initiation of learning when the number of stimulus–reward pairings is limited, conferring significant adaptive advantages in real-world environments.

Key words: **ventral tegmental area; temporal difference algorithm; dopaminergic; extracellular recordings; reward; associative learning**

Introduction

Midbrain dopamine (DA) cells play a central role in reward-mediated learning in animals, and their activity follows classical learning rules (Schultz, 1998, 2002; Waelti et al., 2001). Furthermore, several features of DA cell activity match properties of the prediction error signal of the temporal difference (TD) algorithm for machine learning, leading to the hypothesis that DA cell activity may be providing a teaching signal within a neural analog of a TD learning system in the brain (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Daw et al., 2003; Nakahara et al., 2004). Because the underlying algorithmic processes of TD are well understood (Sutton, 1988; Sutton and Barto, 1998), this link between biological and machine learning processes offers a powerful way to progress our understanding of the neural mechanisms of reward-mediated learning (Barto, 1995; Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Suri and Schultz, 1998, 1999; Rao and Sejnowski, 2001; O'Doherty et al., 2003, 2004; Seymour et al., 2004).

DA cells show responses to unexpected rewards, but after training with cue–reward pairings, they respond to the cue more than to the predicted reward (Ljungberg et al., 1992). A TD mechanism that has been suggested to underlay this process involves a stepwise shift in the timing of prediction error signal, so that it drifts gradually in latency until it follows the cue (Montague et al., 1996; Schultz et al., 1997). Responses of DA cells are therefore predicted to show similar gradual shifts in timing during the process of conditioning. Two additional predictions arise from this proposed mechanism; first, that the response to the reward will decline to zero before the response to the cue develops, and second, that an intermediate redundant cue will not trigger a response during the learning process (Montague et al., 1996; Schultz et al., 1997). These predictions have not been tested rigorously, because few studies have monitored the activity of single dopamine cells during the acquisition of completely novel learning. Thus, the specific temporal representation parameters of TD models remain physiologically under-constrained (Schultz et al., 1997).

In this study, we first investigated in conscious rats whether DA cells conform to these specific patterns of prediction error signaling by recording from single neurons while animals first learned the association between cues and rewards. The results were not consistent with previous model predictions. We then performed a systematic exploration of the parameter space of the TD algorithm to determine whether it can encompass the activity patterns displayed by the DA cells.

Materials and Methods

Electrophysiological methods

Animals. All procedures were approved by the University of Otago Animal Ethics Committee. Fifty-three male, Wistar rats weighing 250–400 g

Received April 15, 2005; revised May 13, 2005; accepted May 14, 2005.

This research was supported by grants from the New Zealand Neurological Foundation, New Zealand Lottery Grants Board, and the Marsden Fund.

Correspondence should be addressed to Dr. Brian Hyland, Department Physiology, Otago School of Medical Sciences, P.O. Box 913, Dunedin 9001, New Zealand. E-mail: brian.hyland@otago.ac.nz.

DOI:10.1523/JNEUROSCI.1478-05.2005

 0270-6474/05/256235-08$15.00/0

were used for this study. Under full anesthesia (sodium pentobarbital; 60 mg/kg, i.p.) and using aseptic technique, a bundle of eight microwire recording electrodes (0.001 inch Formvar-insulated nichrome; A-M Systems, Carlsborg, WA) glued in a 30 gauge stainless steel guide cannula and mounted on an on-head microdrive was implanted through a small burr hole using stereotaxic technique. The electrode tips were left just above the dopamine cell groups of the substantia nigra pars compacta or the ventral tegmental area (5.0–5.5 mm posterior to bregma, 0.5–2.0 mm lateral to midline, and 6.5–7.0 mm below the surface of skull). The electrode assembly was affixed to the skull using six stainless steel screws and dental acrylic. All animals were allowed 1 week to recover from surgery before any recordings.

Behavioral testing. After recovery from surgery, rats were fluid deprived for 24 h and then familiarized with the recording chamber, a clear acrylic box of floor area 25 × 16.5 cm located in a quiet, darkened room. The rats were trained to obtain fluid from a recessed spout in the wall of the chamber. Small volumes (~0.05 ml) of water sweetened with saccharin (0.005 M solution) were delivered to the spout by briefly releasing a solenoid valve (Med Associates, Georgia, VT). The moment of fluid delivery was indicated by a low-frequency click generated by the solenoid. Licking at the spout was detected by the tongue breaking an infrared beam across the spout opening. Initially, the solenoid was manually activated whenever rats spontaneously explored the drinking spout, which usually resulted in extended attention to the spout after only a few minutes in the first session. They were then exposed to an automated random-reward paradigm. In this condition, fluid boluses were delivered at pseudorandom delays (10–20 s) after retrieval of the previous bolus (indicated by detection of licks at the spout). We ran the random-reward condition while searching for cells in case it activated otherwise-quiescent neurons. Animals were therefore exposed to hundreds of solenoid–reward pairings by the time the recordings were made.

DA cells were recorded under a series of control, conditioning, and omission behavioral paradigms. Control paradigms were: random reward (detailed above), to assess baseline responsiveness of the cell to the reward and reward delivery, and cues only, which tested for responses to a novel stimulus outside of any task context (no fluid available). The stimulus was 0.5 or 2.0 s duration and consisted of a 4.5 kHz tone delivered from a speaker (SonAlert; Med Associates) mounted immediately above the drinking spout. Conditioning trials involved exposing the animal to a cued-reward paradigm, in which solenoid activation was preceded by one or two cues. Conditioning with the cued-reward paradigm was continued until a robust response was seen to the cue, or the cell was lost. In one-cue experiments, the solenoid activation occurred either at the end of the cue or after a 1 s delay. In the two-cue experiments, cues were separated by an intercue interval equal to the cue duration, and the solenoid was activated at the end of the second cue. In most two-cue experiments, both cues were tones, but for some cells, house-light illumination was used for the second cue. In all paradigms, successive trials were separated by pseudorandom intertrial intervals of 10–20 s.

Those cells still present after demonstrating conditioned responses to the cues were tested with the omission paradigm. Here, for two-cue tests, on any one trial, there was a probability of 0.6 of a standard cued-reward sequence, as described above, and a probability of 0.2 for each of two oddball sequences. These consisted of one sequence in which there was no solenoid activation after the cues (omit reward) and another in which the second cue was omitted but with activation of the reward solenoid at the usual time relative to the first cue (omit cue 2). In one-cue tests, the paradigm consisted of standard cued-reward (0.8 probability) and omit-reward (0.2 probability) trials.

For cell recordings, signals from the electrodes were amplified (2–10,000×), filtered (0.2–10 kHz bandpass), digitized (20 kHz), and recorded on computer using Discovery software (DataWave Technologies, Berthoud, CO). The extracellularly recorded action potentials were discriminated from each other and from noise based on wave shape using the spike-sorting features of DataWave Personal Scientific Workstation software.

Identification of DA cells. Recorded cells were screened and discarded if the firing rate was >10 Hz or the action potential <1 ms in duration, which represents the minimum cutoff between DA and slow-firing non-DA cells in recordings with the filter settings commonly used in recordings from conscious rats and monkeys (Schultz, 1986; Romo and Schultz, 1990; Ljungberg et al., 1992; Hyland et al., 2002). From this group, only cells that were also profoundly (>50%) inhibited by the dopamine agonist apomorphine (750 μg/kg, i.p.) or the D_2 receptor-selective agonist quinpirole (400 μg/kg, s.c.) were accepted as presumed DA cells (Bunney et al., 1973; Aghajanian and Bunney, 1977; Grace and Bunney, 1980, 1983; Aebischer and Schultz, 1984; Hyland et al., 2002). The position of the electrode tracks was confirmed after completion of the experiments. A lesion was generated at the tip of recording wires by passing DC current (9 V for 1–2 min). After 5–10 d survival time, rats were killed by anesthetic overdose, perfused with saline then formalin solution, the brains sectioned on a freezing microtome, and the position of the marking lesions and cannula tracks mapped on standard atlas sections (Paxinos and Watson, 1997).

Data analyses. Changes in firing rate associated with task events were examined by constructing trial-by-trial dot raster displays and averaged peri-event time histograms. Histograms routinely had bin widths of 25 ms, but 5 ms bin widths were used for measuring latency of responses. Latency was measured from the left edge of the first bin of a peak or trough after an event that was ≥2 SDs of the mean baseline firing rate (calculated from the 2 s before the first event). Population histograms were generated by calculating the average firing rate for equivalent bins across the individual cell histograms, for groups of cells recorded at similar stags of training. To enable comparison of changes in firing rate induced by task events under different conditions across different cells, we normalized histogram firing rates by calculating a modulation index (MI) for each bin i in the original histogram, as follows:

$$MI(i) = \frac{R(i) - R(b)}{R(b)}, \quad (1)$$

where $R(i)$ was the firing rate in bin i, and $R(b)$ was the average firing rate over a 500 ms period beginning 1 s before the time of the first cue (baseline firing rate).

TD modeling

The TD algorithm we used to model DA cell activity was based on that described by Montague et al. (1996). Two stimuli, one at time step 5 and one at time step 15, preceded reward occurrence at time step 20. To prevent signals providing a predictive effect across trials, we implemented pseudorandom gaps between the last time step of one trial and the first time step of the next that were at least as long as trials, so that state vectors were empty by the start of each trial.

The goal of the TD algorithm is to learn to predict future rewards. The future rewards at time step t can be expressed as a value function $V(t)$, which is equal to the expected value ($E[\cdot]$) of the discounted sum of all future rewards in which the rewards $r(i)$ contribute less if they are farther away in time, according to the discount factor γ (Montague et al., 1996; Schultz et al., 1997; Sutton and Barto, 1998), as follows:

$$V(t) = E\left[\sum_{i=t+1}^{\infty} \gamma^{i-t-1} r(i)\right]. \quad (2)$$

Because $V(t)$ is not known, TD models proceed by calculating an estimate $P(t)$ of $V(t)$. To achieve this, in our model, each sensory stimulus l was represented by a state vector $\boldsymbol{x_l}$, with dimension equal to the number of time steps in a trial. Stimuli were represented in the state vectors as a complete serial compound stimulus. In this kind of representation, if a stimulus has occurred, then one component of the vector is "1," and all others are "0." Which of the components is set at 1 at any one time depends on the number of time steps that have passed since stimulus occurrence. Thus, component q of the state vector is 1 if the stimulus presentation was exactly $q - 1$ time steps ago and 0 otherwise (Sutton and Barto, 1990; Montague et al., 1996; Schultz et al., 1997).

Each state vector was associated with a matching weight vector $\boldsymbol{w_l(t)}$. Estimates (predictions) of all future rewards $P_l(t)$ were formed for each stimulus l by the dot product of the state and weight vectors as follows: $P_l(t) = \boldsymbol{x_l(t)} \cdot \boldsymbol{w_l(t)}$.

The total reward prediction provided by all stimuli $P(t)$ was formed by the sum of all stimulus-specific predictions (Suri and Schultz, 1999) as follows:

$$P(t) = \sum_{l} P_l(t). \quad (3)$$

In the temporal difference algorithm, an estimate of the expected reward at a particular time step t is derived from the difference between the prediction of all future rewards at that time step, $P(t)$, and the prediction of all future rewards that was generated at the previous time step, $P(t-1)$ (Sutton, 1988). The difference between these successive predictions represents reward expected to occur at time t. It is this component of the algorithm, which involves calculating a difference between two successive time steps, that gives rise to the term temporal difference. We therefore refer to this value as $TD(t)$: $TD(t) = P(t-1) - \gamma P(t)$. The discount factor γ was set at 0.98 (Suri and Schultz, 1998).

The crucial step is then to compare this expected reward for time step t with the actual reward that occurs at that time step, represented by the scalar value $r(t)$. Any difference between the actual and predicted rewards generates the prediction error $\delta(t)$: $\delta(t) = r(t) - TD(t)$.

All weight vectors were initialized with zeros. The prediction error was used as a teaching signal to update the vector weights according to the following weight change rule: $\Delta \boldsymbol{w_l(t)} = \alpha\delta(t)\boldsymbol{e_l(t)}$, where $0 < \alpha \leq 1$ is the learning rate parameter, and $\boldsymbol{e_l(t)}$ is the eligibility trace for stimulus l. The eligibility trace enables weights associated with previous time steps to be altered by prediction error at time step t and was calculated recursively (Sutton, 1988) by the following: $\boldsymbol{e_l(t+1)} = \lambda \boldsymbol{e_l(t)} + \boldsymbol{x_l(t)}$, where λ is the eligibility trace decay parameter. If $\lambda = 1$, all weights are affected equally [TD(1) model]. If $\lambda = 0$, only weights associated with the immediate previous state are affected [TD(0) model]. The effect when $0 < \lambda < 1$ is to exponentially bias weight change so that vectors representing the most recent events are affected the most by the prediction error signal [TD(λ) model].

The baseline spiking activity of DA cells (~5 Hz) can be equated to a state of 0 prediction error (Schultz et al., 1997). Spike activity can only be suppressed from this baseline down to 0 Hz. On the other hand, DA cells can be transiently excited to instantaneous rates of up to 100 Hz (Hyland et al., 2002). They are therefore very asymmetrical in the range of positive and negative prediction error signal amplitudes they can generate. To model this, we applied a limit to the amplitude of the negative prediction error such that negative and positive prediction errors were scaled in a similar way to the dynamic range of dopamine cell activity. In preliminary runs of the model, we found that the largest positive prediction error values approached 1, which can be equated to the maximum firing rate (100 Hz) in DA cells. The maximum negative prediction error was therefore limited to −0.05.

Results

From 45 neurons that were possibly dopaminergic on electrophysiological criteria, 24 recorded from 11 rats were confirmed DA neurons based on having a clear inhibitory response to dopamine agonist drugs (Fig. 1*A*). This group had an average firing rate of 5.7 ± 3.5 Hz (mean ± SD), action potential duration of 1.5 ± 1.3 ms, and were located in the midbrain dopamine cell fields (Fig. 1*B*) including both medial A9 (substantia nigra pars compacta) and lateral A10 (ventral tegmental area and supramammillary nucleus). There was no regional difference in response patterns, so all cells from all regions were pooled for analysis.

Conditioning of DA cell responses to cues

Dopamine cells displayed rapid response plasticity during classical conditioning of cue–reward associations. DA cells responded to random reward, and then new responses to tone cues developed rapidly when these were paired with reward. Typical features of the development of conditioned responses during learning are illustrated in Figure 2. New short-latency responses (excitations, 74 ± 33 ms; inhibitions, 53 ± 9 ms) began within the first block of training. This cellular conditioning occurred in most cells tested (8 of 11) and developed over a similar time course to the development of conditioned behavior (licking responses to the cues). In 13 other dopamine neurons, recordings were made after acquisition of the behavioral response. Responses to the tone cues were seen in 11 of these cells. These responses were diminished or extinguished after continued ex-

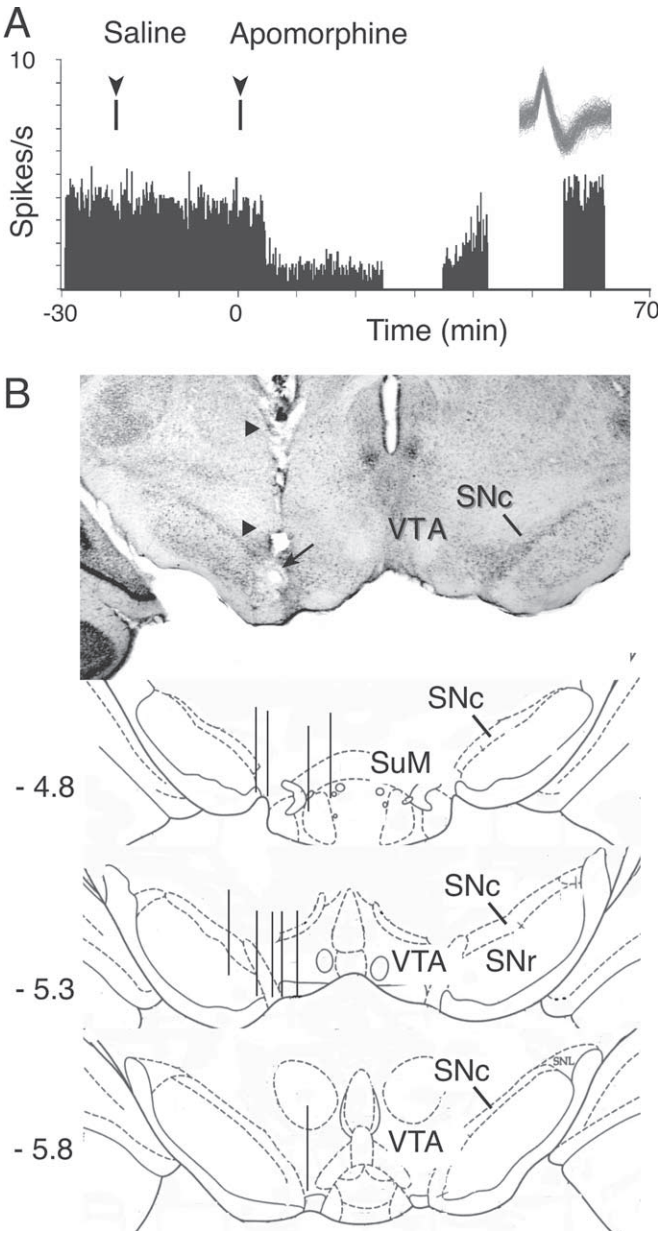

Figure 1. Identification of DA cells. ***A***, Electrophysiological and pharmacological criteria. Rate meter histogram shows low baseline firing rate, lack of response to control saline injection, and inhibitory response of a typical presumed DA neuron to injection of apomorphine (750 μg/kg, i.p.). Gaps in histogram are periods during which recording was suspended. Inset shows overlaid recorded waveforms from this cell, 2 ms total time. ***B***, Location of recorded cells. Histological section shows a cannula track (arrowheads) approaching midbrain DA cell fields and marking lesion (arrow) at the site of recording of a presumed DA neuron. Atlas section diagrams (Paxinos and Watson, 1997) show reconstructed positions of all tracks on which DA cells were recorded (anteroposterior coordinate in mm, relative to bregma, at left). Some tracks yielded more than one cell. SNc, Substantia nigra pars compacta; SNr, substantia nigra pars reticulata; SuM, supramammillary nucleus; VTA, ventral tegmental area.

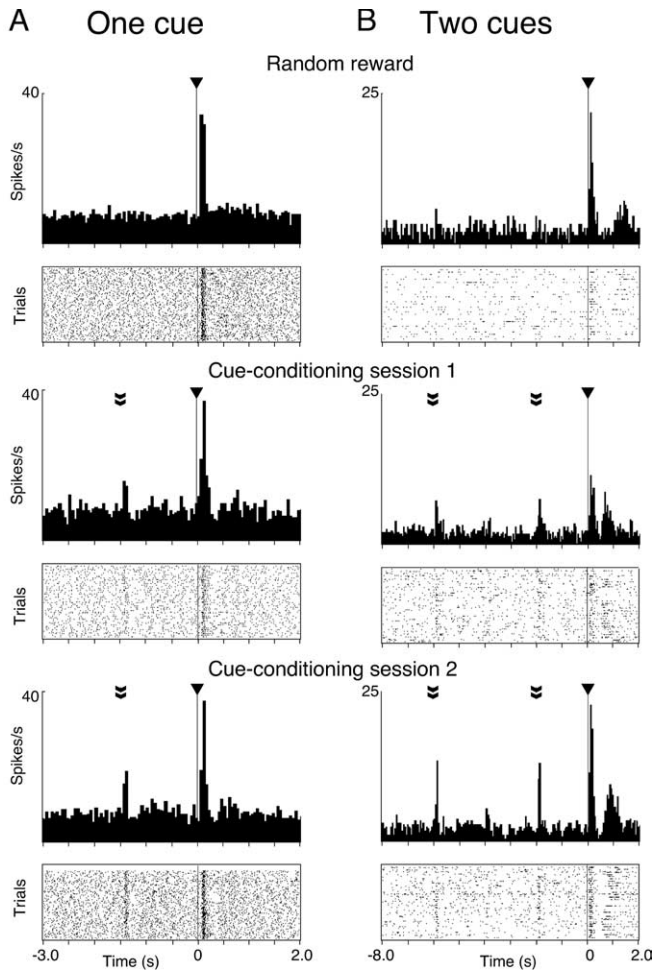

Figure 2. Development of conditioned responses to cues in two different DA neurons. ***A***, DA neuron conditioned with a single tone cue. The top histogram and dot raster show average and trial-by-trial responses to solenoid (filled triangle) in random-reward paradigm. The dot raster shows time of action potentials on individual trials, in original order, first trial at the bottom. The middle and bottom histograms and rasters show the responses of the cell in successive conditioning blocks in which the solenoid was paired with the cue (onset at double arrowhead). ***B***, DA cell conditioned with the two-cue paradigm. Panel layout and labels as for ***A***. Neither cell responded to cues before conditioning (data not shown).

posure to the cues in the absence of rewards, consistent with these also being conditioned responses.

Importantly, in contrast to previous model expectations, we found that new conditioned responses made their first appearance at a constant, short latency after predictive cues in conditioning with either a single (Fig. 2*A*) or double (Fig. 2*B*) cue. These new responses developed rapidly, within a few trials, as can be seen in the dot raster displays of trial-by-trial cell activity. Thus, in these experiments, we found no evidence for a gradual shift in latency of the responses. Similarly, we found that in all cells conditioned with two cues presented in sequence there was a clear response to the second cue, which persisted throughout the experiment (Fig. 2*B*). An additional cell tested with a tone–light cue sequence (data not shown) also responded to both cues.

Finally, we found that responses to reward remained long after the new responses to the predictive cues had developed. This was true both for cells recorded with a single cue (Fig. 2*A*) and in the two-cue paradigm (Fig. 2*B*) and was seen in all cells recorded in early training. However, this effect appeared to be dependent on the duration of training. Figure 3 shows a comparison of single-cell examples and population average data obtained from early (Fig. 3*A*) and late training (Fig. 3*B*). It is notable that unlike the persistent responses seen in DA cells recorded early in training, cells recorded after many sessions of cue–reward pairing were unresponsive to predictable reward delivery (Fig. 3*B*, middle). These cells were still capable of responding to the reward, as demonstrated by robust responses to solenoid activation when delivered without any cues (Fig. 3, left) or on trials in which the second cue was unexpectedly omitted (Fig. 3, right). These data indicate that the development of conditioned responses to cues does not directly lead to abolishment of responses to rewards. Rather, responses to predicted rewards are only completely lost after a period of training that extends beyond that required to first establish conditioned responses to predictive cues.

We also observed that compared with trials in which reward was preceded by two cues, omission of the second cue (Fig. 3, right) restored responses to the solenoid to levels seen with random reward (Fig. 3, left). This was the case both early and late in training. Thus, learning about the intermediate cue was not blocked by the presence of the preceding one, in the sense that the presence or absence of the second cue was taken into account in determining cell responses to the reward.

Prediction error signaling in TD models includes a negative signal at the time of expected rewards and cues if these fail to materialize (Schultz et al., 1997). For DA cell firing at the time of omitted second cue in the two-cue paradigm, this effect was weakly seen in some cells but not others, reflected in the absence of clear inhibitory troughs in population histograms (Fig. 3*A*,*B*). Quantitative analysis of the period 200 ms after the expected time of the second cue compared with baseline in seven cells tested with cue 2 omission in the two-cue paradigm yielded an average modulation index at the time of the expected cue of -0.12 ± 0.17 (mean $\pm$ SD; where 0 = normalized baseline), which represented a nonsignificant difference from baseline firing rate. However, a significant reduction in firing rate was observed when expected rewards were omitted (average modulation index, -0.21 ± 0.20; $p < 0.01$; paired t test; $n = 11$).

TD model

The differences between the predictions of previous TD models and our data raise questions about the ability of TD models to reflect DA cell activity in the rat brain. To investigate whether our observations of DA cell activity can be reconciled with current TD models, we explored the performance of the TD algorithm over a range of parameter settings. Our TD model (Fig. 4*A*) was based on the algorithm described by Montague et al. (1996) but modified by limiting the amplitude of the negative prediction error to match the limited range over which DA cells can be inhibited in activity (see Materials and Methods).

We used a TD model with the same sequence of cues and reward as in our DA cell recording experiments to investigate the role of two key parameters, α and λ. The parameter α (learning rate) sets the magnitude of vector weight changes induced by the prediction error signal. The eligibility-trace decay parameter λ determines to what extent predictions that are farther away in time are altered by the weight update (Sutton and Barto, 1998). We found important effects of the values of these parameters on the behavior of the model. Figure 4*B* shows model prediction error output over learning trials for $\lambda = 0$ and $\alpha = 0.05$. These settings produce a pattern of prediction error output similar to that reported previously (Montague et al., 1996; Schultz et al., 1997). In particular, there is no clear time-locked response to the second cue, and the response to the first cue develops only after a

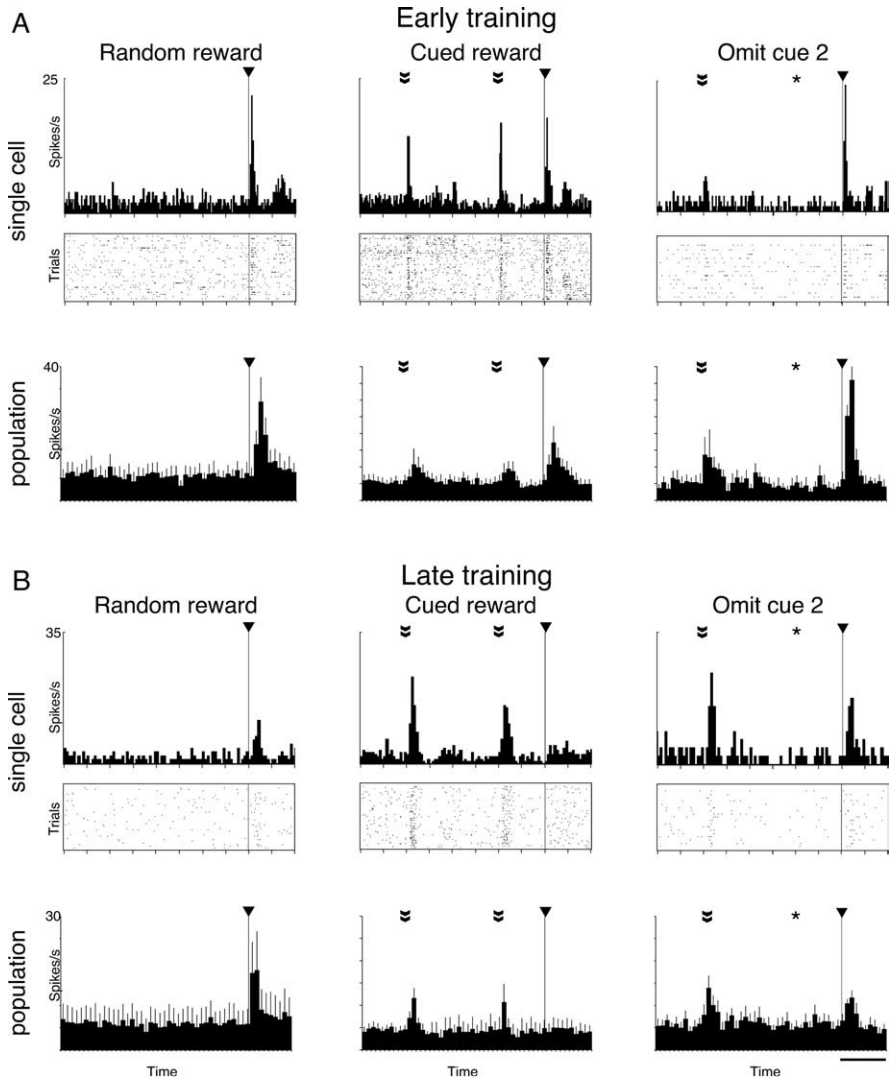

Figure 3. Effect of training on responses of DA cells to reward delivery under different states of predictability. ***A***, Data from animals early (≤6 blocks) in training. The top histograms and dot rasters show a single example cell (same cell as Fig. 2*A*) in the random-reward paradigm (left), cued-reward trials from within the omission paradigm (middle), and omit cue 2 trials of the omission paradigm (right). Population histograms below the rasters were calculated by averaging 50 ms bin counts across all individual histograms ($n = 6, 8$, and 3, respectively) and converting to instantaneous frequency. Error bars represent SEM. ***B***, Data from animals that had been exposed to ≥10 blocks of conditioning (late training). Panel layout and labels as for ***A***. Population histograms were constructed from five, five, and four individual histograms, respectively. Horizontal calibration bar shows 0.5 s for all panels except the example cell data in ***A*** (2 s). Asterisks show time at which cue 2 would normally occur (omit cue 2 trials).

gradual shift in response latency from the time of the reward to the time of the first cue. During this time, there is no response time-locked to either the first cue or the reward, so that there is no overlap in time between cue and reward responses. The pattern is clearly different to that seen in DA cells (Fig. 4*D*).

In contrast, Figure 4*C* shows the result for $\lambda = 0.9$, with $\alpha = 0.005$. With these values, the prediction error signal output reproduced key elements of the pattern seen during learning in our DA cell recording experiments (Fig. 4*D*). In particular, there is a clear time-locked response to the intermediate cue, and new responses appear at their final latency, rather than progressively shifting in time toward the first cue. Early in training, newly developed responses to cue 1 and to reward coexist, with predicted-reward responses only disappearing with additional training, just as seen in the cell data.

Examination of three-dimensional (3-D) plots of model performance over a range of parameter settings (Fig. 5) showed that high values of λ were the critical factor that enabled the model to learn about the significance of cues without gradual trial-by-trial stepwise migration of responses from rewards to cues. As λ was reduced below a threshold level (~0.6 for the number of time steps modeled here), step-wise migration of prediction error signals appeared for all values of α. The setting for α appeared less critical, only failing to produce learning at very high values, as has been noted previously (Sutton and Barto, 1998; Suri, 2001). Otherwise, the value for α, given a suitable setting for λ, determined the number of trials needed for growth of the cue response and for abolition of the reward response and therefore the number of trials over which these responses coexisted. In Figure 4*C*, the setting for α has been chosen so that suppression of responses to predicted rewards occurred only after hundreds of trials, as seen in the cell data.

The relationship between model and cell data are further explored in Figure 6, which compares model prediction errors with DA cell activity when cue 2 was omitted after training in the two-cue paradigm. The model (Fig. 6*A*) generated small negative prediction errors when the expected cue was omitted, replicating the fact that only small inhibitions were seen in DA cell recordings. This was expected because of the limit placed on the possible range of negative prediction errors. Of more interest, this analysis also showed that omitting the second cue in trials in the model restored the amplitude of the prediction error response to the reward toward the level seen when the reward was entirely unpredicted by cues. This occurred in the model despite the presence of a conditioned prediction error response to that cue and was the case both early in training, when reward responses were only mildly suppressed, and later, when reward responses were abolished by the presence of preceding cue signals. However, later in training, the amplitude of the restoration was somewhat less. Thus, in the TD model, the second and seemingly redundant cue develops a role in the prediction of reward. The relevant cell population data from Figure 3 is shown in Figure 6*B*, illustrating the similar restoration of responses and suggesting differential amplitude of restoration depending on stage of training.

Discussion

These results extend previous findings from rats that DA cells respond to sensory cues predicting reward (Miller et al., 1981; Kosobud et al., 1994; Kiyatkin and Rebec, 2001; Hyland et al., 2002) and show for the first time that these are contingent on cue–reward association and arise during acquisition of classically conditioned behavior. We also noted significant depression of DA cell activity at the time of omitted rewards, indicating that rat

DA cells are capable of signaling negative prediction error. The amplitude of inhibitions was small, consistent with the fact that inhibitory responses can be difficult to detect and quantify in neurons that already have a low baseline rate of activity, as noted previously in primate studies (Fiorillo et al., 2003; Morris et al., 2004). The activity of dopamine cells in signaled reward tasks in rats thus appears very similar to that seen in monkeys. Crucially, plasticity of responding as evidenced by the development of conditioned responses during task learning is necessary if DA cells are to provide a reward prediction error function in a TD-type process.

However, at first, the data appeared to be inconsistent with TD approaches, because they failed to match specific predictions from previous modeling. Rat DA cells showed no evidence of step-wise migration of response latencies during initial conditioning, which was predicted. In further contrast to expectations, we consistently observed responding to both the first cue and the reward during initial training. We also noted clear and persistent responses to the intermediate cue in the two-cue paradigm. Neither of these phenomena should occur if a step-wise shift in prediction error explained the overall shift from reward to cue.

Inspection of existing data sets from primates also provides no support for the gradual migration of a response during learning (Mirenowicz and Schultz, 1994; Hollerman and Schultz, 1998; Waelti et al., 2001; Takikawa et al., 2004) and suggests that DA cells show simultaneous responding to both cue and reward, at least early in learning (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz and Schultz, 1994; Fiorillo et al., 2003; Takikawa et al., 2004). On the other hand, early studies comparing DA cell responses early and late in training demonstrated that responses to rewards were diminished in highly trained animals (Ljungberg et al., 1992). We found a similar result here; in rats in which several DA cells were recorded sequentially and that therefore had more exposure to the cue–reward pairing, responses to predicted rewards could eventually be abolished. Together, these data suggest that there are different time courses for acquisition of new conditioned responses to cues and the loss of responses to the rewards predicted by those cues. This stands in clear contrast to the sequential changes and mutually exclusivity of cue and reward responses that are implied by previous TD models of DA cell activity (Montague et al., 1996; Schultz et al., 1997).

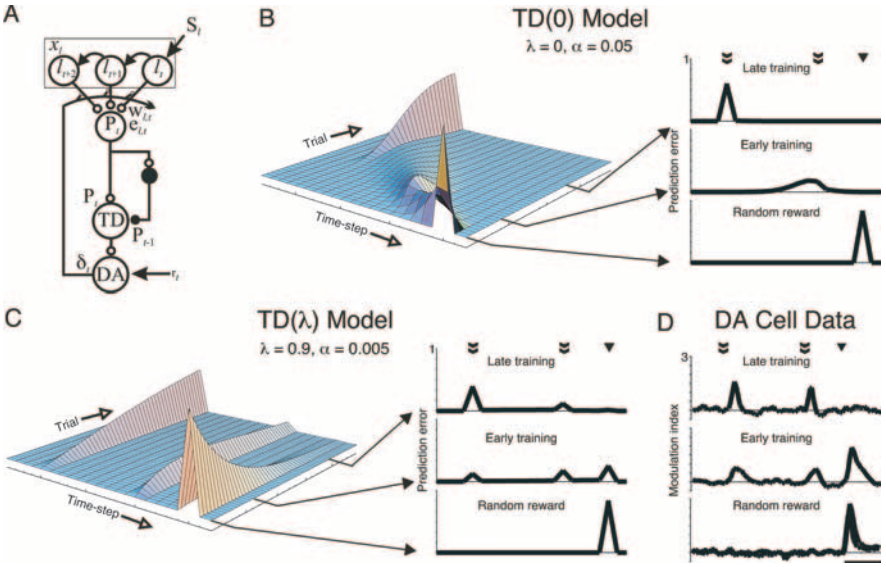

Figure 4. TD models of DA cell activity during learning. ***A***, Simplified network diagram summarizing main features of the TD model. Each sensory signal S_i is represented by a state vector $\boldsymbol{x}_i$, which encodes the signal over time (curved arrows). At any time step t, output of the state vector gives rise to a prediction $P(t)$, which depends on the weight $\boldsymbol{w}_i(t)$ of the component representing the signal at that time. Component weights are eligible for modification after the occurrence of S_i, depending on the value of the eligibility trace $e_i(t)$. Predictions at time step t are subtracted from the prediction of the previous time step to generate the temporal difference $[TD(t)]$. The TD output is compared with the value of the reward signal $r(t)$ to generate the prediction error $\delta(t)$, equated with DA cell activity. This then modifies weights of the state vector $\boldsymbol{x}_i$ representing S_i depending on their eligibility and the learning rate (α). ***B***, Surface plot shows TD prediction error amplitude (vertical axis) during each trial, over the course of learning (400 trials), with $\lambda = 0$ and $\alpha = 0.05$. Grid lines show each time step on every 10th trial. Cues were delivered at time step 5 and 15 and reward at time step 20. Line graphs show prediction error profiles of single trials from the positions on the surface indicated by the arrows, before training (bottom), early in training (middle), and late in training (top). ***C***, Surface plot and single trials for TD learning with $\lambda = 0.9$ and $\alpha = 0.005$ (500 trials). Surface grid lines show every 10th trial. ***D***, Population data from DA cell recordings. The same data from Figure 3*A* (animals with little training) and Figure 3*B* (animals with more extensive training) have been replotted as line graphs, after normalizing for different firing rate by converting to modulation index (see Materials and Methods) and smoothing by a three-step running average. The bottom plot shows responses to unpredicted rewards of both early and late training groups overlaid. The middle plot shows data from cells recorded in animals in early training and top plot data from different cells recorded in animals late in training. It is clear that the cell data matches well the model profiles in ***C*** but not those generated by the parameters used in ***B***. Calibration bar, 500 ms.

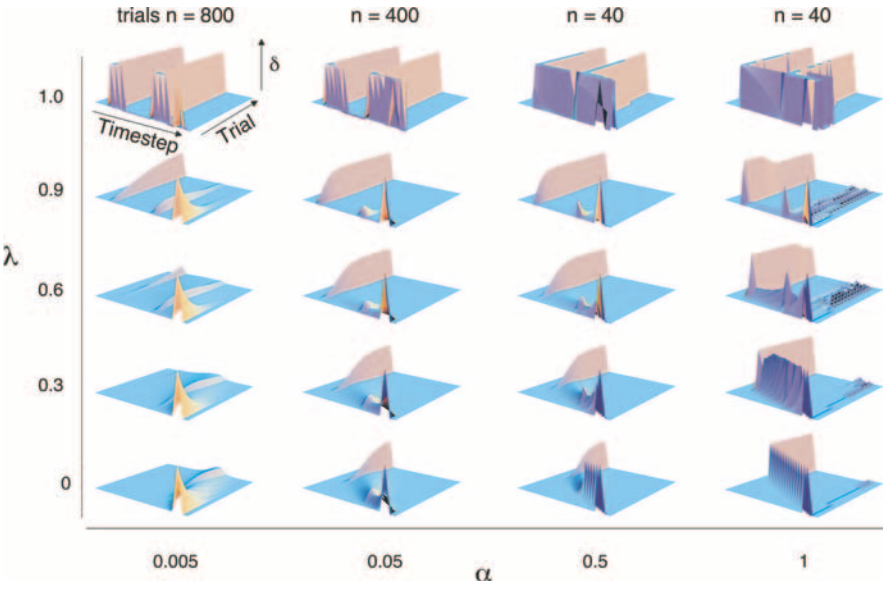

Figure 5. Exploration of the parameter space of the TD algorithm. Each 3-D surface plot shows changes in prediction error output over the course of conditioning for a different value of α and λ. Cues were delivered at time steps 5 and 15 and reward at time step 20. The number of trials shown in each plot (n) was varied for different settings of α, so that similar levels of learning were obtained by the end of the simulation.

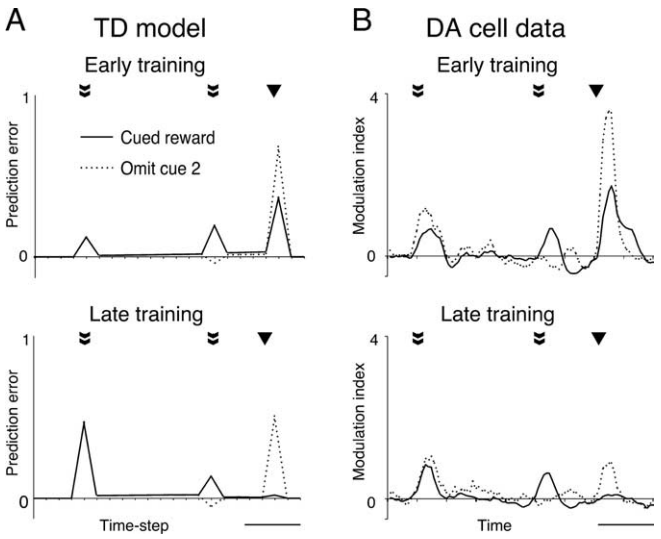

Figure 6. TD modeling of the response of DA cells to omission of an expected intermediate cue signal. ***A***, Prediction error outputs from successive single trials of TD model ($\lambda = 0.9$; $\alpha = 0.005$), in which the second cue was either present (solid lines) or omitted (dotted lines). The two lines completely overlap except at times of cue 2 and reward. The top graph shows trial from early in training (trials 100 and 101) and bottom graph from late in training (trials 400 and 401). The calibration bar indicates five time steps. ***B***, DA cell population data for the early and late training groups from Figure 3, normalized and smoothed as described in the legend for Figure 4. Solid lines show population histograms derived from cued-reward trials within the omission paradigm. Dotted lines show response to cue 2 omission within the same omission paradigm block for the same cells. Calibration, 500 ms.

However, when we investigated the effects of different settings for the parameters α and λ of the TD algorithm, we found that these apparent disparities between TD model prediction errors and DA cell activity are attributable to the particular choice of parameters used in the model. Thus, instead of posing a problem for TD models, the present results show that classical TD mechanisms can account for a wider range of observed DA cell behaviors than previously envisaged. This match of the classical TD model with physiology complements other modifications of the TD method (Daw et al., 2003; Nakahara et al., 2004), which can account for different aspects of neural activity and behavioral performance. It is important to note that our results should not be understood as providing explicit fixed values for neural TD learning parameters, because the specific values that produce the best fit to the data may be different if details of the model structure were changed. For instance, it is not known how sensory stimuli are represented in the brain for reward learning. We followed many previous studies in using a serial compound representation of conditioned stimuli. Alternative representations can be envisaged (Suri and Schultz, 1998, 1999), which might lead to different specific parameter values. Importantly, the present findings constrain TD approaches to brain function by establishing the biologically relevant boundaries for the parameter space that can be linked to physiologically plausible processes.

The parameter λ was of particular importance. Previous studies of the application of the TD algorithm to DA cell data have used TD(0) versions (in which $\lambda = 0$) to account for some aspects of DA cell responses (Montague et al., 1996; Schultz et al., 1997). In TD(0) models, learning occurs progressively across trials with the prediction error signal occurring one time step earlier on each trial until it "arrives" at the cue. Development of a prediction error response to an event k time steps in the past thus requires at least k trials. The advantage of TD(0) in machine learning contexts is that, when the number of trials is not limiting, it is computationally simpler to implement than the other variants (Sutton and Barto, 1998). However, it is the step-wise learning aspect of these models that leads to specific predicted patterns of activity that we failed to find in recorded DA cells. A physiological interpretation of TD(0) is that the system can only hold a memory of an event from one moment to the next. Our findings suggest that this limitation cannot apply to learning in the brain reward pathways.

In contrast, we were able to replicate the observed DA cell responses using TD(λ) models, in which $0 < \lambda < 1$. In TD(λ), vector weights representing a sensory signal remain eligible for modification by prediction error signals for a variable number of time steps after the signal has occurred. This eligibility trace allows bridging between events removed from one another in time within a single trial (Sutton and Barto, 1998). TD(λ) models have been shown in some machine-learning circumstances to be more efficient than TD(0) or TD(1) (Tesauro, 1992; Kaelbling et al., 1996; Sutton and Barto, 1998). Several formal algorithmic models of biological learning phenomena have incorporated eligibility traces (Barto and Sutton, 1982; Sutton, 1988; Sutton and Barto, 1990; Houk et al., 1995), following a suggestion by Klopf [cited in Klopf (1988)]. Eligibility traces have been used in TD models of neural circuits to accelerate learning, but the impact on the pattern of prediction error signaling by DA cells does not appear to have been explored in detail (Suri and Schultz, 1999, 2001; Suri, 2001). The concept that for delayed associative conditioning to occur there must be some memory or trace of the antecedent signal at the time of a subsequent reward is as old as the study of classical conditioning itself (Pavlov, 1927; Hull, 1943). Our finding of an excellent match between DA cell recordings and TD(λ) models strongly suggests that a process producing similar effects to a prolonged eligibility trace probably occurs in the circuits regulating DA cell activity in the brain.

The present finding that λ needs to be set at the high end of its range suggests that the proposed neural eligibility traces must last for a significant portion of the interval between trial events, which in the present study was in the order of seconds. The existence of eligibility traces lasting several seconds in neural circuits is open to experimental verification. Potential mechanisms include sustained firing in reverberating circuits and biochemical mechanisms acting at the synaptic level (Houk et al., 1995).

The parameter α sets the learning rate for weight changes. Low values for α slow the rate at which prediction error signals to cues and rewards develop or are lost. Thus, very low α in combination with high λ generates a prolonged period over which both cue and reward responses occur, as seen in the DA cell data. A physiological interpretation of this low learning rate is that plastic elements in the brain (analogous to the modifiable vector weights of the model) only change by a small proportion of the total dynamic range available to them on any one trial. This seems plausible, given that on each trial there is only a single conjunction between cue and prediction error.

A low learning rate may be considered a disadvantage. However, previous studies in machine learning have found that when λ is set high to enable learning to begin early in training, low values of α improve the stability of learning (Tesauro, 1992; Cichosz, 1995; Kaelbling et al., 1996; Singh and Sutton, 1996). Thus, combining prolonged eligibility for change with low rates for learning rate offers considerable practical advantages that may have provided significant selective pressure for the coevolution of equivalent parameter settings in the brain.

References

Aebischer P, Schultz W (1984) The activity of pars compacta neurons of the monkey substantia nigra is depressed by apomorphine. Neurosci Lett 50:25–29.

Aghajanian GK, Bunney BS (1977) Dopamine "autoreceptors": pharmacological characterization by microiontophoretic single cell recording studies. Naunyn Schmiedebergs Arch Pharmacol 297:1–7.

Barto AG (1995) Adaptive critics and the basal ganglia. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds), pp 215–232. Cambridge, MA: MIT.

Barto AG, Sutton RS (1982) Simulation of anticipatory responses in classical-conditioning by a neuron-like adaptive element. Behav Brain Res 4:221–235.

Bunney BS, Aghajanian GK, Roth RH (1973) Comparison of effects of L-dopa, amphetamine and apomorphine on firing rate of rat dopaminergic neurones. Nat New Biol 245:123–125.

Cichosz P (1995) Truncating temporal differences: on the efficient implementation of TD(λ) for reinforcement learning. J Artif Intell Res 2:287–318.

Daw ND, Courville AC, Touretzky DS (2003) Timing and partial observability in the dopamine system. In: Advances in neural information processing systems (Becker S, Thrun S, Obermayer K, eds), pp 99–106. Cambridge, MA: MIT.

Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299:1898–1902.

Grace AA, Bunney BS (1980) Nigral dopamine neurons: intracellular recording and identification with L-dopa injection and histofluorescence. Science 210:654–656.

Grace AA, Bunney BS (1983) Intracellular and extracellular electrophysiology of nigral dopaminergic neurons–1. Identification and characterization. Neuroscience 10:301–315.

Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309.

Houk JC, Adams JL, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds), pp 249–270. Cambridge, MA: MIT.

Hull CL (1943) Principles of behavior. New York: Appleton-Century.

Hyland BI, Reynolds JNJ, Hay J, Perk CG, Miller R (2002) Firing modes of midbrain dopamine cells in the freely moving rat. Neuroscience 114:475–492.

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4:237–285.

Kiyatkin EA, Rebec GV (2001) Impulse activity of ventral tegmental area neurons during heroin self-administration in rats. Neuroscience 102:565–580.

Klopf AH (1988) A neuronal model of classical conditioning. Psychobiology 16:85–125.

Kosobud AE, Harris GC, Chapin JK (1994) Behavioral associations of neuronal activity in the ventral tegmental area of the rat. J Neurosci 14:7117–7129.

Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol 67:145–163.

Miller JD, Sanghera MK, German DC (1981) Mesencephalic dopaminergic unit activity in the behaviorally conditioned rat. Life Sci 29:1255–1263.

Mirenowicz J, Schultz W (1994) Importance of unpredictability for reward responses in primate dopamine neurons. J Neurophysiol 72:1024–1027.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 43:133–143.

Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O (2004) Dopamine neurons can represent context-dependent prediction error. Neuron 41:269–280.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

Pavlov IP (1927) Conditioned reflexes. Oxford: Oxford UP.

Paxinos G, Watson C (1997) The rat brain in stereotaxic coordinates, Ed 3. London: Academic.

Rao RP, Sejnowski TJ (2001) Spike-timing-dependent Hebbian plasticity as temporal difference learning. Neural Comput 13:2221–2237.

Romo R, Schultz W (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J Neurophysiol 63:592–606.

Schultz W (1986) Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. J Neurophysiol 56:1439–1461.

Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80:1–27.

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241–263.

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. J Neurosci 13:900–913.

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

Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667.

Singh SP, Sutton RS (1996) Reinforcement learning with replacing eligibility traces. Mach Learn 22:123–158.

Suri RE (2001) Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. Exp Brain Res 140:234–240.

Suri RE, Schultz W (1998) Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp Brain Res 121:350–354.

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

Suri RE, Schultz W (2001) Temporal difference model reproduces anticipatory neural activity. Neural Comput 13:841–862.

Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3:9–44.

Sutton RS, Barto AG (1990) Time-derivative models of pavlovian reinforcement. In: Learning and computational neuroscience: foundations of adaptive networks (Gabriel M, Moore J, eds), pp 497–537. Cambridge, MA: MIT.

Sutton RS, Barto AG (1998) Reinforcement learning. Cambridge, MA: MIT.

Takikawa Y, Kawagoe R, Hikosaka O (2004) A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. J Neurophysiol 92:2520–2529.

Tesauro G (1992) Practical issues in temporal difference learning. Mach Learn 8:257–277.

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