LETTER Communicated by Andrew Barto

Long-Term Reward Prediction in TD Models of the Dopamine System

Nathaniel D. Daw
daw@cs.cmu.edu
David S. Touretzky
dst@cs.cmu.edu
Computer Science Department and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

This article addresses the relationship between long-term reward predictions and slow-timescale neural activity in temporal difference (TD) models of the dopamine system. Such models attempt to explain how the activity of dopamine (DA) neurons relates to errors in the prediction of future rewards. Previous models have been mostly restricted to short-term predictions of rewards expected during a single, somewhat artificially defined trial. Also, the models focused exclusively on the phasic pause-and-burst activity of primate DA neurons; the neurons' slower, tonic background activity was assumed to be constant. This has led to difficulty in explaining the results of neurochemical experiments that measure indications of DA release on a slow timescale, results that seem at first glance inconsistent with a reward prediction model. In this article, we investigate a TD model of DA activity modified so as to enable it to make longer-term predictions about rewards expected far in the future. We show that these predictions manifest themselves as slow changes in the baseline error signal, which we associate with tonic DA activity. Using this model, we make new predictions about the behavior of the DA system in a number of experimental situations. Some of these predictions suggest new computational explanations for previously puzzling data, such as indications from microdialysis studies of elevated DA activity triggered by aversive events.

1 Introduction

Recent neurophysiological and modeling work has suggested parallels between natural and artificial reinforcement learning (Montague, Dayan, & Sejnowski, 1996; Houk, Adams, & Barto, 1995; Schultz, Dayan, & Montague, 1997; Suri & Schultz, 1999). Activity recorded from primate dopamine (DA) neurons (Schultz, 1998), a system associated with motivation and addiction, appears qualitatively similar to the error signal from temporal difference (TD) learning (Sutton, 1988). This algorithm allows a system to learn

from experience to predict some measure of reward expected in the future (known as a return).

One aspect of the TD algorithm that has received little attention in the modeling literature is what sort of prediction the system is learning to make. TD algorithms have been devised for learning a variety of returns, differing, for example, as to the time frame over which the predicted rewards accumulate and whether rewards expected far in the future, if they are included at all, are discounted with respect to more immediate rewards. The DA models that have been the subject of simulation (Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1999) all assume for convenience a simple discrete-trial setting, with returns accumulating only over the course of a single trial. Although the modeled experiments have this structure, it is not clear whether or how animals segment a continuous stream of experience into a set of disjoint, repeating trials. And, as this article chiefly explores, the focus on predictions within a single trial eliminates subtler effects that a longer-timescale aspect to predictions would produce in the modeled DA signal.

A second weakness of current TD models of the DA signal is that they focus entirely on phasic activity. DA neurons exhibit tonic background firing punctuated by phasic bursts and pauses. Previous DA models explain the locations of the phasic events but treat the background firing rate as constant. Advocates of the TD models sometimes explicitly distinguish tonic DA as a separate phenomenon to which the models do not apply (Schultz, 1998). Nonetheless, experimental methods that measure DA activity over a very slow timescale have revealed effects that appear hard to reconcile with the TD models. Most critically, microdialysis studies have found evidence of elevated DA activity in response to aversive situations (see Horvitz, 2000, for a review). That such activity seems inconsistent with an error signal for reward prediction has fueled criticism of the TD models of DA (Redgrave, Prescott, & Gurney, 1999; Horvitz, 2000). Another slow-timescale effect that is not predicted by the TD models is the habituation, seen in voltammetry recordings, of DA release over the course of about a minute of unpredictable, rewarding brain stimulation (Kilpatrick, Rooney, Michael, & Wightman, 2000).

This article addresses both of these weaknesses of existing TD models together. We show that rather than being inapplicable to tonic DA, these models have all along contained the germ of an account of it in the relationship between slow-timescale DA release and long-term reward predictions. We study the tonic behavior of TD models modified only so far as necessary to allow them to predict a longer-term return that includes rewards expected in future trials. When the artificial horizon on the return is removed, the TD error includes a slowly changing background term, which we associate with tonic DA. These slow changes suggest a new computational explanation for the seemingly paradoxical data on DA responses to aversive events and for the slow habituation of DA release in brain stimulation experiments. We

also suggest a number of predictions about experiments that have not yet been performed. Finally, we offer some thoughts about what these computational considerations suggest about the functional anatomy of DA and related brain systems.

2 TD Models of the Dopamine System

2.1 Return Definitions for TD Models. TD learning is a reward prediction algorithm for Markov decision processes (MDPs; Sutton, 1988). In a series of discrete time steps, the process moves through a series of states, and the goal is to learn the value of these states. (In the context of modeling animal brains, time steps can be taken to correspond to intervals of real time and states to animals' internal representations of sensory observations.) Knowledge of values can, for instance, guide a separate action selection mechanism toward the most lucrative areas of the environment. More formally, systems use the TD algorithm to learn to estimate a value function $V(s(t))$, which maps the state at some instant, $s(t)$, to a measure of the expected reward that will be received in the future. We refer to the function's value at time t, $V(s(t))$, using the abbreviated notation $V(t)$. There are a number of common measures of future reward, known as returns, differing mainly in how rewards expected at different times are combined.

The simplest return, used in the DA models of Montague et al. (1996) and Schultz et al. (1997), is expected cumulative undiscounted reward,

$$V(t) = E\left[\sum_{\tau>t} r(\tau)\right], \tag{2.1}$$

where $r(\tau)$ is the reward received at time τ. This return makes sense only in finite horizon settings—those in which time is bounded. It is not informative to sum rewards over a period without end, since $V(t)$ can then be infinite. To usefully measure the value of a state in an infinite horizon problem, it is necessary either to modify the problem—by subdividing events into a series of time-bounded episodes, with returns accumulating only within these trials—or to introduce a different notion of value that takes into account predictions stretching arbitrarily far into the future.

The most common return that accommodates unbounded time discounts rewards exponentially in their delays; these discounted values can be summed over an infinite window without diverging:

$$V_{exp}(t) = E\left[\sum_{\tau>t} \gamma^{\tau-t} r(\tau)\right]. \tag{2.2}$$

The parameter $\gamma < 1$ controls the steepness of discounting.

The published TD models were all simulated using a return truncated on trial boundaries (Montague et al., 1996; Schultz et al., 1997; Suri & Schultz,

1999), except for the model of Houk et al. (1995), which included no simulations. The horizon obviated the need for discounting, and indeed most of the experiments and expositions used the undiscounted cumulative return, equation 2.1. All of the reports also mentioned that V_{exp} could be used to extend the models to a more realistic infinite-horizon situation, and Suri and Schultz (1999) used that return in their simulations. However, since even these investigations took place in an episodic setting, this work left unexamined how the behavior of the modeled DA system would be affected by longer-term predictions about rewards expected after the trial was complete. As we show here, the introduction of an infinite horizon gives rise to long-term predictions that would produce subtle variations in DA behavior, some previously puzzling indications of which have been observed in experiment.

To demonstrate this, we examine a model based on a different return that is closely related to equation 2.2 but makes the predictions we will be examining more obvious. This return assumes a constant baseline expectation of ρ reward per time step and measures the sum of differences between observed rewards and the baseline (Schwartz, 1993; Mahadevan, 1996; Tsitsiklis & Van Roy, 1999):

$$V_{rel}(t) = E\left[\sum_{\tau>t}(r(\tau) - \rho)\right]. \tag{2.3}$$

This relative value function is finite if ρ is taken to be the long-term future average reward per time step, $\lim_{n\to\infty}(1/n)E[\sum_{\tau=t}^{t+n-1} r(\tau)]$. This value ρ, the discrete analog of the long-term reward rate, is the same for all t under some assumptions about the structure of the MDP (Mahadevan, 1996).

For horizonless problems, V_{rel} (rather than the standard undiscounted return V) properly extends V_{exp} to the undiscounted case where $\gamma = 1$. Tsitsiklis and Van Roy (2002) prove that their TD algorithm for learning V_{rel} represents the limit as the discounting parameter γ approaches one in a version of exponentially discounted TD. This proof assumes V_{exp} is estimated with a linear function approximator, using a constant bias term in the discounted algorithm but not in the average reward one. Assuming function approximators that meet these constraints, there can thus be no in-principle difference between TD models based on lightly discounted V_{exp} and those based on V_{rel}, so approximate versions of the DA neuron behaviors we predict in section 3 are also expected under the exponentially discounted return in an infinite horizon setting. We use V_{rel} throughout this article to make it easier to examine the role of long-term predictions, since it segregates one component of such predictions in the ρ term. We return to the details of the relationship between the two models in the discussion.

2.2 Specifying a TD Model of Dopamine. In this article, we analyze the DA model of Montague et al. (1996), modified to use the TD rule of

Tsitsiklis and Van Roy (1999) to learn an estimate $\hat{V}_{rel}$ of V_{rel} instead of V. The TD algorithm defines an error signal by which a function approximator can be trained to learn the value function. The model uses a linear function approximator,

$$\hat{V}_{rel}(t) = \mathbf{w}(t) \cdot \mathbf{s}(t),$$

where $\mathbf{w}$ is a trainable weight vector and $\mathbf{s}$ is a state vector.

At each time step, weights are updated according to the delta rule,

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \nu \cdot \mathbf{s}(t) \cdot \delta_{rel}(t),$$

where ν is a learning rate and δ_{rel} is the average reward TD error:

$$\delta_{rel}(t) = \hat{V}_{rel}(t+1) - \hat{V}_{rel}(t) + r(t) - \rho(t). \tag{2.4}$$

In this equation, the average reward $\rho(t)$ is time dependent because it must be estimated on-line; we use an exponentially windowed running average with learning rate $\kappa \ll \nu$:

$$\rho(t+1) = \kappa r(t) + (1-\kappa)\rho(t).$$

Under the model, DA neurons are assumed to fire with a rate proportional to $\delta_{rel}(t) + b$ for some background rate b. Negative prediction error is thus taken to correspond to firing at a rate slower than b.

Finally, we must define the state vector $\mathbf{s}(t)$. The simple conditioning experiments we consider here involve at most a single discrete stimulus, whose timing provides all available information for predicting rewards. We can thus delineate all possible states of the tasks by how long ago the stimulus was last seen. We represent this interval using a tapped delay line. Specifically, we define the state vector's ith element $s_i(t)$ to be one if the stimulus was last seen at time $t - i$, and zero otherwise. This effectively reduces the linear function approximator to a table lookup representation in the complete state-space of the tasks. We do not use a constant bias term, which would be superfluous since the average reward TD algorithm learns V_{rel} only up to an arbitrary additive constant. For similar reasons, in the very simple experiments we discuss where rewards or punishments are delivered randomly without any stimuli, we leave $\mathbf{s}(t)$ empty, which means that all predictions learned in this context reside in $\rho(t)$ rather than $\mathbf{w}(t)$. (This is reasonable since reward delivery in these tasks can be viewed as controlled by a one-state Markov process whose V_{rel} is arbitrary.)

Representation of state and timing is one of the more empirically underconstrained aspects of DA system models, and we have elsewhere suggested an alternative scheme based on semi-Markov processes (Daw, Courville, & Touretzky, 2002). We use the tapped delay line architecture here for its simplicity, generality, and consistency with previous models. In particular, this is essentially the same stimulus representation used by Montague et al.

(1996), following the suggestion of Sutton and Barto (1990). The one important difference between the representational scheme we use here and that of Montague et al. is in the length of $\mathbf{s}(t)$. In their experiments, this vector was shorter than the interval between trials (which they took to be constant). The effect was that at some point during the intertrial interval, $\mathbf{s}(t)$ became a vector of all zeros, and so the prediction $\hat{V}(t) = \mathbf{w}(t) \cdot \mathbf{s}(t)$ was also necessarily zero until the next stimulus occurred. This prevented predictions of rewards in subsequent trials from backing up over the intertrial interval, effectively enforcing a horizon of a single trial on the learned value function. (Backups might still have occurred had the model contained either a constant bias term in $\mathbf{s}(t)$ or eligibility traces in the learning rule, but it did not.) In the Montague et al. (1996) model, then, the learned value function had a finite horizon as a side effect of the state representation and learning rule being used, even though it was exposed to a continuous series of trials. In our model, we instead take $\mathbf{s}(t)$ to be long enough that it never empties out between trials. This requires us to randomize the interval between trials (which was done in most studies of DA neurons but not in previous models), since otherwise, stimulus delivery would itself be predictable from events in the previous trial, and a reward-predicting stimulus would not, after learning, continue to evoke prediction error. In fact, supporting our view that predictions should encompass more than a single trial, in highly overtrained monkeys performing a task with minimal variability in the intertrial interval, DA neurons ceased responding to reward-predicting stimuli (Ljunbgerg, Apicella, & Schultz, 1992).

3 Results

We first verify that the average reward TD version of the DA model we examine here preserves the ability of previous models to capture the basic phasic response properties of DA neurons. Figure 1 shows the modeled DA signal (as the points of the discrete signal connected by lines) for three trials during the acquisition and extinction of a stimulus-reward association, reproducing some key qualitative properties of phasic DA neuron responses (Schultz, 1998). The results are not visibly different from those reported by Montague et al. (1996) and others, which is to be expected given the close relationship of the models.

In the rest of this section, we examine how long-term reward predictions would affect DA system behavior, by finding differences between the undiscounted TD model (which is useful only in an episodic setting and was the basis of most previous investigations) and the average reward TD model (which learns to predict an infinite-horizon return; we show later that all of the same properties are expected under an exponentially discounted TD model). The undiscounted TD error signal, used by Montague et al. (1996) to model the DA response, is $\delta(t) = V(t+1) - V(t) + r(t)$. This differs from

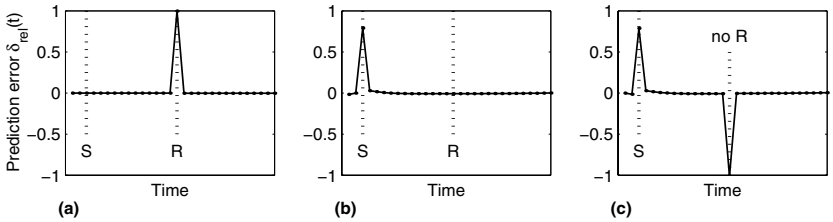

Figure 1: Error signal $\delta_{rel}(t)$ from an average reward TD model, displaying basic phasic properties of DA neurons. S: stimulus, R: reward. (a) Before learning, activation occurs in response to a reward but not to the stimulus that precedes it. (b) After learning, activation transfers to the time of the stimulus, and there is no response to the reward. (c) When a predicted reward is omitted, negative prediction error, corresponding to a pause in background firing, occurs at the expected time of the reward.

δ_{rel} (see equation 2.4) only in that $\rho(t)$ is subtracted from the latter. The extra effect of learning this infinite horizon return is to depress the TD error signal by the slowly changing long-term reward estimate $\rho(t)$. Manifestations of this effect would be most noticeable when $\rho(t)$ is large or changing, neither of which is true of the experiment modeled above.

The average reward TD simulations pictured in Figures 2 through 4 demonstrate the predicted behavior of the DA response in situations where the effect of $\rho(t)$ should be manifest. Figure 2 depicts the basic effect, a decrease in the DA baseline as the experienced reward rate increases. Empirically, DA neurons respond phasically to randomly delivered, unsignaled rewards; the figure shows how the tonic inhibitory effect of $\rho(t)$ on the error signal should reduce DA activity when such rewards are delivered

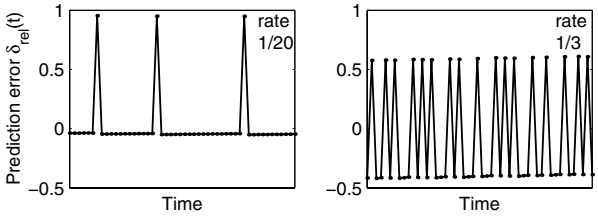

Figure 2: Tonic effects in modeled DA signal. Increasing the rate of randomly delivered, unsignaled rewards should depress baseline neuronal activity, an effect that is also visible as a lower peak response to rewards. (Left) Modeled DA response to rewards delivered at a low rate. (Right) Modeled DA response to rewards delivered at a higher rate, showing depressed activity. These snapshots are from the end of a long sequence of rewards delivered at the respective rates.

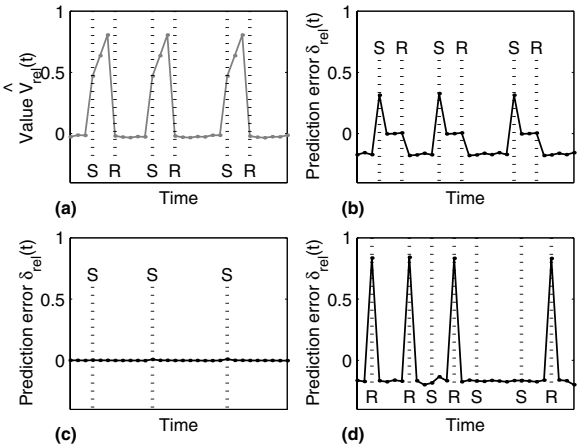

Figure 3: Tonic effects in modeled DA signal after conditioning and extinction. S: stimulus, R: reward. (a) After conditioning, value prediction $\hat{V}_{rel}(t)$ jumps when the stimulus is received, ramps up further in anticipation of reward, and drops back to baseline when the reward is received. (b) The corresponding TD error signal—the modeled DA response—shows the classic phasic burst to the stimulus and is then zero until the reward is received. Between trials, the error signal is negative, corresponding to a predicted tonic depression in DA firing below its normal background rate. (c) After extinction by repeated presentation of stimuli alone, the tonic depression disappears (along with the phasic responses) and the error signal is zero everywhere. (d) If extinction is instead accomplished by presenting unpaired rewards and stimuli, the error signal remains tonically depressed to a negative baseline between phasic responses to the unpredicted rewards.

at a fast Poisson rate compared to a slow one. If the experienced reward rate changes suddenly, the estimate $\rho(t)$ will slowly adapt to match it, and DA activity will smoothly adapt accordingly; we show the system's stable behavior when this adaptation asymptotes. (Note that although the prediction parameters have reached asymptotic values, the rewards themselves are unpredictable, and so they continue to evoke phasic prediction error.) In Figure 2, the inhibition resulting from a high average reward estimate $\rho(t)$ can be seen in both the level of tonic background firing between rewards and the absolute peak magnitude of the phasic responses to rewards. Below, we discuss some evidence for the latter seen in voltammetry experiments measuring DA concentrations in the nucleus accumbens.

A more complex effect emerges when randomly delivered rewards are signaled by a stimulus preceding them by a constant interval. In this case, the phasic DA response is known from experiment to transfer to the predictive stimulus. In addition, according to the average reward TD model, the

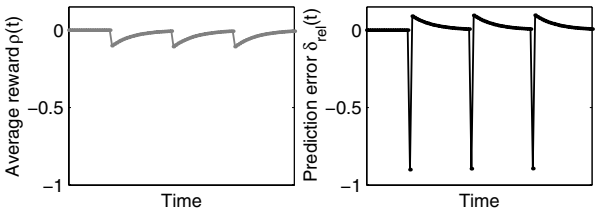

Figure 4: Tonic effects in modeled DA signal in an aversive situation. Receipt of primary punishment (such as shock) decreases the average reward rate estimate, producing a small but long-lasting elevation in modeled DA activity. (Left) The reward rate estimate $\rho(t)$ responding to three punishments. (Right) The corresponding TD error signal—the modeled DA response—shows a phasic depression followed by a prolonged excitation.

inhibitory effect of $\rho(t)$ should manifest itself as before in a sustained negative prediction error (i.e., a reduced baseline firing rate) between trials (see Figure 3b). During the fixed interval between stimulus onset and reward, the learning algorithm can eliminate this prediction error by smoothly ramping up $\hat{V}_{rel}(t)$ (see Figure 3a), so that $\hat{V}_{rel}(t+1) - \hat{V}_{rel}(t)$ exactly cancels the $-\rho(t)$ term in equation 2.4 (see Figure 3b). But outside the stimulus-reward interval, the algorithm cannot predict events, since the interval between trials is random, so it cannot similarly compensate for the negative prediction error due to $\rho(t)$. Hence, $\rho(t)$ drives $\delta_{rel}(t)$ slightly negative between trials. The predicted pattern of results—a jump in the DA baseline on stimulus presentation, followed by a drop when the reward is received—would be noticeable only when the rate of reward delivery is high enough to make its inhibitory effect between trials visible. For instance, it cannot be seen in Figure 1b, which was produced using the same task but at a much lower reward rate. This may be why, to our knowledge, it has not so far been observed in experiments.

Another experimental situation involving a change of reward rate is the extinction of a stimulus-reward association. Previous TD models suggest that when a stimulus-reward association is extinguished by repeatedly presenting the stimulus but omitting the reward, the phasic DA response to the formerly predictive stimulus should wane and disappear. The concomitant reduction of the experienced reward rate means that under the average reward model, the inhibition of the DA baseline by $\rho(t)$ should also extinguish, slowly increasing the baseline firing rate between trials until it matches that within trials (see Figure 3c). Were extinction instead accomplished while preserving the reward rate—by presenting both stimuli and rewards, but unpaired—the reinforcers, now unpredictable, would trigger phasic excitation, but the tonic inhibitory effect of $\rho(t)$ would remain (see Figure 3d).

A final prediction about DA activity in an infinite horizon model emerges if we treat aversive stimuli such as shock as equivalent to negative reward. This is an oversimplification, but some behavioral evidence that it is nonetheless reasonable comes from animal learning experiments in which a stimulus that predicts the absence of an otherwise expected reward (a conditioned inhibitor for reward) can, when subsequently presented together with a neutral stimulus and a shock, block the learning of an association between the neutral stimulus and the aversive outcome (Dickinson & Dearing, 1979; Goodman & Fowler, 1983). This is presumably because the conditioned inhibitor, predictive of a different but still aversive outcome, is assumed to account for the shock. If aversive stimuli are treated as negative rewards, their appearance will decrease the predicted reward rate $\rho(t)$, causing a prolonged increase in the tonic DA baseline. This effect is shown in Figure 4. In the simulations, phasic negative error is also seen to the negative rewards. As we discuss below, there is some experimental evidence for both sorts of responses.

4 Discussion

Previous models of the DA system were exercised in a fairly stylized setting, where events were partitioned into a set of independent trials and the system learned to predict rewards only within a trial. Though it is computationally straightforward to extend these models to a more realistic setting with longer-term predictions and no such partitions, and indeed the original articles all suggested how this could be done, no previous work has examined how such an improvement to the model would affect its predictions about the behavior of the DA system. By using a version of the TD algorithm that explicitly segregates long-term average reward predictions, we show here that such an extension suggests several new predictions about slow timescale changes in DA activity. Some of these predictions have never been tested, while others explain existing data that had been puzzling under older trial-based models.

In our model, the long-term reward predictions that are learned if returns are allowed to accumulate across trial boundaries would manifest themselves largely as slow changes in the tonic DA baseline. Though little attention has been paid to tonic firing rates in the electrophysiological studies that the DA models address—and so the predictions we suggest here have for the most part not been addressed in this literature—a related experimental technique may offer some clues. Microdialysis studies examine chemical evidence of DA activity in samples taken very slowly, on the order of once per 10 minutes. It is likely that at least part of what these experiments measure is tonic rather than phasic activity.

A key mystery from the microdialysis literature that has been used to criticize TD models of DA is evidence of increased DA activity in target brain areas, such as the striatum, in response to aversive events such as foot

shocks (reviewed by Horvitz, 2000). Were this activity phasic, it would contradict the notion of DA as an error signal for the prediction of reward. But we show here that the slower tonic component of the modeled DA signal should indeed increase in the face of aversive stimuli, because they should reduce the average reward prediction ρ (see Figure 4). This may explain the paradox of an excitatory response to aversive events in a putatively appetitive signal. In the model as presented here, to be fair, the small and slow positive error would on average be offset by a larger but shorter duration burst of negative error, timelocked to the aversive stimulus. However, what microdialysis actually measured in such a situation would depend on such vagaries as the nonlinearities of transmitter release and reuptake. Also, instead of being fully signaled by DA pauses, phasic information about aversive events may partially be carried by excitation in a parallel, opponent neural system (Kakade, Daw, & Dayan, 2000; Daw, Kakade, & Dayan, in press). All of these ideas about the nature of the dopaminergic response to aversive stimuli would ideally be tested with unit recordings of DA neurons in aversive paradigms; the very sparse and problematic set of such recordings so far available reveals some DA neurons with excitatory and some with inhibitory responses (Schultz & Romo, 1987; Mirenowicz & Schultz, 1996; Guarraci & Kapp, 1999). In accord with our explanation, there is some evidence that the excitatory responses tend to last longer (Schultz & Romo, 1987).

Another set of experiments used a much faster voltammetry technique to measure DA release in the striatum with subsecond resolution, in response to bursts of rewarding intracranial stimulation (Kilpatrick et al., 2000). In these experiments, no phasic DA release is measured when stimulation is triggered by the animal's own lever presses, but DA responses are seen to (presumably unpredictable) stimulation bouts delivered on the schedule of lever presses made previously by another animal. In this case, transmitter release habituates over time, gradually reducing the magnitude of the phasic DA responses to stimulations over the course of a train lasting about a minute. This is essentially a dynamic version of the experiment shown in Figure 2, where a sudden increase in the delivered reward rate gives rise to a gradual increase in $\rho(t)$ and a corresponding decrease in modeled DA activity. Under a TD model involving only short-term predictions, it would be difficult to find any computational account for this habituation, but it is roughly consistent with what we predict from the average-reward model. However, the effect seen in these experiments seems to be mainly identifiable with a decline in the peak phasic response to the rewards (which is only part of what is predicted in Figure 2); the corresponding decline in tonic baseline DA levels, if any, is exceedingly small. This may also have to do with the nonlinearity of transmitter release, or with a floor effect in the measurements or in extracellular DA concentrations.

The interpretation of this experiment points to an important and totally unresolved issue of timescale in the model presented here. Voltam-

metry experiments reveal information about DA activity on an intermediate timescale—slower than electrophysiology but faster than microdialysis. Voltammetry is inappropriate for monitoring the kind of long-term changes in basal DA concentrations measured by microdialysis; perhaps for this reason, the two methodologies give rather different pictures of DA activity during intracranial stimulation, with microdialysis suggesting a very long-lasting excitation (Kilpatrick et al., 2000; Fiorino, Coury, Fibiger, & Phillips, 1993). We have argued that understanding DA responses to aversive stimuli requires distinguishing between phasic and tonic timescales, but in reality there are surely many timescales important to the DA system and at which the system might show distinct behaviors. The model we have presented here is too simple to cope with such a proliferation of timescales; it works by contrasting predictions learned at only two fixed timescales: quickly adapting phasic reward predictions that vary from state to state, against a slowly changing average reward prediction that is the same everywhere. The speed at which the average reward estimate $\rho(t)$ adapts controls the timescale of tonic effects in the model; it is unclear to exactly what real-world timescale this should correspond. A more realistic model might learn not just a single "long-term" average reward prediction, but instead predictions of future events across a spectrum of different timescales. We speculate that in such a model, the opponent interactions between fast and slow predictions that we study in this article could survive as competition to account for the same observations between predictions with many different characteristic timescales.

The rigidity of timescale in our model (and other TD models) also makes it an unsatisfying account of animal learning behavior, since animals seem much more flexible about timescales. Not only can they cope with events happening on timescales ranging from seconds to days, but their learning processes, as exhibited behaviorally, are often invariant to a very wide range of dilations or contractions of the speed of events (Gallistel & Gibbon, 2000). These considerations again suggest a multiscale model.

The idea that animals are learning to predict a horizonless return rather than just the rewards in the current trial has also appeared in the behavioral literature. Kacelnik (1997) uses this idea to explain animals' choices on experiments designed to measure their discounting of delayed rewards (Mazur, 1987). Under Kacelnik's model, however, it is necessary to assume that animals neglect the intervals between trials in computing their long-term expectations, an assumption that would be difficult to incorporate convincingly in a TD model like the one presented here. Results like this suggest that the notion of a trial is not as behaviorally meaningless as we treat it here. More work on this point is needed.

We have presented all of our results in terms of the average reward TD algorithm, because it is easiest to see how long- and short-term predictions interact in this context. However, with some caveats, all of these effects would also be seen in the more common exponentially discounted formu-

lation of TD learning, which was used for at least one previously published DA model (Suri & Schultz, 1999). This will be true if the algorithm is used in an infinite horizon setting and with discounting light enough to allow rewards expected in future trials to contribute materially to value estimates. We briefly sketch here why this holds asymptotically (i.e., when the value functions are well learned, so that $\hat{V} = V$) using a table representation for the value function; Tsitsiklis and Van Roy (2000) give a more sophisticated proof that the algorithms are equivalent on an update-by-update basis in the limit as $\gamma \to 1$, assuming linear value function approximation and that the state vector in the discounted case contains a constant bias element whose magnitude is chosen so that its learned weight plays a role similar to that of ρ in the average reward model.

The key point is that with an infinite horizon, V_{exp} grows as the reward rate increases, since it sums all discounted future rewards; a portion of the value of each state thus encodes the long-term future reward rate, weighted by the degree of discounting. In particular, the exponentially discounted value function V_{exp} can be viewed as the sum of a state-dependent term that approximates V_{rel} (better as γ increases) and a state-independent baseline magnitude $\rho/(1-\gamma)$ that encodes the long-term future reward rate expected everywhere. All of the DA behaviors we consider depend on ρ being subtracted in the error signal $\delta_{rel}(t)$. Though this term seems to be missing from the exponentially discounted error signal $\delta_{exp}(t) = \gamma \hat{V}_{exp}(t+1) - \hat{V}_{exp}(t) + r(t)$, it is actually implicit: the state-independent portions of $\hat{V}_{exp}(t)$ and $\hat{V}_{exp}(t+1)$ are equal to $\rho/(1-\gamma)$, so the state-independent portion of $\gamma \hat{V}_{exp}(t+1) - \hat{V}_{exp}(t)$ in the error signal is equal to $(\gamma - 1) \cdot \rho/(1-\gamma)$, which is $-\rho$.

This is easiest to see when rewards are Poisson with rate ρ, as in Figure 2. In this case, $V_{exp}(t)$ is exactly $\rho/(1-\gamma)$ for all t. The exponentially discounted TD error signal is

$$\begin{aligned}\delta_{exp}(t) &= r(t) + \gamma \hat{V}_{exp}(t+1) - \hat{V}_{exp}(t) \\ &= r(t) + (\gamma - 1)\hat{V}_{exp}(t) \\ &= r(t) - \rho,\end{aligned}$$

just as in the average reward case. While this example holds for any γ, in more structured state-spaces, the value of γ can matter. For the predicted DA behaviors described in this article, the effect of steep discounting would be to modulate the tonic inhibition depending on proximity to reward; the correspondence between discounted and average reward models improves smoothly as γ increases.

In summary, the introduction of an infinite horizon increases the value of V_{exp} at all states, leading to the subtraction of approximately ρ from the error signal and to just the sort of tonic behaviors we have discussed in this

article. Note that the equivalence between the models depends on the exponentially discounted model having a sufficient state representation; as we have discussed, the model of Montague et al. (1996) effectively had a finite horizon due to gaps in its state representation and the lack of a bias term.

One significant model of DA neurons that is physiologically more detailed than the TD models is that of Brown, Bullock, and Grossberg (1999). Though this is an implementational model and not grounded in reinforcement learning theory, it actually has much the same computational structure as a TD model. In particular, the DA response is modeled as the sum of a phasic primary reward signal like $r(t)$ and the positively rectified time derivative of a striatal reward anticipation signal that, like $\hat{V}(t)$, shows sustained elevation between the occurrence of a stimulus and the reward it predicts. The chief structural elaboration of this model over the TD models is that it separates into a distinct pathway the transient inhibitory effects on DA that result, in TD models, from the negatively rectified portion of the time difference $\hat{V}(t+1) - \hat{V}(t)$. This distinction vanishes at the more algorithmic level of analysis of this article, but the Brown et al. model provides a detailed proposal for how the brain may implement TD-like computations.

The average reward TD model presented here would require additional physiological substrates. In particular, the DA system would need the average reward estimate $\rho(t)$ to compute the error signal. One candidate substrate for this signal is serotonin (5HT), a neuromodulator that seems to act as an opponent to DA. Kakade et al. (2000) and Daw et al. (in press) detail this proposal under the assumption that 5HT reports $\rho(t)$ directly to DA target structures rather than (as we envision here) channeling this portion of the error signal through DA. Since 5HT is known to oppose DA not only at the level of target structures but also by inhibiting DA firing directly (reviewed by Kapur & Remington, 1996), it seems likely that any effects of the average reward would be visible in the DA system as well, as we predict here.

Alternatively, the average reward could be computed entirely within the DA system, through activity habituation mechanisms in DA neurons. In a habituation model, some variable internal to DA neurons, such as calcium concentration or autoreceptor activation, could track $\rho(t)$ (by accumulating with neuronal firing) and inhibit further spiking or transmitter release. Intriguingly, if the inhibition worked at the level of transmitter release, the effects of the average reward signal predicted in this article might be visible only using methods that directly measured transmitter release rather than spiking. Such is the case for the two pieces of experimental evidence so far best consistent with our predictions: dialysis studies showing elevated DA activity in aversive situations (Horvitz, 2000) and voltammetry measurements of DA release habituating to high rates of intracranial stimulation (Kilpatrick et al., 2000).

These considerations about how the DA system computes the hypothesized error signal could be tested by direct comparison of voltammetric

and electrophysiological measurements of DA activity. More generally, our modeling suggests that greater attention should be paid to DA activity over slower timescales, in electrophysiological as well as neurochemical recordings. We particularly suggest that this would be useful in experiments that manipulate the rates of delivery of reward or punishment since these, in the model, control tonic DA release.

Acknowledgments

This research was supported by NSF IRI-9720350, IIS-9978403, DGE-9987588, and an NSF Graduate Fellowship. The authors particularly thank Peter Dayan, and also Rudolf Cardinal, Barry Everitt, and Sham Kakade for many helpful discussions and comments on this work. We also thank two anonymous reviewers for copious and substantive comments that helped shape the final article.

References

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience, 19*(23), 10502–10511.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2002). Dopamine and inference about timing. In *Proceedings of the Second International Conference on Development and Learning* (pp. 271–276). Los Alamitos, CA: IEEE Computer Society.

Daw, N. D., Kakade, S., & Dayan, P. (in press). Opponent interactions between serotonin and dopamine. *Neural Networks*.

Dickinson, A., & Dearing, M. F. (1979). Appetitive-aversive interactions and inhibitory processes. In A. Dickonson & R. A. Boakes (Eds.), *Mechanisms of learning and motivation* (pp. 203–231). Hillsdale, NJ: Erlbaum.

Fiorino, D. F., Coury, A., Fibiger, H. C., & Phillips, A. G. (1993). Electrical stimulation of reward sites in the ventral tegmental area increases dopamine transmission in the nucleus accumbens of the rat. *Behavioral Brain Research, 55*, 131–141.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review, 107*(2), 289–344.

Goodman, J., & Fowler, H. (1983). Blocking and enhancement of fear conditioning by appetitive CSs. *Animal Learning and Behavior, 11*, 75–82.

Guarraci, F., & Kapp, B. (1999). An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential Pavlovian fear conditioning in the awake rabbit. *Behavioral Brain Research, 99*, 169–179.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience, 96*, 651–656.

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.

Kacelnik, A. (1997). Normative and descriptive models of decision making: Time discounting and risk sensitivity. In G. R. Bock & G. Cardew (Eds.), *Characterizing human psychological adaptations* (pp. 51–70). New York: Wiley.

Kakade, S., Daw, N. D., & Dayan, P. (2000). Opponent interactions between serotonin and dopamine for classical and operant conditioning. *Society for Neuroscience Abstracts, 26*, 1763.

Kapur, S., & Remington, G. (1996). Serotonin-dopamine interaction and its relevance to schizophrenia. *American Journal of Psychiatry, 153*, 466–476.

Kilpatrick, M. R., Rooney, M. B., Michael, D. J., & Wightman, R. M. (2000). Extracellular dopamine dynamics in rat caudate-putamen during experimenter-delivered and intracranial self-stimulation. *Neuroscience, 96*, 697–706.

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

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms and empirical results. *Machine Learning, 22*, 1–38.

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior* (Vol. 5, pp. 55–73). Hillsdale, NJ: Erlbaum.

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

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short latency dopamine burst too short to signal reinforcement error? *Trends in Neurosciences, 22*, 146–151.

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

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

Schultz, W., & Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology, 57*, 201–217.

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

Suri, R. E., & Schultz, W. (1999). A neural network 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 method of temporal differences. *Machine Learning, 3*, 9–44.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Tsitsiklis, J. N., & Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica, 35*, 319–349.

Tsitsiklis, J. N., & Van Roy, B. (2002). On average versus discounted reward temporal-difference learning. *Machine Learning, 49*, 179–191.

Received June 30, 2000; accepted April 30, 2002.