LETTER Communicated by Mark Ungless

Representation and Timing in Theories of the Dopamine System

Nathaniel D. Daw
daw@gatsby.ucl.ac.uk
UCL, Gatsby Computational Neuroscience Unit, London, WC1N3AR, U.K.

Aaron C. Courville
aaronc@cs.cmu.edu
Carnegie Mellon University, Robotics Institute and Center for the Neural Basis of Cognition, Pittsburgh, PA 15213, U.S.A.

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

Although the responses of dopamine neurons in the primate midbrain are well characterized as carrying a temporal difference (TD) error signal for reward prediction, existing theories do not offer a credible account of how the brain keeps track of past sensory events that may be relevant to predicting future reward. Empirically, these shortcomings of previous theories are particularly evident in their account of experiments in which animals were exposed to variation in the timing of events. The original theories mispredicted the results of such experiments due to their use of a representational device called a tapped delay line.

Here we propose that a richer understanding of history representation and a better account of these experiments can be given by considering TD algorithms for a formal setting that incorporates two features not originally considered in theories of the dopaminergic response: partial observability (a distinction between the animal's sensory experience and the true underlying state of the world) and semi-Markov dynamics (an explicit account of variation in the intervals between events). The new theory situates the dopaminergic system in a richer functional and anatomical context, since it assumes (in accord with recent computational theories of cortex) that problems of partial observability and stimulus history are solved in sensory cortex using statistical modeling and inference and that the TD system predicts reward using the results of this inference rather than raw sensory data. It also accounts for a range of experimental data, including the experiments involving programmed temporal variability and other previously unmodeled dopaminergic response phenomena,

which we suggest are related to subjective noise in animals' interval timing. Finally, it offers new experimental predictions and a rich theoretical framework for designing future experiments.

1 Introduction

The responses of dopamine neurons in the primate midbrain are well characterized by a temporal difference (TD) (Sutton, 1988) reinforcement learning (RL) theory, in which neuronal spiking is supposed to signal error in the prediction of future reward (Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). Although such theories have been influential, a key computational issue remains: How does the brain keep track of those sensory events that are relevant to predicting future reward, when the rewards and their predictors may be separated by long temporal intervals?

The problem traces to a disconnect between the physical world and the abstract formalism underlying TD learning. The formalism is the Markov process, a model world that proceeds stochastically through a series of states, sometimes delivering reward. The TD algorithm learns to map each state to a prediction about the reward expected in the future. This is possible because, in a Markov process, future states and rewards are conditionally independent of past events, given only the current state. There is thus no need to remember past events: the current state contains all information relevant to prediction. This assumption is problematic when it comes to explaining experiments on dopamine neurons, which often involve delayed contingencies. In a typical experiment, a monkey learns that a transient flash of light signals that, after a 1-second delay, a drop of juice will be delivered. Because of this temporal gap, the animal's immediate sensory experiences (gap, flash, gap, juice) do not by themselves correspond to the states of a Markov process. This example also demonstrates that these issues of memory are tied up with issues of timing—in this case, marking the passage of the 1 second interval.

Existing TD theories of the dopamine system address these issues using variations on a device called a *tapped delay line* (Sutton & Barto, 1990), which redefines the state to include a buffer of previous sensory events within some time window. If the window is large enough to encompass relevant history, which is assumed in the dopamine theories, then the augmented states form a Markov process, and TD learning can succeed. Clearly, this approach fudges an issue of selection: How can the brain adaptively decide which events should be remembered, and for how long? In practice, the tapped delay line is also an awkward representation for predicting events whose timing can vary. As a result, the theory incorrectly predicted the firing of dopamine neurons in experiments in which the timing of events was

varied (Hollerman & Schultz, 1998; Fiorillo & Schultz, 2001). This problem has received only limited attention (Suri & Schultz, 1998, 1999; Daw, 2003).

In this letter, we take a deeper look at these issues by adopting a more appropriate formalism for the experimental situation. In particular, we propose modeling the dopamine response using a TD algorithm for a partially observable semi-Markov process (also known as a hidden semi-Markov model), which generalizes the Markov process in two ways. This richer formalism incorporates variability in the timing between events (semi-Markov dynamics; Bradtke & Duff, 1995) and a distinction between the sensory experience and the underlying but only partially observable state (Kaelbling, Littmann, & Cassandra, 1998). The established theory of RL with partial observability offers an elegant approach to maintaining relevant sensory history. The idea is to use Bayesian inference, with a statistical description ("world model") of how the hidden process evolves, to infer a probability distribution over the likely values of the unobservable state. If the world model is correct, this inferred state distribution incorporates all relevant history (Kaelbling et al., 1998) and can itself be used in place of the delay line as a state representation for TD learning.

Applied to theories of the dopamine system, this viewpoint casts new light on a number of issues. The system is viewed as making predictions using an inferred state representation rather than raw sensory history. This reframes the problem of representing adequate stimulus history in the computationally more familiar terms of learning an appropriate world model. It also situates the dopamine neurons in a broader anatomical and functional context, since predominant models of sensory cortex envision it performing the sort of world modeling and hidden state inference we require (Doya, 1999; Rao, Olshausen, & Lewicki, 2002). Combined with a number of additional assumptions (notably, about the relative strength of positive and negative error representation in the dopamine response; Niv, Duff, & Dayan, 2005; Bayer & Glimcher, 2005), the new model accounts for puzzling results on the responses of dopamine neurons when event timing is varied; further, armed with this account of temporal variability, we consider the effect of noise in internal timing processes and show that this can address other experimental phenomena. Previous models can be viewed as approximations to the new one under appropriate limits.

The rest of the letter is organized as follows. In section 2 we discuss previous models and how they cope with temporal variability. We realize our own account of the system in several stages. We begin in section 3 with a general overview of the pieces of the model. In section 4, we develop and simulate a fully observable semi-Markov TD model, and in the following section we generalize it to the partially observable case. As a limiting case of the complete, partially observable model, the simpler model is appropriate for analyzing the complete model's behavior in many situations. After presenting results about the behavior of each model, we discuss to what

extent its predictions are upheld experimentally. Finally, in section 6, we conclude with more general discussion.

2 Previous Models

In this section, we review the TD algorithm and its use in models of the dopamine response, focusing on the example of a key, problematic experiment.

These models address unit recordings of dopamine neurons in primates performing a variety of appetitive conditioning tasks (for review, see Schultz, 1998). These experiments can largely be viewed as variations on Pavlovian trace conditioning, a procedure in which a transient cue such as a flash of light is followed after some interval by a reinforcer, regardless of the animal's actions. In fact, some of the experiments were conducted using delay conditioning (in which the initial stimulus is not punctate but rather prolonged to span the gap between stimulus onset and reward) or involved instrumental requirements (that is, the cue signaled the monkey to perform some behavioral response such as a key press to obtain a reward). For most of the data considered here, no notable differences in dopamine behavior have been observed between these methodological variations, to the extent that comparable experiments have been done. Thus, here, and in common with much prior modeling work on the system, we will neglect action selection and stimulus persistence and idealize the tasks as Pavlovian trace conditioning.

2.1 The TD Algorithm. The TD theory of the dopamine response (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997) involves modeling the experimental situation as a Markov process and drawing on the TD algorithm for reward prediction in such processes (Sutton, 1988). Such a process comprises a set $\mathcal{S}$ of states, a transition function T, and a reward function R. The process has discrete dynamics; at each time step t, a real-valued reward $\mathbf{r}_t$ and a successor state $\mathbf{s}_{t+1} \in \mathcal{S}$ are drawn. The distributions over $\mathbf{r}_t$ and $\mathbf{s}_{t+1}$ are specified by the functions T and R and depend on only the value of the current state $\mathbf{s}_t$. In modeling the experimental situation, the process time steps are taken to correspond to short, constant-length blocks of real time, and the state corresponds to some representation, internal to the animal, of relevant experimental stimuli.

We can define a value function mapping states to expected cumulative discounted future reward,

$$V_s \equiv E\left[\sum_{\tau=t}^{t_{end}} \gamma^{\tau-t}\mathbf{r}_\tau \mid \mathbf{s}_t = s\right], \tag{2.1}$$

where the expectation is taken with respect to stochasticity in the state transitions and reward magnitudes, t_{end} is the time the current trial ends, and γ is a parameter controlling the steepness of temporal discounting.

The goal of the TD algorithm is to use samples of states $\mathbf{s}_t$ and rewards $\mathbf{r}_t$ to learn an approximation $\hat{V}$ to the true value function V. If such an estimate were correct, it would satisfy

$$\hat{V}_{\mathbf{s}_t} = E[\mathbf{r}_t + \gamma \hat{V}_{\mathbf{s}_{t+1}} \mid \mathbf{s}_t], \tag{2.2}$$

which is just the value function definition rewritten recursively. The TD learning rule is based on this relation: given a sample of a pair of adjacent states and an intervening reward, the TD algorithm nudges the estimate $\hat{V}_{\mathbf{s}_t}$ toward $\mathbf{r}_t + \gamma \hat{V}_{\mathbf{s}_{t+1}}$. The change in $\hat{V}_{\mathbf{s}_t}$ is thus proportional to the TD error,

$$\delta_t = \mathbf{r}_t + \gamma \hat{V}_{\mathbf{s}_{t+1}} - \hat{V}_{\mathbf{s}_t}, \tag{2.3}$$

with values updated as $\hat{V}_{\mathbf{s}_t} \leftarrow \hat{V}_{\mathbf{s}_t} + \nu \cdot \delta_t$ for learning rate ν. In this article, we omit consideration of *eligibility traces*, as appear in the TD-λ algorithm (Sutton, 1988; Houk et al., 1995; Sutton & Barto, 1998). These would allow error at time t directly to affect states encountered some time steps before, an elaboration that can speed up learning but does not affect our general argument.

2.2 TD Models of the Dopamine Response.

2.2.1 Model Specification. The TD models of the dopamine response assume that dopamine neurons fire at a rate proportional to the prediction error δ_t added to some constant background activity level, so that positive δ_t corresponds to neuronal excitation and negative δ_t to inhibition. They differ in details of the value function definition (e.g., whether discounting is used) and how the state is represented as a function of the experimental stimuli. Here we roughly follow the formulation of Montague et al. (1996; Schultz et al., 1997), on which most subsequent work has built.

The state is taken to represent both current and previous stimuli, represented using tapped delay lines (Sutton & Barto, 1990). Specifically, assuming the task involves only a single stimulus, the state $\mathbf{s}_t$ is defined as a binary vector, whose ith element is one if the stimulus was last seen at time $t - i$ and zero otherwise. For multiple stimuli, the representation is the concatenation of several such history vectors, one for each stimulus. Importantly, reward delivery is not represented with its own delay line. In fact, reward delivery is assumed to have no effect on the state representation. As illustrated in Figure 1, stimulus delivery sets off a cascade of internal states, whose progression, once per time step, tracks the time since stimulus

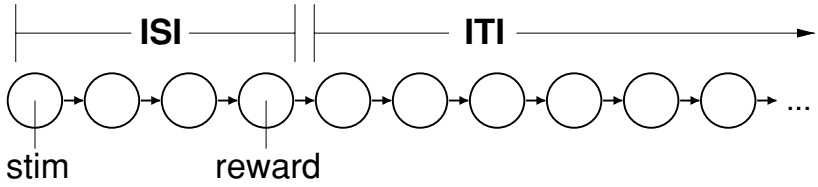

Figure 1: The state space for a tapped delay line model of a trace conditioning experiment. The stimulus initiates a cascade of states that mark time relative to it. If the interstimulus interval is deterministic, the reward falls in one such state. ISI: interstimulus interval; ITI: intertrial interval.

delivery. These time steps are taken to correspond to constant slices of real time, of duration perhaps 100 ms.

The value is estimated linearly as the dot product of the state vector with a weight vector: $\hat{V}_{\mathbf{s}_t} = \mathbf{s}_t \cdot w_t$. For the case of a single stimulus, this is equivalent to a table maintaining a separate value for each of the marker states shown in Figure 1.

2.2.2 Account for Basic Findings. The inclusion of the tapped delay line enables the model to mimic basic dopamine responses in trace conditioning (Montague et al., 1996; Schultz et al., 1997). Dopamine neurons burst to unexpected rewards or reward-predicting signals, when δ_t is positive, and pause when an anticipated reward is omitted (and δ_t is negative). The latter is a timed response and occurs in the model because value is elevated in the marker states intervening between stimulus and reward. If reward is omitted, the difference $\gamma \hat{V}_{\mathbf{s}_{t+1}} - \hat{V}_{\mathbf{s}_t}$ in equation 2.3 is negative at the state where the reward was expected, so negative error is seen when that state is encountered without reward.

2.2.3 Event Time Variability. This account fails to predict the response of dopamine neurons when the timing of rewards is varied from trial to trial (Hollerman & Schultz, 1998; see also Fiorillo & Schultz, 2001). Figure 2 (left) shows the simulated response when a reward is expected 1 second after the stimulus but instead delivered 0.5 second early (top) or late (bottom) in occasional probe trials. The noteworthy case is what follows early reward. Experiments (Hollerman & Schultz, 1998) show that the neurons burst to the early reward but do not subsequently pause at the time reward was originally expected. In contrast, because reward arrival does not affect the model's stimulus representation, the delay line will still subsequently arrive in the state in which reward is usually received. There, when the reward is not delivered again, the error signal will be negative, predicting (contrary to experiment) a pause in dopamine cell firing at the time the reward was originally expected.

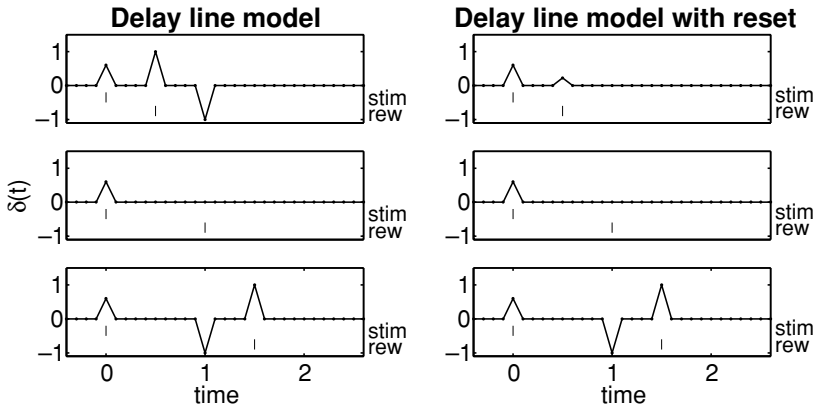

Figure 2: TD error (simulated dopamine response) in two tapped delay line models ($\gamma = 0.95$) of a conditioning task in which reward was delivered early (top) or late (bottom) in occasional probe trials. (Middle: reward delivered at normal time.) (Left) In the standard model, positive error is seen to rewards arriving at unexpected times and negative error is seen on probe trials at the time reward had been originally expected. (Right) In a modified version of the model in which reward resets the delay line, no negative error is seen following an early reward.

It might seem that this problem could be solved simply by adding a second delay line to represent the time since reward delivery, so that the model could learn not to expect a second reward after an early one. However, in the experiment discussed here, the model might not have had the opportunity for such learning. Since early rewards were delivered only in occasional probe trials, value predictions were presumably determined by experience with the reward occurring at its normal time. Further, even given extensive experience with early rewards, two tapped delay lines plus a linear value function estimator could never learn the appropriate discrimination, because (as is easy to verify) the expected future value at different points in the task is a nonlinear function of the two-delay-line state representation.

A number of authors have proposed fixing this misprediction by assuming that reward delivery resets the representational system, for example, by clearing the delay line representing time since the stimulus (Suri & Schultz, 1998, 1999; Brown, Bullock, & Grossberg, 1999). This operation negates all pending predictions and avoids negative TD error when they fail to play out. Figure 2 (right) verifies that this device eliminates the spurious inhibition after an early reward. However, it is unclear from the original work under what circumstances such a reset is justified or appropriate, and doubtful that this simple, ad hoc rule generalizes properly to other situations. We return to these considerations in the discussion. Here, we investigate a more

systematic approach to temporal representation in such experiments, based on the view of stimulus history taken in work on partially observable Markov processes (Kaelbling et al., 1998). On this view, a (in principle, learnable) world model is used to determine relevant stimulus history. Therefore, we outline an appropriate family of generative models for reinforcer and stimulus delivery in conditioning tasks: the partially observable semi-Markov process.

3 A New Model: A Broad Functional Framework

In this article, we specify a new TD model of the dopamine system incorporating semi-Markov dynamics and partial observability. Our theory envisions the dopaminergic value learning system as part of a more extensive framework of interacting learning systems than had been previously considered in dopaminergic theories. Figure 3 lays out the pieces of the model; those implemented in this article are shown in black.

The idea of the theory is to address prediction in the face of partial observability by using a statistical model of the world's contingencies to infer a probability distribution over the world's (unobservable) state and then to use this inferred representation as a basis for learning to predict values using a TD algorithm. Thus, we have:

- A **model learning** system that learns a forward model of state transitions, state dwell times, and observable events. Similar functions are often ascribed to cortical areas, particularly prefrontal cortex (Owen, 1997; Daw, Niv, & Dayan, 2005).

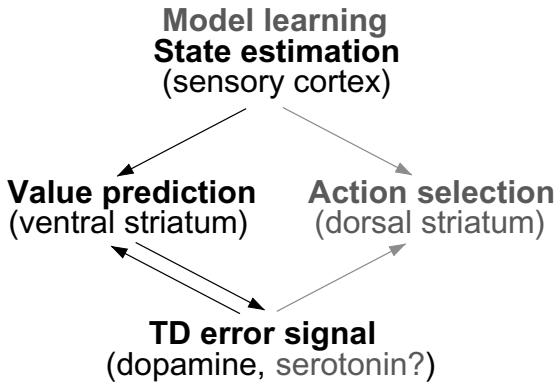

Figure 3: Schematic of the interaction between multiple learning systems suggested in this article. Those discussed in detail here are shown in black.

- A **state estimation** system that infers the world's state (and related latent variables) using sensory observations and the world model. This broadly corresponds to cortical sensory processing systems (Rao et al., 2002).
- A **value prediction** system that uses a TD error signal to map this inferred state representation to a prediction of future reward. This portion of the system works similarly to previous TD models of the dopamine system, except that we assume semi-Markov rather than Markov state transition dynamics. We associate this aspect of the model with the dopamine neurons and their targets (Schultz, 1998). Additionally, as discussed below, information about negative errors or aversive events, which may be missing from the dopaminergic error signal, could be provided by other systems such as serotonin (Daw, Kakade, & Dayan, 2002).

In this article, since we are studying asymptotic dopamine responding, we assume the correct model has already been learned and do not explicitly perform model learning (though we have studied it elsewhere in the context of theories of conditioning behavior; Courville & Touretzky, 2001; Courville, Daw, Gordon, & Touretzky, 2003; Courville, Daw, & Touretzky, 2004). Model fitting can be performed by variations on the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977); a version for hidden semi-Markov models is presented by Guedon and Cocozza-Thivent (1990). Here we would require an online version of these methods (as in Courville & Touretzky, 2001).

RL is ultimately concerned with action selection in Markov decision processes, and it is widely assumed that the dopamine system is involved in control as well as prediction. In RL approaches such as actor-critic (Sutton, 1984), value prediction in a Markov process (as studied here) is a subproblem useful for learning action selection policies. Hence we assume that there is also:

- An **action selection** system that uses information from the TD error (or perhaps the learned value function) to learn an action selection policy. Traditionally, this is associated with the dopamine system's targets in the dorsal striatum (O'Doherty et al., 2004).

As we are focused on dopamine responses in Pavlovian tasks, we do not address policy learning in this article.

4 A New Model: Semi-Markov Dynamics

We build our model in two stages, starting with a model that incorporates semi-Markov dynamics but not partial observability. This simplified model is useful for both motivating the description of the complete model and studying its behavior, since the simplified model is easier to analyze and

represents a good approximation to the complete model's behavior under certain conditions.

4.1 A Fully Observable Semi-Markov Model. A first step toward addressing issues of temporal variability in dopamine experiments is to adopt a formalism that explicitly models such variability. Here we generalize the TD models presented in section 2 to use TD in a semi-Markov process, which adds temporal variability in the state transitions.

In a semi-Markov process, state transitions occur as in a Markov process, except that they occur irregularly. The dwell time for each state visit is randomly drawn from a distribution associated with the state. In addition to transition and reward functions (T and R), semi-Markov models contain a function D specifying the dwell time distribution for each state. The process is known as semi-Markov because although the identities of successor states obey the Markov conditional independence property, the probability of a transition at a particular instant depends not just on the current state but on the time that has already been spent there. We model rewards and stimuli as instantaneous events occurring on the transition into a state.

We require additional notation. It can at times be useful to index random variables either by their time t or by a discrete index k that counts state transitions. The time spent in state $\mathbf{s}_k$ is $\mathbf{d}_k$, drawn conditional on $\mathbf{s}_k$ from the distribution specified by the function D. If the system entered that state at time τ, delivering reward $\mathbf{r}_k$, then we can also write that $\mathbf{s}_t = \mathbf{s}_k$ for all $\tau \leq t < \tau + \mathbf{d}_k$ and $\mathbf{r}_t = \mathbf{r}_k$ for $t = \tau$ while $\mathbf{r}_t = 0$ for $\tau < t < \tau + \mathbf{d}_k$.

It is straightforward to adapt standard reinforcement learning algorithms to this richer formal framework, a task first tackled by Bradtke and Duff (1995). Our formulation is closer to that of Mahadevan, Marchalleck, Das, & Gosavi (1997; Das, Gosavi, Mahadevan, & Marchalleck, 1999). We use the value function

$$\hat{V}_{\mathbf{s}_k} = E[\mathbf{r}_{k+1} - \rho \mathbf{d}_k + \hat{V}_{\mathbf{s}_{k+1}} \mid \mathbf{s}_k], \tag{4.1}$$

where the expectation is now taken additionally with respect to randomness in the dwell time $\mathbf{d}_k$. There are two further changes to the formulation here. First, for bookkeeping purposes, we omit the reward $\mathbf{r}_k$ received on entering state $\mathbf{s}_k$ from that state's value. Second, in place of the exponentially discounted value function of equation 2.2, we use an average reward formulation, in which $\rho \equiv lim_{n\to\infty}(1/n) \cdot \sum_{\tau=t}^{t+n-1} \mathbf{r}_\tau$ is the average reward per time step. This represents a limit of the exponentially discounted case as the discounting factor $\gamma \to 1$ (for details, see Tsitsiklis & Van Roy, 2002) and has some useful properties for modeling dopamine responses (Daw & Touretzky, 2002; Daw et al., 2002). Following that work, we will henceforth assume that the value function is infinite horizon, that is, when written in

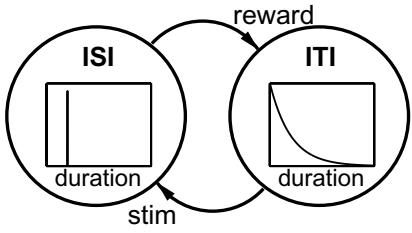

Figure 4: The state space for a semi-Markov model of a trace conditioning experiment. States model intervals of time between events, which vary according to the distributions sketched in the insets. Stimulus and reward are delivered on state transitions. ISI: interstimulus interval; ITI: intertrial interval. Here, the ISI is constant, while the ITI is drawn from an exponential distribution.

the unrolled form of equation 2.1 as a sum of rewards, the sum does not terminate on a trial boundary but rather continues indefinitely.

In TD for semi-Markov processes (Bradtke & Duff, 1995; Mahadevan et al., 1997; Das et al., 1999), value updates occur irregularly, whenever there is a state transition. The error signal is

$$\delta_k = \mathbf{r}_{k+1} - \rho_k \cdot \mathbf{d}_k + \widehat{V}_{\mathbf{s}_{k+1}} - \widehat{V}_{\mathbf{s}_k}, \tag{4.2}$$

where ρ_k is now subscripted because it must be estimated separately (e.g., by the average reward over the last n states, $\rho_k = \sum_{k'=k-n+1}^{k+1} \mathbf{r}_{k'} / \sum_{k'=k-n}^{k} \mathbf{d}_{k'}$)

4.2 Connecting This Theory to the Dopamine Response. Here we discuss a number of issues related to simulating the dopamine response with the algorithm described in the previous section.

4.2.1 State Representation. To connect equation 4.2 to the firing of dopamine neurons, we must relate the states $\mathbf{s}_k$ to the observable events. In the present, fully observable case, we take them to correspond one to one. In this model, a trace conditioning experiment has a very simple structure, consisting of two states that capture the intervals of time between events (see Figure 4; compare Figure 1). The CS is delivered on entry into the state labeled ISI (for interstimulus interval), while the reward is delivered on entering the ITI (intertrial interval) state. This formalism is convenient for reasoning about situations in which interevent intervals can vary, since such variability is built into the model. Late or early rewards, for instance, just correspond to longer or shorter times spent in the ISI state.

Although the model assumes input from a separate timing mechanism—in order to measure the elapsed interval $\mathbf{d}_k$ between events used in the update equation—the passage of time does not by itself have any effect on

the modeled dopamine signal. Instead, TD error is triggered only by state transitions, which are here taken to be always signaled by external events. Thus, this simple scheme cannot account for the finding that dopamine neurons pause when reward is omitted (Schultz et al., 1997). (It would instead assume the world remains in the ISI state with zero TD error until another event occurs, signaling a state transition and triggering learning.) In section 5, we handle these cases by using the assumption of partial observability to infer a state transition in the absence of a signal; however, when states are signaled reliably, that model will reduce to this one. We thus investigate this model in the context of experiments not involving reward omission.

4.2.2 Scaling of Negative Error. Because the background firing rates of dopamine neurons are low, excitatory responses have a much larger magnitude (measured by spike count) than inhibitory responses thought to represent the same absolute prediction error (Niv, Duff, et al., 2005). Recent work quantitatively comparing the firing rate to estimated prediction error confirms this observation and suggests that the dopamine response to negative error is rescaled or partially rectified (Bayer & Glimcher, 2005; Fiorillo, Tobler, & Schultz, 2003; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004).

This fact can be important when mean firing rates are computed by averaging dopamine responses over trials containing both positive and negative prediction errors, since the negative errors will be underrepresented (Niv, Duff, et al., 2005). To account for this situation, we assume that dopaminergic firing rates are proportional to $\delta + \psi$, positively rectified, where ψ is a small background firing rate. We average this rectified quantity over trials to simulate the dopamine response. The pattern of direct proportionality with rectification beneath a small level of negative error is consistent with the experimental results of Bayer and Glimcher (2005).

Note that we assume that values are updated based on the complete error signal, with the missing information about negative errors reported separately to targets (perhaps by serotonin; Daw et al., 2002). An alternative possibility (Niv, Duff, et al., 2005) is that complete negative error information is present in the dopaminergic signal, though scaled differently, and targets are properly able to decode such a signal. There are as yet limited data to support or distinguish among these suggestions, but the difference is not material to our argument here. This article explores the implications for dopaminergic recordings of the asymmetry between bursts and pauses. Such asymmetry is empirically well demonstrated and distinct from speculation as to how dopamine targets might cope with it.

4.2.3 Interval Measurement Noise. We will in some cases consider the effects of internal timing noise on the modeled dopamine signal. In the model, time measurements enter the error signal calculation through the estimated dwell time durations $\mathbf{d}_k$. Following behavioral studies (Gibbon,

1977), we assume that for a constant true duration, these vary from trial to trial with a standard deviation proportional to the length of the true interval. We take the noise to be normally distributed.

4.3 Results: Simulated Dopamine Responses in the Model. Here we present simulations demonstrating the behavior of the model in various conditions. We consider unsignaled and signaled reward and the effect of externally imposed or subjective variability in the timing of events. Finally, we discuss experimental evidence relating to the model's predictions.

4.3.1 Results: Free Reward Delivery. The simplest experimental finding about dopamine neurons is that they burst when animals receive random, unsignaled reward (Schultz, 1998). The semi-Markov model's explanation for this effect is different from the usual TD account.

This "free reward" experiment can be modeled as a semi-Markov process with a single state (see Figure 5, bottom right). Assuming Poisson delivery of rewards with magnitude r, mean rate λ, and mean interreward interval $\theta = 1/\lambda$, the dwell times are exponentially distributed. We examine the TD error, using equation 4.2. The state's value $\widehat{\mathbf{V}}$ is arbitrary (since it only appears subtracted from itself in the error signal) and $\rho = r/\theta$ asymptotically. The TD error on receiving a reward of magnitude r after a delay d is thus

$$\delta = r - \rho d + \widehat{\mathbf{V}} - \widehat{\mathbf{V}} \tag{4.3}$$

$$= r(1 - d/\theta), \tag{4.4}$$

which is positive if $d < \theta$ and negative if $d > \theta$, as illustrated in Figure 5 (left). That is, the TD error is relative to the expected delay θ: rewards occurring earlier than usual have higher value than expected, and conversely for later-than-average rewards.

Figure 5 (right top) confirms that the semi-Markov TD error averaged over multiple trials is zero. However, due to the partial rectification of inhibitory responses, excitation dominates in the average over trials of the simulated dopamine response (see Figure 5, right middle), and net phasic excitation is predicted.

4.3.2 Results: Signaled Reward and Timing Noise. When a reward is reliably signaled by a stimulus that precedes it, dopaminergic responses famously transfer from the reward to the stimulus (Schultz, 1998). The corresponding semi-Markov model is the two-state model of Figures 4 and 6a. (We assume the stimulus timing is randomized.)

As in free-reward tasks, the model predicts that the single-trial response to an event can vary from positive to negative depending on the interval preceding it, and, if sufficiently variable, the response averaged over trials will skew excitatory due to partial rectification of negative errors.

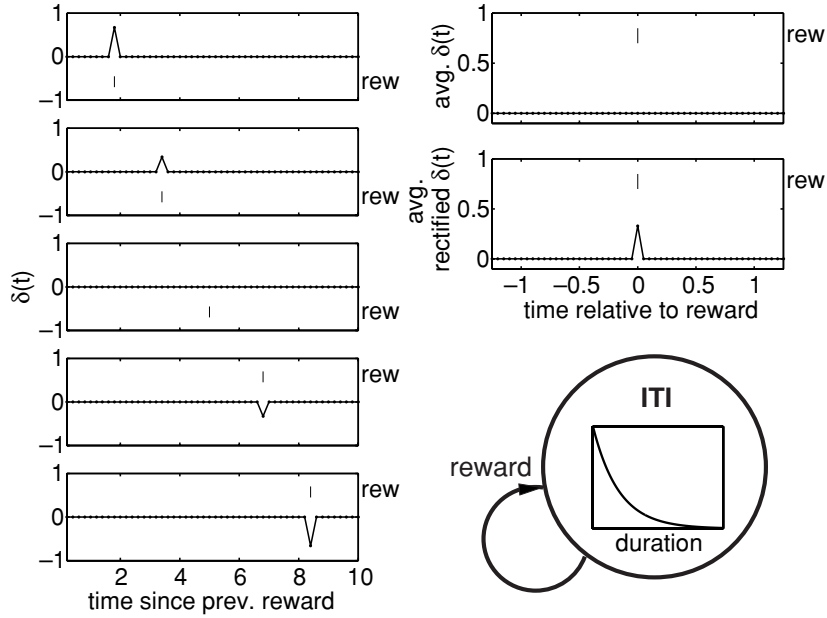

Figure 5: TD error to rewards delivered freely at Poisson intervals, using the semi-Markov TD model of equation 4.2. The state space consists of a single state (illustrated bottom right), with reward delivered on entry. (Left) Error in a single trial ranges from strongly positive through zero to strongly negative (top to bottom), depending on the time since the previous reward. Traces are aligned with respect to the previous reward. (Right) Error averaged over trials, aligned on the current reward. Right top: Mean TD error over trials is zero. Right middle: Mean TD error over trials with negative errors partially rectified (simulated dopamine signal) is positive. Mean interreward interval: 5 sec; reward magnitude: 1; rectification threshold: -0.1.

Analogous to rewards, this is evident here also for cues, whose occurrence (but not timing) in this task is signaled by the reward in the previous trial. As shown in Figure 6a, the model thus predicts the transfer of the (trial-averaged) response from the wholly predictable reward to the temporally unpredictable stimulus. (Single-trial cue responses covary with the preceding intertrial interval in a manner exactly analogous to reward responses in Figure 5 and are not illustrated separately.)

Variability in the stimulus-reward interval has analogous effects. If the stimulus-reward interval is jittered slightly (see Figure 6b), there is no effect on the average simulated dopamine response. This is because, in contrast to the situations considered thus far, minor temporal jitter produces only small negative and positive prediction errors, which fail to reach the threshold

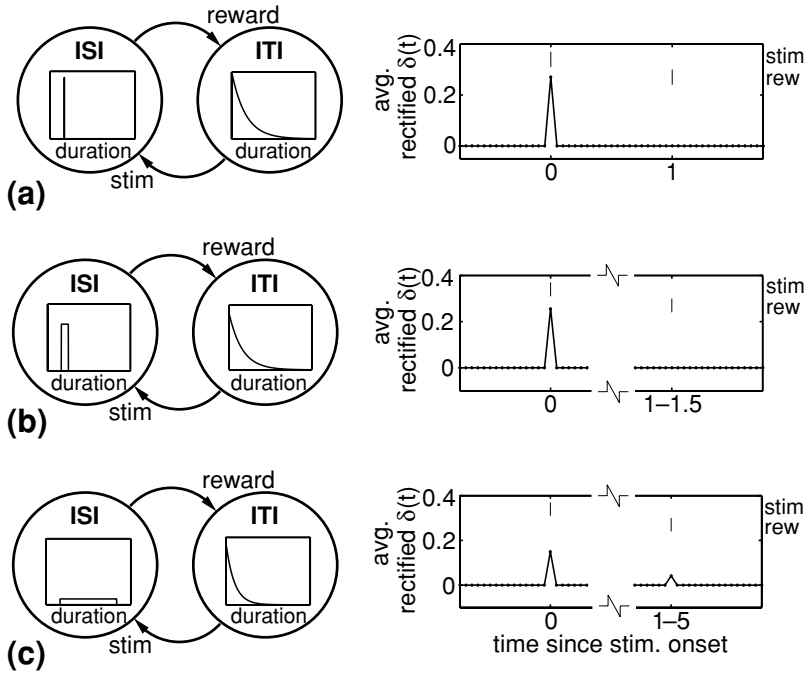

Figure 6: Semi-Markov model of experiments in which reward delivery is signaled. Tasks are illustrated as semi-Markov state spaces next to the corresponding simulated dopaminergic response. When stimulus-reward interval is (a) deterministic or (b) only slightly variable, excitation is seen to stimulus but not reward. (c) When the stimulus-reward interval varies appreciably, excitation is seen in the trial-averaged reward response as well. (ITI changed between conditions to match average trial lengths.) (a) Mean ITI: 5 sec, ISI: 1 sec; (b) Mean ITI: 4.75 secs, ISI: 1–1.5 secs uniform, (c) Mean ITI: 3 secs, ISI: 1–5 secs uniform; reward magnitude: 1; rectification threshold: –0.1.

of rectification and thus cancel each other out in the average. But if the variability is substantial, then a response is seen on average (see Figure 6c), because large variations in the interstimulus interval produce large positive and negative variations in the single-trial prediction error, exceeding the rectification threshold. Responding is broken out separately by delay in Figure 7. In general, the extent to which rectification biases the average dopaminergic response to be excitatory depends on how often, and by how much, negative TD error exceeds the rectification threshold. This in turn depends on the amount of jitter in the term $-\rho_k \cdot \boldsymbol{d}_k$ in equation 4.2, with larger average rewards ρ and more sizable jitter in the interreward intervals d promoting a net excitatory response.

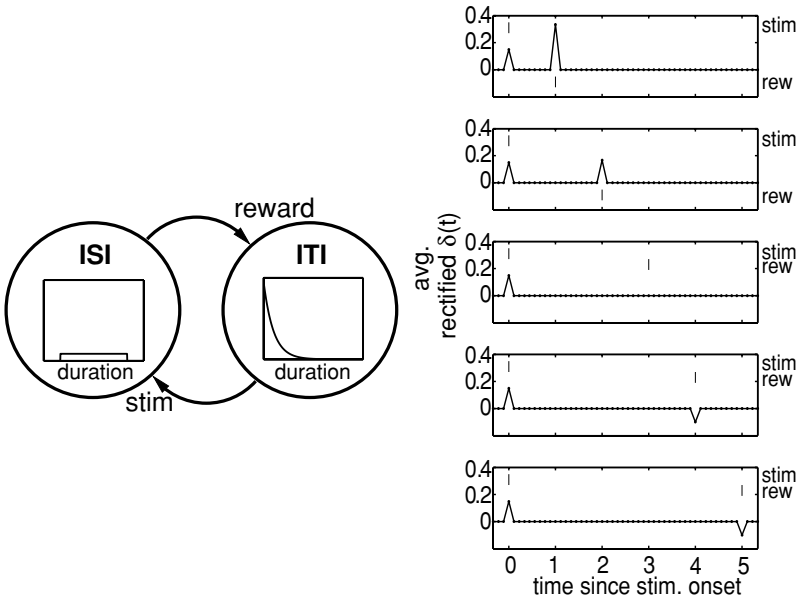

Figure 7: Semi-Markov TD error to rewards occurring at different delays after a stimulus (earlier to later, top to bottom); same task as Figure 6c. (Left) Task illustrated as a semi-Markov state space; rewards arrive at random, uniformly distributed intervals after the stimulus. (Right) Model predicts a decline in reward response with delay, with inhibition for rewards later than average. Parameters as in Figure 6c.

Finally, a parallel effect can be seen when we consider the additional effect of subjective time measurement noise. We repeat the conditioning experiment with a deterministic programmed ITI but add variability due to modeled subjective noise in timing. Figure 8 demonstrates that this noise has negligible effect when the delay between stimulus and reward is small, but for a longer delay, the modeled dopaminergic response to the reward reemerges and cannot be trained away. This is because timing noise has a constant coefficient of variation (Gibbon, 1977) and is thus more substantial for longer delays.

4.4 Discussion: Data Bearing on These Predictions. We have shown that the semi-Markov model explains dopaminergic responses to temporally unpredictable free rewards and reward predictors (Schultz, 1998). Compared to previous models, this account offers a new and testably different pattern of explanation for these excitatory responses. On stimulus or reward receipt, the "cost" $-\rho_k \cdot \boldsymbol{d}_k$ of the delay preceding it is subtracted from the phasic dopamine signal (see equation 4.2). Because of this

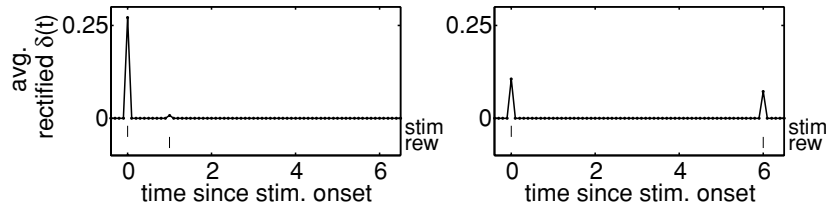

Figure 8: Effect of timing noise on modeled dopaminergic response to signaled reward depends on the interval between stimulus and reward (ISI). (Left) For ISI = 1 sec, error variation due to timing noise is unaffected by rectification and response to reward is minimal. (Right) For ISI = 6 sec, error variation from timing noise exceeds rectification level, and response to reward emerges. Mean ITI: 3 sec; reward magnitude: 1; rectification threshold: −0.1.; coefficient of variation of timing noise: 0.5.

subtraction, the model predicts that the single-trial phasic dopamine response should decline as the intertrial or interstimulus interval preceding it increases. This prediction accords with results (albeit published only in abstract form) of Fiorillo and Schultz (2001) from a conditioning experiment in which the stimulus-reward interval varied uniformly between 1 and 3 seconds. The theory further predicts that the response to later-than-average rewards should actually become negative; the available data are ambiguous on this point.[1] However, suggestively, all published dopamine neuron recordings exhibit noticeable trial-to-trial variability in excitatory responses (e.g., to temporally unpredictable stimuli), including many trials with a response at or below baseline. The suggestion that this variability reflects the preceding interevent interval could be tested with reanalysis of the data.

These phenomena are not predicted by the original tapped delay line model. This is because, unlike the semi-Markov model, it assesses the cost of a delay not all together in the phasic response to an event (so that variability in the delay impacts on the event response) but instead gradually during the interval preceding it, on the passage through each marker state. (In particular, at each time step, the error includes a term of $-\rho$ in the average reward formulation, or in the exponentially discounted version a related penalty arising from discounting; Daw & Touretzky, 2002.) On that account, rewards or reward predictors arriving on a Poisson or uniformly distributed random schedule should generally excite neurons regardless of their timing,

[1] The publicly presented data from Fiorillo and Schultz (2001) include spike counts for different stimulus-reward delays, supporting the conclusion that the mean response never extends below baseline. However, the accompanying spike rasters suggest that this conclusion may depend on the length of the time window over which the spike counts are taken.

and the ubiquitous response variability must be attributed to unmodeled factors.

The new theory also predicts that dopamine neurons should not respond when small amounts of temporal jitter precede an event, but that an excitatory response should emerge, on average, when the temporal variability is increased. This is consistent with the available data. In the experiment discussed above involving 1 to 3 second interstimulus intervals, Fiorillo and Schultz (2001) report net excitation to reward. Additionally, in an experiment involving a sequence of two stimuli predicting reward, neurons were excited by the second stimulus only when its timing varied (Schultz, Apicella, & Ljungberg, 1993). There is also evidence for tolerance of small levels of variability. In an early study (Ljungberg, Apicella, & Schultz, 1992), dopamine neurons did not respond to rewards ("no task" condition) or stimuli ("overtrained" condition) whose timing varied somewhat. Finally, the model predicts similar effects of subjective timing noise, and unpublished data support the model's prediction that it should be impossible to train away the dopaminergic response to a reward whose timing is deterministically signaled by a sufficiently distant stimulus (C. Fiorillo, personal communication, 2002).

Thus, insofar as data are available, the predictions of the theory discussed in this section appear to be borne out. A number of these phenomena would be difficult to explain using a tapped delay line model. The major gap in the theory as presented so far is the lack of account for experiments involving reward omission. We now show how to treat these as examples of partial observability. The model discussed so far follows as a limiting case of the resulting, more complex model whenever the world's state is directly observable.

5 A New Model: Partial Observability

Here we extend the model described in the previous section to include partial observability. We specify the formalism and discuss algorithms for state inference and value learning. Next, we present simulation results and analysis for several experiments involving temporal variability and reward omission. Finally, we discuss how the model's predictions fare in the light of available data.

5.1 A Partially Observable Semi-Markov Model. Partial observability results from relaxing the one-to-one correspondence between states and observables that was assumed above. We assume that there is a set $\mathcal{O}$ of possible observations (which we take, for presentational simplicity, as each instantaneous and binary) and that reward is simply a special observation. The state evolves as before, but it is not observable; instead, each state is associated with a multinomial distribution over $\mathcal{O}$, specified by an observation function O. Thus, when the process enters state $\mathbf{s}_k$, it emits an

accompanying observation $\mathbf{o}_k \in \mathcal{O}$ according to a multinomial conditioned on the state. One observation in $\mathcal{O}$ is the null observation, $\emptyset$. If no state transition occurs at time t, then $\mathbf{o}_t = \emptyset$. That is, nonempty observations can occur only on state transitions. Crucially, the converse is not true: state transitions can occur silently. This is how we model omitted reward in a trace conditioning experiment.

We wish to find a TD algorithm for value prediction in this formalism. Most of the terms in the error signal of equation 4.2 are unavailable, because the states they depend on are unobservable. In fact, it is not even clear when to apply the update, since the times of state transitions are themselves unobservable. Extending a standard approach to partial observability in Markov processes (Chrisman, 1992; Kaelbling et al., 1998) to the semi-Markov case, we assume that the animal learns a model of the hidden process (that is, the functions T, O, and D determining the conditional probabilities of state transitions, observations, and dwell time durations). Such a model can be used together with the observations to infer estimates of the unavailable quantities. Note that given such a model, the values of the hidden states could in principle be computed offline using value iteration. (Since the hidden process is just a normal semi-Markov process, partial observability does not affect the solution.) We return to this point in the discussion; here, motivated by evidence of dopaminergic involvement in error-driven learning, we present an online TD algorithm for learning the same values by sampling, assisted by the model.

The new error signal has a form similar to the fully observable case:

$$\delta_{s,t} = \beta_{s,t}(\mathbf{r}_{t+1} - \rho_t \cdot E[\mathbf{d}_t] + E[\widehat{V}_{s_{t+1}}] - \widehat{V}_s). \tag{5.1}$$

We discuss the differences, from left to right.

First, the new error signal is state as well as time dependent. We compute an error signal for each state s at each time step t, on the hypothesis that the process transitioned out of s at t. The error signal is weighted by the probability that this is actually the case:

$$\beta_{s,t} \equiv P(\mathbf{s}_t = s, \phi_t = 1 | \mathbf{o}_1, \ldots, \mathbf{o}_{t+1}), \tag{5.2}$$

where ϕ_t is a binary indicator that takes the value one if the state transitioned between times t and $t+1$ (self-transitions count) and zero otherwise. Note that this determination is conditioned on observations made through time $t+1$; this is chosen to parallel the one-time-step backup in the TD algorithm. β can be tracked using a version of the standard forward-backward recursions for hidden Markov models; equations are given in the appendix.

The remaining changes in the error signal of equation 5.1 are the expected dwell time $E[d_t]$ and expected value of the successor state $E[\widehat{\mathbf{V}}_{s_{t+1}}]$. These

are computed from the observations using the model, again conditioned on the hypothesis that the system left state s at time t:

$$E[\mathbf{d}_t] \equiv \sum_{d=1}^{\infty} d \cdot P(\mathbf{d}_t = d | \mathbf{s}_t = s, \phi_t = 1, \mathbf{o}_1, \ldots, \mathbf{o}_{t+1}) \tag{5.3}$$

$$E[\widehat{V}_{s_{t+1}}] \equiv \sum_{s' \in \mathcal{S}} \widehat{V}_{s'} P(\mathbf{s}_{t+1} = s' | \mathbf{s}_t = s, \phi_t = 1, \mathbf{o}_{t+1}). \tag{5.4}$$

Expressions for computing these quantities using the inference model are given in the appendix.

Assuming the inference model is correct (i.e., that it accurately captures the process generating the observations), this TD algorithm for value learning is exact in that it has the same fixed point as value iteration using the inference model. The proof is sketched in the appendix. Note also that in the fully observable limit (i.e., when $\mathbf{s}$, $\mathbf{d}$, and ϕ can be inferred with certainty), the algorithm reduces exactly to the semi-Markov rule of equation 4.2. Simulations (not reported here) demonstrate that in general, when the posterior distributions over these parameters are relatively well specified (i.e., when uncertainty is moderate), this algorithm behaves similarly to the semi-Markov algorithm described in section 4. The main differences come about in cases of considerable state uncertainty, as when reward is omitted.

We have described the TD error computation for learning values in a partially observable semi-Markov process. It may be useful to review how the computation actually proceeds. At each time step, the system receives a (possibly empty) observation or reward, and the representational system uses this to update its estimate of the state departure distribution β and other latent quantities. The TD learning system uses these estimates to compute the TD error δ, which is reported by the dopamine system (perhaps assisted by the serotonin system). Stored value estimates $\widehat{\mathbf{V}}$ are updated to reduce the error, and the cycle repeats.

5.2 Connecting This Theory to the Dopamine Response. In order to finalize the specification of the model, we discuss several further issues about simulating the dopamine response with the partially observable semi-Markov algorithm.

5.2.1 Vector vs. Scalar Error Signals. As already mentioned, equation 5.1 is a vector rather than a scalar error signal, since it contains an error for each state's value. Previous models have largely assumed that the dopamine response reports a scalar error signal, supported by experiments showing striking similarity between the responses of most dopaminergic neurons (Schultz, 1998). However, there is some variability between neurons.

Notably, only a subset (55–75%) of neurons displays any particular sort of phasic response (Schultz, 1998). Also, several studies report sizable subsets of neurons showing qualitatively different patterns of responding in the same situation (e.g., excitation versus inhibition; Schultz & Romo, 1990; Mirenowicz & Schultz, 1996; Waelti, Dickinson, & Schultz, 2001; Tobler, Dickinson, & Schultz, 2003, though see Ungless, Magill, & Bolam, 2004).

We suggest that dopamine neurons might code a vector error signal like equation 4.2 in some distributed manner and that this might account for response variation between neurons. Absent data from experiments designed to constrain such a hypothesis, we illustrate for this article the dopamine response as a scalar, cumulative error over all the states:

$$\delta_t = \sum_{s \in \mathcal{S}} \delta_{s,t}. \tag{5.5}$$

This quantity may be interpreted in terms of either a vector or scalar model of the dopamine signal. If dopamine neurons uniformly reported this scalar signal, then values could be learned by apportioning the state-specific error according to $\beta_{s,t}$ at targets (with $\widehat{V}_s$ updated proportionally to $\delta_t \cdot \beta_{s,t} / \sum_{s' \in \mathcal{S}} \beta_{s',t}$). This is a reasonable approximation to the full algorithm so long as there is only moderate state uncertainty and works well in our experience (simulations not reported here). The simplest vector signal would have different dopamine neurons reporting the state-dependent error $\delta_{s,t}$ for different states; the scalar error δ_t could then be viewed as modeling the sum or average over a large group of neurons. It is noteworthy that articles on dopamine neuron recordings predominantly report data in terms of such averages, accompanied by traces of a very few individual neurons. However, such a sparsely coded vector signal is probably unrealistic given the relative homogeneity reported for individual responses. A viable compromise might be a more coarsely coded vector scheme in which each dopamine neuron reports the cumulative TD error over a random subset of states. For large enough subsets (e.g, more than 50% of states per neuron), such a scheme would capture the overall homogeneity but limited between-neuron variability in the single-unit responses. In this case, the aggregate error signal from equation 5.5 would represent both the average over neurons and, roughly, a typical single-neuron response.

5.2.2 World Models and Asymptotic Model Uncertainty. As already mentioned, because our focus is on the influence of an internal model on asymptotic dopamine responding rather than on the process of learning such a model, for each task we take as given a fixed world model based on the actual structure that generated the task events. For instance, for trace-conditioning experiments, the model is based on Figure 4. Importantly, however, we assume that animals never become entirely certain

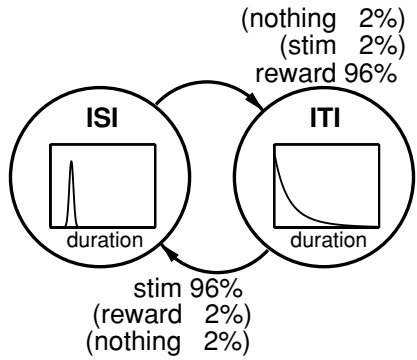

Figure 9: State space for semi-Markov model of trace conditioning experiment, with asymptotically uncertain dwell time distributions and observation models. For simplicity, analogous noise in the transition probabilities (small chance of self-transition) is not illustrated.

about the world's precise contingencies; each model is thus systematically altered to include asymptotic uncertainty in its distributions over dwell times, transitions, and observations. This variance could reflect irreducible sensor noise (e.g., time measurement error) and persistent uncertainty due to assumed nonstationarity in the contingencies being modeled (Kakade & Dayan, 2000, 2002). We represent even deterministic dwell-time distributions as gaussians with nonzero variance. Similarly, the multinomials describing observation and state transition probabilities attribute nonzero probability even to anomalous events (such as state self-transitions or reward omission). The effects of these modifications on the semi-Markov model of trace conditioning are illustrated in Figure 9.

5.3 Results: Simulated Dopaminergic Responding in the Model. We first consider the basic effect of reward omission. Figure 10 (left top) shows the effect on a trace-conditioning task, with the inferred state distribution illustrated by a shaded bar under the trace. As time passes without reward, inferred probability mass leaks into the ITI state (shown as the progression from black to white in the bar, blown up on the inset), accompanied by negative TD error. Because the model's judgment that the reward was omitted occurs progressively, the predicted dopaminergic inhibition is slightly delayed and prolonged compared to the expected time of reward delivery.

Repeated reward omission also reduces the value predictions attributed to preceding stimuli, which in turn has an impact on the dopaminergic responses to the stimuli and to the subsequent rewards when they arrive. Figure 11 shows how, asymptotically, the degree of partial reinforcement trades off relative responding between stimuli and rewards.

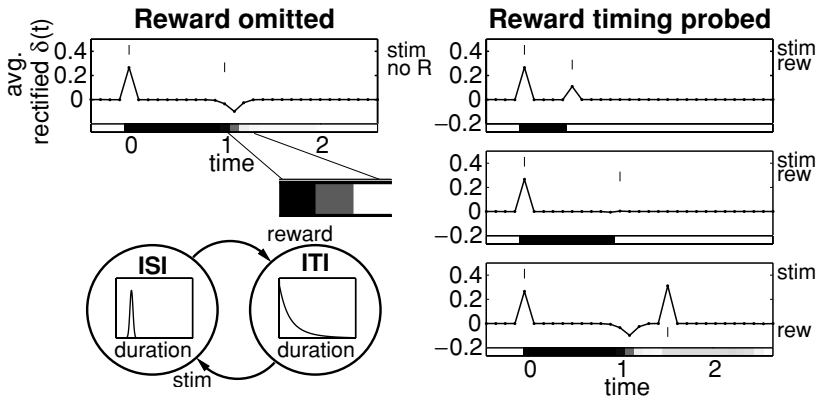

Figure 10: Simulated dopamine responses from partially observable semi-Markov model of trace conditioning, with reward omitted or delivered at an unexpected time. (Left top) Reward omission inhibits response, somewhat after the time reward was expected. (Left bottom) State space of inference model. (Right) When reward is delivered earlier (top) or later (bottom) than expected, excitation is seen to reward. No inhibition is seen after early rewards, at the time reward was originally expected. Shaded bars show inferred state (white: ITI, black: ISI). Mean ITI: 5 sec; ISI: 1 sec; reward magnitude: 1; rectification threshold: −0.1; probability of anomalous event in inference model: 2%; CV of dwell time uncertainty: 0.08.

Hollerman and Schultz (1998) generalized the reward omission experiment to include probe trials in which the reward was delivered a half-second early or late. Figure 10 (right) shows the behavior of the partially observable semi-Markov model on this task. In accord with the findings discussed in section 2, the semi-Markov model predicts no inhibition at the time the reward was originally expected. As shown by the shaded bar underneath the trace, this is because the model assumes that the early reward has signaled an early transition into the ITI state, where no further reward is expected. While the inference model judges such an event unlikely, it is a better explanation for the observed data than any other path through the state space.

5.4 Discussion: Data Bearing on These Predictions. The partially observable model behaves like the fully observable model for the experiments reported in the previous section (simulations not shown) and additionally accounts for dopamine responses when reward is omitted (Schultz et al., 1997) or delivered at unexpected times (Hollerman & Schultz, 1998). The results are due to the inference model making judgments about the likely progression of the hidden state when observable signals are absent

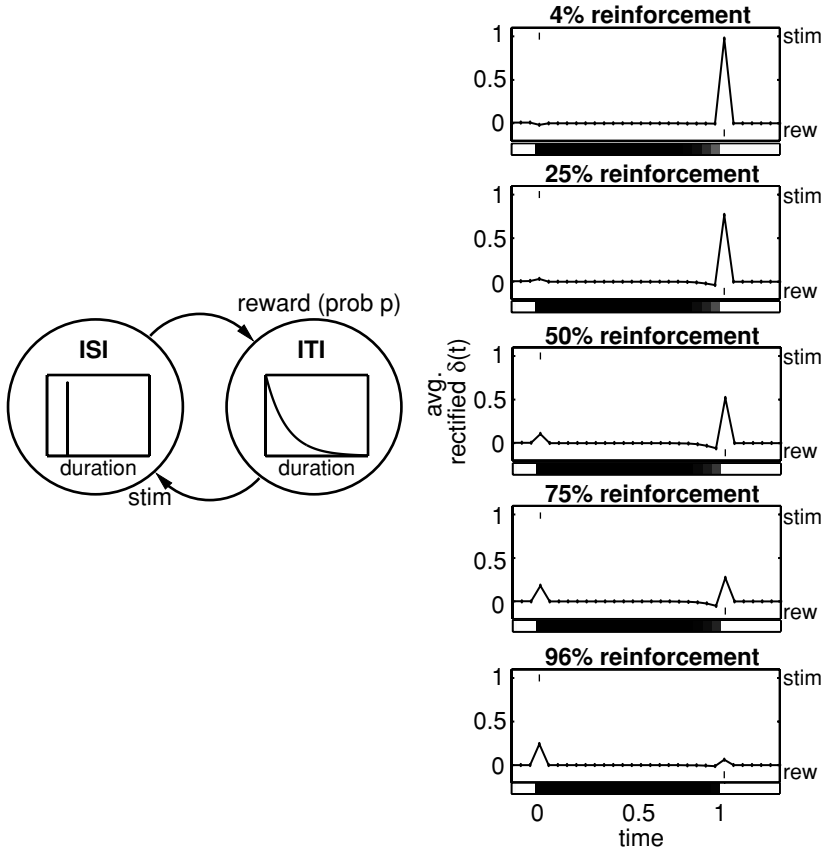

Figure 11: Simulated dopamine responses from partially observable semi-Markov model of trace conditioning with different levels of partial reinforcement. (Left) State space for inference. (Right) As chance of reinforcement increases, phasic responding to the reward decreases while responding to the stimulus increases. Shaded bars show inferred state (white: ITI, black: ISI). Mean ITI: 5 sec; ISI: 1 sec; reward magnitude: 1; rectification threshold: -0.1; probability of anomalous event in inference model: 2%; CV of dwell time uncertainty: 0.08.

or unusual. Since such judgments unfold progressively with the passage of time, simulated dopaminergic pauses are both later and longer than bursts. This difference is experimentally verified by reports of population duration and latency ranges (Schultz, 1998; Hollerman & Schultz, 1998) and is unexplained by delay line models.

However, in order to obtain pause durations similar to experimental data, it was necessary to assume fairly low levels of variance in the inference model's distribution over the interstimulus interval. That this variance (0.08) was much smaller than the level of timing noise suggested by behavioral experiments (0.3–0.5 according to Gibbon, 1977, or even 0.16 reported more recently by Gallistel, King, & McDonald, 2004), goes against the natural assumption that the uncertainty in the inference model captures the noise in internal timing processes. It is possible that because of the short, 1-second interevent durations used in the recordings, the monkeys are operating in a more accurate timing regime that seems behaviorally to dominate for subsecond intervals (see Gibbon, 1977). In this respect, it is interesting to compare the results of Morris et al. (2004), who recorded dopamine responses after slightly longer (1.5 and 2 seconds) deterministic trace intervals and reported noticeably more prolonged inhibitory responses. A fuller treatment of these issues would likely require both a more realistic theory that includes spike generation and as yet unavailable experimental analysis of the trial-to-trial variability in dopaminergic pause responses.

The new model shares with previous TD models the prediction that the degree of partial reinforcement should trade off phasic dopaminergic responding between a stimulus and its reward (see Figure 11). This is well verified experimentally (Fiorillo et al., 2003; Morris et al., 2004). One of these studies (Fiorillo et al., 2003) revealed an additional response phenomenon: a tonic "ramp" of responding between stimulus and reward, maximal for 50% reinforcement. Such a ramp is not predicted by our model (in fact, we predict very slight inhibition preceding reward, related to the possibility of an early, unrewarded state transition), but we note that such ramping is seen only in delay, and not trace, conditioning (Fiorillo et al., 2003; Morris et al., 2004). Therefore, it might reflect some aspect of the processing of temporally extended stimuli that our theory (which incorporates only instantaneous stimuli) does not yet contemplate. Alternatively, Niv, Duff, et al. (2005) suggest that the ramp might result from trial averaging over errors on the progression between states in a tapped-delay line model, due to the asymmetric nature of the dopamine response. This explanation can be directly reproduced in our semi-Markov framework by assuming that there is at least one state transition during the interstimulus interval, perhaps related to the persistence of the stimulus (simulations not reported; Daw, 2003). Finally, we could join the authors of the original study (Fiorillo et al., 2003) in assuming that the ramp is an entirely separate signal multiplexed with the prediction error signal. (Note, however, that although they associate the ramp with "uncertainty," this is not the same kind of uncertainty that we mainly discuss in this article. Our uncertainty measures posterior ignorance about the hidden state; Fiorillo et al. are concerned with known stochasticity or "risk" in reinforcer delivery.)

6 Discussion

We have developed a new account of the dopaminergic response that builds on previous ones in a number of directions. The major features in the model are a partially observable state and semi-Markov dynamics; these are accompanied by a number of further assumptions (including asymptotic uncertainty, temporal measurement noise, and the rectification of negative TD error) to produce new or substantially different explanations for a range of dopaminergic response phenomena. We have focused particularly on a set of experiments that exercise both the temporal and hidden state aspects of the model—those involving the state uncertainty that arises when an event can vary in its timing or be altogether omitted.

6.1 Comparing Different Models. The two key features of our model, partial observability and semi-Markov dynamics, are to a certain extent separable and each interesting in its own right. We have already shown how a fully observable semi-Markov model with interesting properties arises as a limiting case of our model when, as in many experimental situations, the observations are unambiguous. Another interesting relative, which has yet to be explored, would include partial observability and state inference but not semi-Markov dynamics. One way to construct such a model is to note that any discrete-time semi-Markov process of the sort described here has a hidden Markov model that is equivalent to it for generating observation sequences. This can be obtained by subdividing each semi-Markov state into a series of discrete time marker states, each lasting one time step (see Figure 12). State inference and TD are straightforward in this setting (in

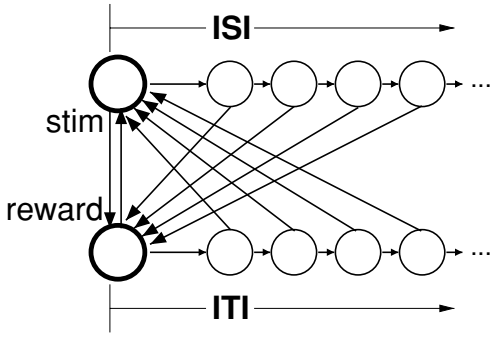

Figure 12: Markov model equivalent to the semi-Markov model of a trace conditioning experiment from Figure 4. The ISI and ITI states are subdivided into a series of substates that mark the passage of time. Stimuli and rewards occur only on transitions from one set of states into the other; dwell time distributions are encoded by the transition probabilities from each marker state.

particular, standard TD can be performed directly on the state posteriors, which are Markov; Kaelbling et al., 1998). Moreover, the fully observable limit of this model (equivalently, the model obtained by subdividing states in our fully observable semi-Markov model) is just a version of the familiar tapped delay line model, in which a series of marker states marks time from each event. Since each new event launches a new cascade of marker states and stops the old one, this model includes a "reset" device similar to those discussed in section 2. An interesting avenue for exploration would be intermediate hybrids, in which long delays are subdivided more coarsely into a series of a few temporally extended, semi-Markov states. A somewhat similar division of long delays into a series of several, temporally extended (internal) substates of different length is a feature of some behavioral theories of timing (Killeen & Fetterman, 1988; Machado, 1997), in part because animals engage in different behaviors during different portions of a timed interval.

Although the Markov and semi-Markov formalisms are equivalent as generative models, the TD algorithms for value learning in each are qualitatively different, because of their different approaches for "backing up" reward predictions over a long delay. The semi-Markov algorithms do so directly, while Markov algorithms employ a series of intermediate marker states. There is thus an empirical question as to which sort of algorithm best corresponds to the dopaminergic response.

A key empirical signature of the semi-Markov model would be its prediction that when reward or stimulus timing can vary, later-than-average reward-related events should inhibit dopamine neurons (e.g., see Figure 7). Markov models generally predict no such inhibition, but instead excitation that could shrink toward but not below baseline. In principle, this question could be addressed with trial-by-trial reanalysis of any of a number of extant data sets, since the new model predicts the same pattern of dopaminergic excitation and inhibition as a function of the preceding interval in a number of experiments, including ones as simple as free reward delivery (see Figure 5). Unfortunately, the one study directly addressing this issue was published only as an abstract (Fiorillo & Schultz, 2001). Though the original interpretation followed the Markov account, the publicly presented data appear ambiguous. (See the discussion in section 4 and note 1.)

A signature of many (though not all; Pan, Schmidt, Wickens, & Hyland, 2005) tapped delay line algorithms would be phasic dopaminergic activity during the period between stimulus and reinforcer, reflecting value backing up over intermediate marker states during initial acquisition of a stimulus-reward association. No direct observations have been reported of such intermediate activity, which may not be determinative since it would be subtle and fleeting. As already mentioned Niv, Duff, et al. (2005) have suggested that the tonic ramp observed in the dopamine response by Fiorillo et al. (2003) might reflect the average over trials of such a response. Should

this hypothesis be upheld by a trial-by-trial analysis of the data, it would be the best evidence for Markov over semi-Markov TD.

A final consideration regarding the trade-off between Markov and semi-Markov approaches is that, as Niv, Duff, and Dayan (2004) point out, Markov models are quite intolerant of timing noise. Our results suggest that semi-Markov models are more robustly able to account for dopaminergic response patterns in the light of presumably noisy internal timing.

It is worth discussing the contributions of two other features of our model that differ from the standard TD account. Both have also been studied separately in previous work. First, the asymmetry between excitatory and inhibitory dopamine responses has appreciable effects only when averaging over trials with different prediction errors. Thus, it is crucial in the present semi-Markov account, where most of the burst responses we simulate originate from the asymmetric average over errors that differ from trial to trial due to differences in event timing. Delay line models do not generally predict similar effects of event timing on error, and so in that context, asymmetric averaging has mainly been important in understanding experiments in which error varies due to differential reward delivery (as in Figure 7; Niv, Duff, et al., 2005).

In contrast, the use of an average-reward TD rule (rather than the more traditional discounted formulation) plays a more cosmetic role in this work. In the average reward formulation, trial-to-trial variability in delays $\mathbf{d}_k$ affects the prediction error (see equation 4.2) through the term $-\rho_k \cdot \mathbf{d}_k$; in a discounted version, analogous effects would occur due to long delays being more discounted (as $\gamma^{\mathbf{d}_k}$). One advantage of the average reward formulation is that it is invariant to dilations or contractions of the timescale of events, which may be relevant to behavior (discussed below). Previous work on average-reward TD in the context of delay line models has suggested that this formulation might shed light on tonic dopamine, dopamine-serotonin interactions, and motivational effects on response vigor (Daw & Touretzky, 2002; Daw et al., 2002; Niv, Daw, & Dayan, 2005).

6.2 Behavioral Data. Our model is concerned with the responses of dopamine neurons. However, insofar as dopaminergically mediated learning may underlie some forms of both classical and instrumental conditioning (e.g., Parkinson et al., 2002; Faure, Haberland, Condé, & Massioui, 2005), the theory suggests connections with behavioral data as well. Much more work will be needed to develop such connections fully, but we mention a few interesting directions here.

Our theory generalizes and provides some justification (in limited circumstances) for the "reset" hypothesis that had previously been proposed, on an ad hoc basis, to correct the TD account of the Hollerman and Schultz (1998) experiment (Suri & Schultz, 1998, 1999; Brown et al., 1999). In our theory, "reset" (here, an inferred transition into the ITI state) occurs after reward because this is consistent with the inference model for that experiment. But

in other circumstances, for instance, those in which a stimulus is followed by a sequence of more than one reward, information about the stimulus remains predictively relevant after the first reward, and our theory (unlike its predecessors) would not immediately discard it. Dopamine neurons have not been recorded in such situations, but behavioral experiments offer some clues. Animals can readily learn that a stimulus predicts multiple reinforcers; in classical conditioning, this has been repeatedly, though indirectly, demonstrated by showing that adding or removing reinforcers to a sequence has effects on learning (upward and downward unblocking; Holland & Kenmuir, 2005; Holland, 1988; Dickinson & Mackintosh, 1979; Dickinson, Hall, & Mackintosh, 1976). No such learning would be possible in the tapped delay line model if the first reward reset the representation. Because it has somewhat different criteria for what circumstances trigger a reset, the Suri and Schultz (1998, 1999) model would also have problems learning a task known variously as feature-negative occasion setting or sequential conditioned inhibition (Yin, Barnet, & Miller, 1994; Bouton & Nelson, 1998; Holland, Lamoureux, Han, & Gallagher, 1999). However, we should note that a different sort of behavioral experiment does support a reward-triggered reset in one situation. In an instrumental conditioning task requiring animals to track elapsed time over a period of about 90 seconds, reward arrival seems to reset animals' interval counts (Matell & Meck, 1999). It is unclear how to reconcile this last finding with the substantial evidence from classical conditioning.

Gallistel and Gibbon (2000) have argued that behavioral response acquisition in classical conditioning experiments is timescale invariant in the sense that contracting or dilating the speed of all events does not affect the number of stimulus-reward pairings before a response is seen. This would not be true for responses learned by simple tapped delay line TD models (since doubling the speed of events would halve the number of marker states and thereby speed acquisition), but this property naturally holds for our fully observable semi-Markov model (and would hold for the partially observable model if the unspecified model learning phase were itself timescale invariant).

There is also behavioral evidence that may relate to our predictions about the relationship between the dopamine response and the preceding interval. The latency to animals' behavioral responses across many instrumental tasks is correlated with the previous interreinforcement interval, with earlier responding after shorter intervals ("linear waiting"; Staddon & Cerutti, 2003). Given dopamine's involvement in response vigor (e.g., Dickinson, Smith, & Mirenowicz, 2000; Satoh, Nakai, Sato, & Kimura, 2003; McClure, Daw, & Montague, 2003; Niv, Daw, et al., 2005), this effect may reflect enhanced dopaminergic activity after shorter intervals, as our model predicts. However, the same reasoning applied to reward omission (after which dopaminergic responding is transiently suppressed) would incorrectly predict slower responding. In fact, animals respond earlier following reward

omission (Staddon & Innis, 1969; Mellon, Leak, Fairhurst, & Gibbon, 1995); we thus certainly do not have a full account of the many factors influencing behavioral response latency, particularly on instrumental tasks.

Particularly given these complexities, we stress that our theory is not intended as a theory of behavioral timing. To the contrary, it is adjunct to such a theory: it assumes an extrinsic timing signal. We have investigated how information about the passage of time can be combined with sensory information to make inferences about future reward probabilities and drive dopaminergic responding. We have explored the resulting effect on the dopaminergic response of one feature common to most behavioral timing models—scalar timing noise (Gibbon, 1977; Killeen & Fetterman, 1988; Staddon & Higa, 1999)—but we are otherwise rather agnostic about the timing substrate. One prominent timing theory, BeT (Killeen & Fetterman, 1988), assumes animals time intervals by probabilistically visiting a series of internally generated behavioral states of extended duration (see also LeT; Machado, 1997). While these behavioral "states" may seem to parallel the semi-Markov states of our account, it is important to recall that in our theory, the states are not actual internal states of the animal but rather are notional features of the animal's generative model of events in the external world. That generative model, plus extrinsic information about the passage of time, is used to infer a distribution over the external state. One concrete consequence of this difference can be seen in our simulations, in which even though the world is assumed to change abruptly and discretely, the internal representation is continuous and can sometimes change gradually (unlike the states of BeT).

Finally, both behavioral and neuroscientific data from instrumental conditioning tasks suggest that depending on the circumstances, animals seem to evaluate potential actions using either TD-style model-free or dynamic-programming-style model-based RL methods (Dickinson & Balleine, 2002; Daw, Niv, & Dayan, in press). This apparent heterogeneity of control is puzzling and relates to a potential objection to our framework. Specifically, our assumption that animals maintain a full world model may seem to make redundant the use of TD to learn value estimates, since the world model itself already contains the information necessary to derive value estimates (using dynamic programming methods such as value iteration). This tension may be resolved by considerations of efficiency and of balancing the advantages and disadvantages of both RL approaches. Given that the world model is in principle learned online and thus subject to ongoing change, it is computationally more frugal and often not appreciably less accurate to use TD to maintain relatively up-to-date value estimates rather than to repeat laborious episodes of value iteration. This simple idea is elaborated in RL algorithms like prioritized sweeping (Moore & Atkeson, 1993) and Dyna-Q (Sutton, 1990). In fact, animals behaviorally exhibit characteristics of each sort of planning under circumstances to which that method is computationally well suited (Daw et al., 2005).

6.3 Future Theoretical Directions. A number of authors have previously suggested schemes for combining a world model with TD theories of the dopamine system (Dayan, 2002; Dayan & Balleine, 2002; Suri, 2001; Daw et al., in press; Daw et al., 2005; Smith, Becker, & Kapur, 2005). However, all of this work concerns the use of a world model for planning actions. The present account explores a separate, though not mutually exclusive, function of a world model: for state inference to address problems of partial observability. Our work thus situates the hypothesized dopaminergic RL system in the context of a broader view of the brain's functional anatomy, in which the subcortical RL systems receive a refined, inferred sensory representation from cortex (similar frameworks have been suggested by Doya, 1999, 2000 and by Szita & Lorincz, 2004). Our Bayesian, generative view on this sensory processing is broadly consistent with recent theories of sensory cortex (Lewicki & Olshausen, 1999; Lewicki, 2002). Such theories may suggest how to address an important gap in the present work: we have not investigated how the hypothesized model learning and inference functions might be implemented in brain tissue. In the area of cortical sensory processing, such implementational questions are an extremely active area of research (Gold & Shadlen, 2002; Deneve, 2004; Rao, 2004; Zemel, Huys, Natarajan, & Dayan, 2004). Also, in the context of a generative model whose details are closer to our own, it has been suggested that a plausible neural implementation might be possible using sampling rather than exact inference for the state posterior (Kurth-Nelson & Redish, 2004). Here, we intend no claim that animal brains are using the same methods as we have to implement the calculations; our goal is rather to identify the computations and their implications for the dopamine response.

The other major gap in our presentation is that while we have discussed reward prediction in partially observable semi-Markov processes, we have not explicitly treated action selection in the analogous decision processes. It is widely presumed that dopaminergic value learning ultimately supports the selection of high-valued actions, probably by an approach like actor-critic algorithms (Sutton, 1984; Sutton & Barto, 1998), which use TD-derived value predictions to influence a separate process of learning action selection policies. There is suggestive evidence from functional anatomy that distinct dopaminergic targets in the ventral and dorsal striatum might subserve these two functions (Montague et al., 1996; Voorn, Vanderschuren, Groenewegen, Robbins, & Pennartz, 2004; Daw et al., in press; but see also Houk et al., 1995; Joel, Niv, & Ruppin, 2002, for alternative views). This idea has also recently received more direct support from fMRI (O'Doherty et al., 2004) and unit recording (Daw, Touretzky, & Skaggs, 2004) studies.

We do not envision the elaborations in the present theory as substantially changing this picture. That said, hidden state deeply complicates the action selection problem in Markov decision processes (i.e., partially observable MDPs or POMDPs; Kaelbling et al., 1998). The difficulty, in a nutshell, is that the optimal action when the state is uncertain may be different

from what would be the optimal action if the agent were certainly in any particular state (e.g., for exploratory or information-gathering actions)—and, in general, the optimal action varies with the state posterior, which is a continuous-valued quantity unlike the manageably discrete state variable that determines actions in a fully observable MDP. Behavioral and neural recording experiments on monkeys in a task involving a simple form of state uncertainty suggest that animals indeed use their degree of state uncertainty to guide their behavior (Gold & Shadlen, 2002).

Since the appropriate action can vary continuously with the state posterior, incorporating action selection in the present model would require approximating the policy (and value) as functions of the full belief state, preferably nonlinearly.

6.4 Future Experimental Directions. Our work enriches previous theories of dopaminergic responding by identifying two important theoretical issues that should guide future experiments: the distinction between Markov and semi-Markov backup and partial observability. We have already discussed how the former issue suggests new experimental analyses; similarly, issues of hidden state are ripe for future experiment.

Partial observability suggests a particular question: whether dopaminergic neurons report an aggregate error signal or separate error signals tied to different hypotheses about the world's state (see section 5.2). This could be studied more or less directly, for instance, by placing an animal in a situation where ambiguous reward expectancy (e.g., one reward, or two, or three) resolved into a situation where the intermediate reward was expected with certainty. On a scalar error code, dopamine neurons should not react to this event; with a vector error code, different neurons should report both positive and negative error.

More generally, it would be interesting to study how dopaminergic neurons behave in many situations of sensory ambiguity (as in noisy motion discriminations, e.g., Gold & Shadlen, 2002, where much is known about how cortical neurons track uncertainty but there is no indication how, if at all, the dopaminergic system is involved). The present theory and the antecedent theory of partially observable Markov decision processes suggest a framework by which many such experiments could be designed and analyzed.

Appendix

Here we present and sketch derivations for the formulas for inference in the partially observable semi-Markov model. Inference rules for similar hidden semi-Markov models have been described by Levinson (1986) and Guedon & Cocozza-Thivent (1990). We also sketch the proof of the correctness of the TD algorithm for the model.

Below, we use abbreviated notation for the transition, dwell time, and observation functions. We write the conditional transition probabilities as $T_{s',s} \equiv P(\mathbf{s}_k = s|\mathbf{s}_{k-1} = s')$; the conditional dwell time distributions as $D_{s,d} \equiv P(\mathbf{d}_k = d|\mathbf{s}_k = s)$; and the observation probabilities as $O_{s,o} \equiv P(\mathbf{o}_k = o|\mathbf{s}_k = s)$. These functions, together with the observation sequence, are given; the goal is to infer posterior distributions over the unobservable quantities needed for TD learning.

The chief quantity necessary for the TD learning rule of equation 5.1 is $\beta_{s,t} = P(\mathbf{s}_t = s, \phi_t = 1|\mathbf{o}_1 \ldots \mathbf{o}_{t+1})$, the probability that the process left state s at time t. To perform the one time step of smoothing in this equation, we use Bayes' theorem on the subsequent observation:

$$\beta_{s,t} = \frac{P(\mathbf{o}_{t+1}|\mathbf{s}_t = s, \phi_t = 1) \cdot P(\mathbf{s}_t = s, \phi_t = 1|\mathbf{o}_1 \ldots \mathbf{o}_t)}{P(\mathbf{o}_{t+1}|\mathbf{o}_1 \ldots \mathbf{o}_t)}. \tag{A.1}$$

In this equation and several below, we have made use of the Markov property: the conditional independence of $\mathbf{s}_{t+1}$ and $\mathbf{o}_{t+1}$ from the previous observations and states given the predecessor state $\mathbf{s}_t$. In semi-Markov processes (unlike Markov processes), this property holds only at a state transition, that is, when $\phi_t = 1$.

The first term of the numerator of equation A.1 can be computed by integrating over $\mathbf{s}_{t+1}$ in the model: $P(\mathbf{o}_{t+1}|\mathbf{s}_t = s, \phi_t = 1) = \sum_{s' \in \mathcal{S}} T_{s,s'} O_{s',\mathbf{o}_{t+1}}$.

Call the second term of the numerator of equation A.1 $\alpha_{s,t}$. Computing it requires integrating over the possible durations of the stay in state s:

$$\alpha_{s,t} = P(\mathbf{s}_t = s, \phi_t = 1|\mathbf{o}_1 \ldots \mathbf{o}_t) \tag{A.2}$$

$$= \sum_{d=1}^{\infty} P(\mathbf{s}_t = s, \phi_t = 1, \mathbf{d}_t = d|\mathbf{o}_1 \ldots \mathbf{o}_t) \tag{A.3}$$

$$= \sum_{d=1}^{\infty} \frac{P(\mathbf{o}_{t-d+1} \ldots \mathbf{o}_t|\mathbf{s}_t = s, \phi_t = 1, \mathbf{d}_t = d, \mathbf{o}_1 \ldots \mathbf{o}_{t-d}) \cdot P(\mathbf{s}_t = s, \phi_t = 1, \mathbf{d}_t = d|\mathbf{o}_1 \ldots \mathbf{o}_{t-d})}{P(\mathbf{o}_{t-d+1} \ldots \mathbf{o}_t|\mathbf{o}_1 \ldots \mathbf{o}_{t-d})} \tag{A.4}$$

$$= \sum_{d=1}^{\infty} \frac{O_{s,\mathbf{o}_{t-d+1}} D_{s,d} P(\mathbf{s}_{t-d+1} = s, \phi_{t-d} = 1|\mathbf{o}_1 \ldots \mathbf{o}_{t-d})}{P(\mathbf{o}_{t-d+1} \ldots \mathbf{o}_t|\mathbf{o}_1 \ldots \mathbf{o}_{t-d})}, \tag{A.5}$$

where the sum need not actually be taken out to infinity, but only until the last time a nonempty observation was observed (where a state transition must have occurred). The derivation makes use of the fact that the observation $\mathbf{o}_t$ is empty with probability one except on a state transition. Thus, under the hypothesis that the system dwelled in state s from time $t - d + 1$ through time t, the probability of the sequence of null observations during that period equals just the probability of the first, $O_{s,\mathbf{o}_{t-d+1}}$.

Integrating over predecessor states, the quantity $P(\mathbf{s}_{t-d+1} = s, \phi_{t-d} = 1|\mathbf{o}_1 \ldots \mathbf{o}_{t-d})$, the probability that the process *entered* state s at time $t - d + 1$, equals

$$\sum_{s' \in \mathcal{S}} T_{s',s} \cdot P(\mathbf{s}_{t-d} = s', \phi_{t-d} = 1|\mathbf{o}_1 \ldots \mathbf{o}_{t-\tau}) = \sum_{s' \in \mathcal{S}} T_{s',s} \cdot \alpha_{s',t-d}. \tag{A.6}$$

Thus, α can be computed recursively, and prior values of α back to the time of the last nonempty observation can be cached, allowing dynamic programming analogous to the Baum-Welch procedure for hidden Markov models (Baum, Petrie, Soulds, & Weiss, 1970).

Finally, the normalizing factors in the denominators of equations A.5 and A.1 can be computed by similar recursions, after integrating over the state occupied at $t - d$ (see equation A.5) or t (see equation A.1) and the value of ϕ at those times. Though we do not make use of this quantity in the learning rules, the belief state over state occupancy, $B_{s,t} = P(\mathbf{s}_t = s|\mathbf{o}_1 \ldots \mathbf{o}_t)$, can also be computed by a recursion on α exactly analogous to equation A.2, substituting $P(\mathbf{d}_t \geq d|\mathbf{s}_t = s)$ for $D_{s,d}$.

The two expectations in the TD learning rule of equation 5.1 are:

$$E[\widehat{V}_{\mathbf{s}_{t+1}}] = \sum_{s' \in \mathcal{S}} \widehat{V}_{s'} P(\mathbf{s}_{t+1} = s'|\mathbf{s}_t = s, \phi_t = 1, \mathbf{o}_{t+1}) \tag{A.7}$$

$$= \sum_{s' \in \mathcal{S}} \widehat{V}_{s'} \frac{T_{s,s'} O_{s',\mathbf{o}_{t+1}}}{\sum_{s''} T_{s,s''} O_{s'',\mathbf{o}_{t+1}}} \tag{A.8}$$

and

$$E[\mathbf{d}_t] = \sum_{d=1}^{\infty} d \cdot P(\mathbf{d}_t = d|\mathbf{s}_t = s, \phi_t = 1, \mathbf{o}_1 \ldots \mathbf{o}_{t+1}) \tag{A.9}$$

$$= \sum_{d=1}^{\infty} d \cdot P(\mathbf{d}_t = d|\mathbf{s}_t = s, \phi_t = 1, \mathbf{o}_1 \ldots \mathbf{o}_t) \tag{A.10}$$

$$= \frac{\sum_{d=1}^{\infty} d \cdot P(\mathbf{s}_t = s, \mathbf{d}_t = d, \phi_t = 1|\mathbf{o}_1 \ldots \mathbf{o}_t)}{\alpha_{s,t}}, \tag{A.11}$$

where the sum can again be truncated at the time of the last nonempty observation, and $P(\mathbf{s}_t = s, \mathbf{d}_t = d, \phi_t = 1|\mathbf{o}_1 \ldots \mathbf{o}_t)$ is computed as on the right-hand side of equation A.2.

The proof that the TD algorithm of equation 5.1 has the same fixed point as value iteration is sketched below. We assume the inference model correctly matches the process generating the samples. With each TD update, $\widehat{V}_s$ is nudged toward some target value with some step size $\beta_{s,t}$. It is easy

to show that, analogous to the standard stochastic update situation with constant step sizes, the fixed point is the average of the targets, weighted by their probabilities and their step sizes. Fixing some arbitrary t, the update targets and associated step sizes β are functions of the observations $\mathbf{o}_1, \ldots, \mathbf{o}_{t+1}$, which are, by assumption, samples generated with probability $P(\mathbf{o}_1, \ldots, \mathbf{o}_{t+1})$ by a semi-Markov process whose parameters match those of the inference model. The fixed point is

$$\widehat{V}_s = \frac{\sum_{\mathbf{o}_1 \ldots \mathbf{o}_{t+1}} [P(\mathbf{o}_1 \ldots \mathbf{o}_{t+1}) \cdot \beta_{s,t} \cdot (r(\mathbf{o}_{t+1}) - \rho_t \cdot E[\mathbf{d}_t] + E[\widehat{V}_{\mathbf{s}_{t+1}}])]}{\sum_{\mathbf{o}_1 \ldots \mathbf{o}_{t+1}} [P(\mathbf{o}_1 \ldots \mathbf{o}_{t+1}) \cdot \beta_{s,t}]}, \tag{A.12}$$

where we have written the reward $\mathbf{r}_{t+1}$ as a function of the observation, $r(\mathbf{o}_{t+1})$, because rewards are just a special case of observations in the partially observable framework. The expansions of $\beta_{s,t}$, $E[\mathbf{d}_t]$, and $E[\widehat{V}_{\mathbf{s}_{t+1}}]$ are all conditioned on $P(\mathbf{o}_1, \ldots, \mathbf{o}_{t+1})$, where this probability from the inference model is assumed to match the empirical probability appearing in the numerator and denominator of this expression. Thus, we can marginalize out the observations in the sums on the numerator and denominator, reducing the fixed-point equation to

$$\widehat{V}_s = \sum_{s' \in \mathcal{S}} \left[T_{s,s'} \left(\sum_{o \in \mathcal{O}} [O_{s',o} \cdot r(o)] + \widehat{V}_{s'} \right) \right] - \sum_d \rho_t \cdot d \cdot D_{s,d}, \tag{A.13}$$

which (assuming $\rho_t = \rho$) is Bellman's equation for the value function, and is also by definition the same fixed point as value iteration.

Acknowledgments

This work was supported by National Science Foundation grants IIS-9978403 and DGE-9987588. A.C. was funded in part by a Canadian NSERC PGS B fellowship. N.D. is funded by a Royal Society USA Research Fellowship and the Gatsby Foundation. We thank Sham Kakade, Yael Niv, and Peter Dayan for their helpful insights, and Chris Fiorillo, Wolfram Schultz, Hannah Bayer, and Paul Glimcher for helpfully sharing with us, often prior to publication, their thoughts and experimental observations.

References

Baum, L. E., Petrie, T., Soulds, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164–171.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Bouton, M. E., & Nelson, J. B. (1998). Mechanisms of feature-positive and feature-negative discrimination learning in an appetitive conditioning paradigm. In N. A. Schmajuk & P. C. Holland (Eds.), *Occasion setting: Associative learning and cognition in animals* (pp. 69–112). Washington, DC: American Psychological Association.

Bradtke, S. J., & Duff, M. O. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems, 7* (pp. 393–400). Cambridge, MA: MIT Press.

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.

Chrisman, L. (1992). Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)* (pp. 183–188). San Jose, CA: AAAI Press.

Courville, A. C., Daw, N. D., Gordon, G. J., & Touretzky, D. S. (2003). Model uncertainty in classical conditioning. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems, 16* Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2004). Similarity and discrimination in classical conditioning: A latent variable account. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17*. Cambridge, MA: MIT Press.

Courville, A. C., & Touretzky, D. S. (2001). Modeling temporal structure in classical conditioning. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 3–10). Cambridge, MA: MIT Press.

Das, T., Gosavi, A., Mahadevan, S., & Marchalleck, N. (1999). Solving semi-Markov decision problems using average reward reinforcement learning. *Management Science*, *45*, 560–574.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications*. Unpublished doctoral dissertation, School of Computer Science, Carnegie Mellon University.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*, 603–616.

Daw, N. D., Niv, Y., & Dayan, P. (in press). Actions, values, policies, and the basal ganglia. In E. Bezard (Ed.), *Recent breakthroughs in basal ganglia research*. New York: Nova Science.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, *14*, 2567–2583.

Daw, N., Touretzky, D., & Skaggs, W. (2004). Contrasting neuronal correlates between dorsal and ventral striatum in the rat. In *Cosyne04 Computational and Systems Neuroscience Abstracts*, Vol. 1.

Dayan, P. (2002). Motivated reinforcement learning. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 11–18). Cambridge, MA: MIT Press.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation and reinforcement learning. *Neuron*, *36*, 285–298.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, *39*, 1–38.

Deneve, S. (2004). Bayesian inference in spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17*. Cambridge, MA: MIT Press.

Dickinson, A., & Balleine, B. (2002). The role of learning in motivation. In C. R. Gallistel (Ed.), *Stevens' handbook of experimental psychology* (3rd ed.), *Vol. 3: Learning, motivation and emotion* (pp. 497–533). New York: Wiley.

Dickinson, A., Hall, G., & Mackintosh, N. J. (1976). Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*, 313–322.

Dickinson, A., & Mackintosh, N. J. (1979). Reinforcer specificity in the enhancement of conditioning by posttrial surprise. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*, 162–177.

Dickinson, A., Smith, J., & Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behavioral Neuroscience*, *114*, 468–483.

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

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*, 732–739.

Faure, A., Haberland, U., Condé, F., & Massioui, N. E. (2005). Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *Journal of Neuroscience*, *25*, 2771–2780.

Fiorillo, C. D., & Schultz, W. (2001). The reward responses of dopamine neurons persist when prediction of reward is probabilistic with respect to time or occurrence. In *Society for Neuroscience Abstracts, 27*, 827.5.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902.

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

Gallistel, C. R., King, A., & McDonald, R. (2004). Sources of variability and systematic error in mouse timing behavior. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*, 3–16.

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

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*, 299–308.

Guedon, Y., & Cocozza-Thivent, C. (1990). Explicit state occupancy modeling by hidden semi-Markov models: Application of Derin's scheme. *Computer Speech and Language*, *4*, 167–192.

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

Holland, P. C., & Kenmuir, C. (2005). Variations in unconditioned stimulus processing in unblocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*, 155–171.

Holland, P. C., Lamoureux, J. A., Han, J., & Gallagher, M. (1999). Hippocampal lesions interfere with Pavlovian negative occasion setting. *Hippocampus*, *9*, 143–157.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*, 304–309.

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.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*, 535–547.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Kakade, S., & Dayan, P. (2000). Acquisition in autoshaping. In S. A. Solla, T. K. Leen, & K. R. Muller (Eds.), *Advances in neural information processing systems, 12*. Cambridge, MA: MIT Press.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review*, *109*, 533–544.

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*, 274–295.

Kurth-Nelson, Z., & Redish, A. (2004). μagents: Action-selection in temporally dependent phenomena using temporal difference learning over a collective belief structure. *Society for Neuroscience Abstracts, 30*, 207.1.

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, *1*, 29–45.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*, 356–363.

Lewicki, M. S., & Olshausen, B. A. (1999). A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *16*, 1587–1601.

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

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological Review*, *104*, 241–265.

Mahadevan, S., Marchalleck, N., Das, T., & Gosavi, A. (1997). Self-improving factory simulation using continuous-time average-reward reinforcement learning. In *Proceedings of the 14th International Conference on Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Matell, M. S., & Meck, W. H. (1999). Reinforcement-induced within-trial resetting of an internal clock. *Behavioural Processes*, *45*, 159–171.

McClure, S. M., Daw, N. D., & Montague, P. R. (2003). A computational substrate for incentive salience. *Trends in Neurosciences*, *26*, 423–428.

Mellon, R. C., Leak, T. M., Fairhurst, S., & Gibbon, J. (1995). Timing processes in the reinforcement-omission effect. *Animal Learning and Behavior*, *23*, 286–296.

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.

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, *13*, 103–130.

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.

Niv, Y., Daw, N. D., & Dayan, P. (2005). How fast to work: Response vigor, motivation, and tonic dopamine. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17*. Cambridge, MA: MIT Press.

Niv, Y., Duff, M. O., & Dayan, P. (2004). The effects of uncertainty on TD learning. In *Cosyne04—Computational and Systems Neuroscience Abstracts*, vol. 1.

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty, and TD learning. *Behavioral and Brain Functions*, *1*, 6.

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

Owen, A. M. (1997). Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*, *53*, 431–450.

Pan, W. X., Schmidt, R., Wickens, J., & Hyland, B. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience*, *25*, 6235–6242.

Parkinson, J. A., Dalley, J. W., Cardinal, R. N., Bamford, A., Fehnert, B., Lachenal, G., Rudarakanchana, N., Halkerston, K., Robbins, T. W., & Everitt, B. J. (2002). Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: Implications for mesoaccumbens dopamine function. *Behavioral Brain Research*, *137*, 149–163.

Rao, R. P. N. (2004). Hierarchical Bayesian inference in networks of spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17*. Cambridge, MA: MIT Press.

Rao, R. P. N., Olshausen, B. A., & Lewicki, M. S. (2002). *Probabilistic models of the brain: Perception and neural function*. Cambridge, MA: MIT Press.

Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *Journal of Neuroscience*, *23*, 9913–9923.

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

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

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

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

Smith, A. J., Becker, S., & Kapur, S. (2005). A computational model of the functional role of the ventral-striatal D2 receptor in the expression of previously acquired behaviors. *Neural Computation*, *17*, 361–395.

Staddon, J. E. R., & Cerutti, D. T. (2003). Operant conditioning. *Annual Reviews of Psychology*, *54*, 115–144.

Staddon, J. E. R., & Higa, J. J. (1999). Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior*, *71*, 215–251.

Staddon, J. E., & Innis, N. K. (1969). Reinforcement omission on fixed-interval schedules. *Journal of the Experimental Analysis of Behavior*, *12*, 689–700.

Suri, R. E. (2001). Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Experimental Brain Research*, *140*, 234–240.

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements with dopamine-like reinforcement signal in neural network model. *Experimental Brain Research*, *121*, 350–354.

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. (1984). *Temporal credit assignment in reinforcement learning*. Unpublished doctoral dissertation, University of Massachusetts.

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

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

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.

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

Szita, I., & Lorincz, S. (2004). Kalman filter control embedded into the reinforcement learning framework. *Neural Computation*, *16*, 491–499.

Tobler, P., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, *23*, 10402–10410.

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

Ungless, M. A., Magill, P. J., & Bolam, J. P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, *303*, 2040–2042.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neuroscience*, *27*, 468–474.

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

Yin, H., Barnet, R. C., & Miller, R. R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: Operational similarities and differences. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*, 419–428.

Zemel, R., Huys, Q., Natarajan, R., & Dayan, P. (2004). Probabilistic computation in spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17*. Cambridge, MA: MIT Press.

Received February 24, 2005; accepted September 29, 2005.