LETTER Communicated by Andrew Barto

Temporal Difference Model Reproduces Anticipatory Neural Activity

Roland E. Suri
Wolfram Schultz
Institut de Physiologie, 1700 Fribourg, Switzerland

Anticipatory neural activity preceding behaviorally important events has been reported in cortex, striatum, and midbrain dopamine neurons. Whereas dopamine neurons are phasically activated by reward-predictive stimuli, anticipatory activity of cortical and striatal neurons is increased during delay periods before important events. Characteristics of dopamine neuron activity resemble those of the prediction error signal of the temporal difference (TD) model of Pavlovian learning (Sutton & Barto, 1990). This study demonstrates that the prediction signal of the TD model reproduces characteristics of cortical and striatal anticipatory neural activity. This finding suggests that tonic anticipatory activities may reflect prediction signals that are involved in the processing of dopamine neuron activity.

1 Introduction

In a famous experiment by Pavlov (1927), a dog was trained with the ringing of a bell (stimulus) followed by food delivery (reinforcer). In the first trial, the animal salivated when food was presented. After several trials, salivation started when the bell was rung. This finding suggests that the salivation response following the bell ring reflects anticipation of food delivery. A large body of experimental evidence led to the hypothesis that Pavlovian learning is dependent on the degree of unpredictability of the reinforcer (Rescorla & Wagner, 1972; Dickinson, 1980). According to this hypothesis, reinforcers become progressively less efficient for behavioral adaptation as their predictability grows during the course of learning. The difference between the actual occurrence and the prediction of the reinforcer is usually referred to as the error in the reinforcer prediction. This concept has been employed in the temporal difference model (TD model) of Pavlovian learning (Sutton & Barto, 1990). The TD model uses reinforcement prediction errors for learning a reinforcement prediction signal. This signal was compared to anticipatory responses. As animals seem to optimize the sum of reinforcement over time (Mackintosh, 1974; Dickinson, 1980), it is the goal of the TD model to compute a desired prediction signal that reflects the sum of future reinforcement. If the reinforcement is food intake,

this desired prediction signal reflects the sum of available food in the future. After training of the TD model with a stimulus followed by a reinforcer, the prediction error signal increases phasically when the stimulus is presented, and the prediction signal is tonically increased during the intratrial interval. Recent studies relate the TD model to neural information processing because the reward prediction error of the TD model resembles dopamine neuron activity in situations with unpredicted rewards, fully predicted rewards, reward-predicting stimuli, and unexpectedly omitted rewards. The comparison between basal ganglia anatomy and the architecture of the TD model suggests that cortico-striatonigral pathways are involved in adaptation of dopamine neuron activities (Barto, 1995; Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Suri & Schultz, 1999).

Anticipatory activity is related to an upcoming event that is prerepresented as a result of a retrieval action of antedating events, in contrast to activity reflecting memorized features of a previously experienced event (Wagner, 1978). Therefore, as in Pavlov's experiment, anticipatory activity precedes a future event irrespective of the physical features of the antedating events, which make this future event predictable (see Figure 1A). Phasic activity anticipating rewards was reported in midbrain dopamine neurons (Ljungberg, Apicella, & Schultz, 1992; Schultz, Apicella, & Ljungberg, 1993; Schultz et al., 1997; Mirenowicz & Schultz, 1994). Tonic delay period activity that anticipates stimuli, rewards, or the animal's own actions was termed anticipatory, preparatory, or predictive and has been reported in the striatum (Hikosaka, Sakamoto, & Usui, 1989; Alexander & Crutcher, 1990a, 1990b; Apicella, Scarnati, Ljungberg, & Schultz, 1992; Schultz & Romo, 1992; Kermadi & Joseph, 1995; Tremblay, Hollerman, & Schultz, 1998; Hollerman, Tremblay, & Schultz,1998), supplementary motor area (Alexander & Crutcher, 1990a, 1990b; Romo & Schultz, 1992), prefrontal cortex (Watanabe, 1996), orbitofrontal cortex (Tremblay & Schultz, 1999, 2000; Schultz, Tremblay, & Hollerman, 2000), premotor cortex (Mauritz & Wise, 1986), and primary motor cortex (Alexander & Crutcher, 1990a, 1990b).

The TD model was usually applied to learn to predict one reinforcer. In situations with two different anticipated rewards, we use two TD models, each processing one reward. In order to investigate the relations between anticipatory neural activity and predictive signals of the TD model (Sutton & Barto, 1990), we compare simulated predictive signals with anticipatory neural activities.

2 Description of Anticipatory Neural Activity

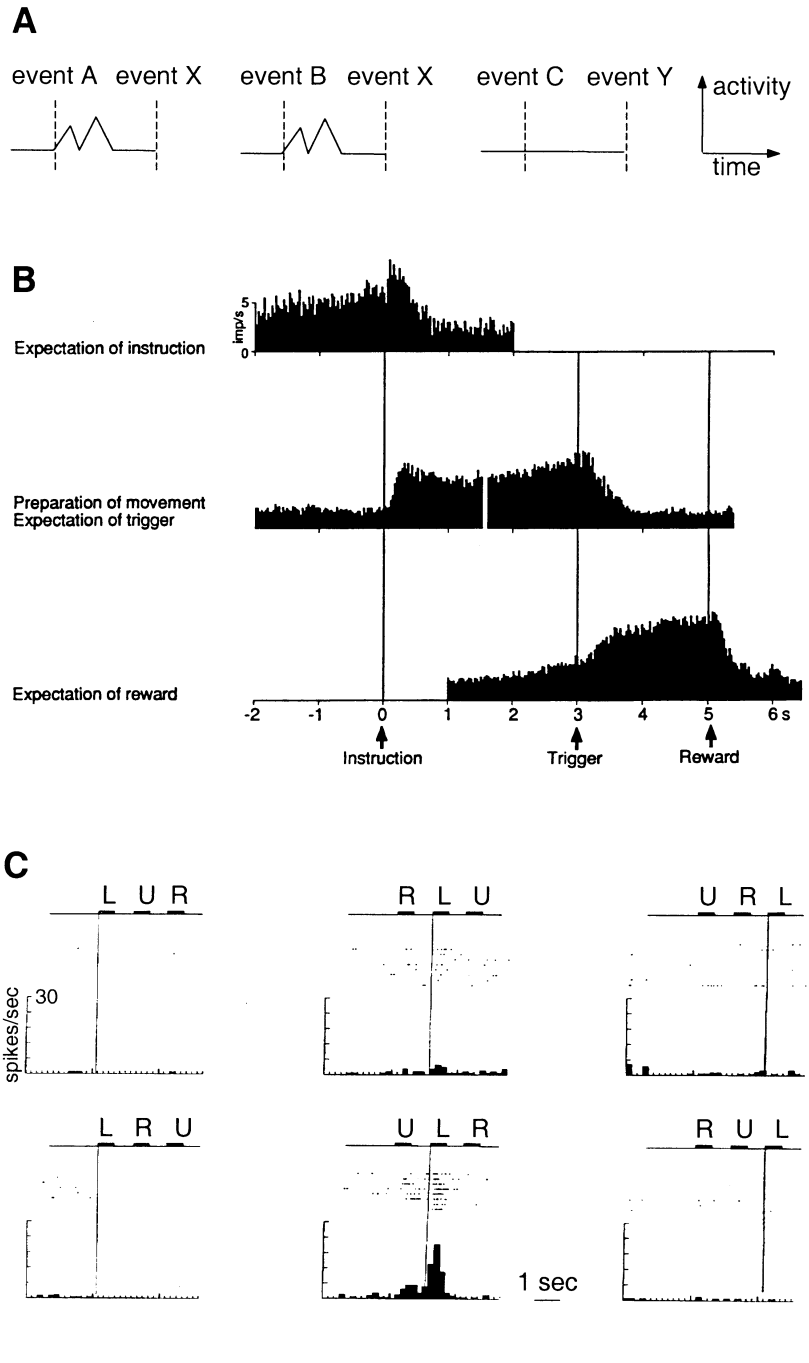
2.1 Phasic Anticipatory Activity. Phasic anticipatory neural responses of about 100 msec duration were reported for midbrain dopamine neurons. These neurons are activated by unpredicted rewards and by a reward following a stimulus for the first time. After repeated presentations of the

stimulus followed by the reward, the activation elicited by the reward decreases and entirely disappears after learning is completed. These neurons become activated instead by the stimulus (Ljungberg et al., 1992; Schultz, Apicella, & Ljungberg, 1993; Schultz et al., 1997; Mirenowicz & Schultz, 1994; Schultz, 1998).

2.2 Tonic Anticipatory Activity. Tonic anticipatory activity was found in subsets of cortical and striatal neurons. Before learning, such neurons often respond to specific events. When this event becomes predicted in the course of learning, these responses seem to become progressively preceded by anticipatory neural activity (Hikosaka et al., 1989; Tremblay et al., 1998). After learning, anticipatory neural activity often progressively increases between the first (predicting) event and the second (predicted) event. This activity starts increasing at the onset of the predicting event or during the interevent interval. Alternatively to this progressive increase, the responses can be limited to the predictive and to the predicted event as shown in Figure 4C (Hikosaka et al., 1989; Alexander & Crutcher, 1990a, 1990b; Apicella et al., 1992; Kermadi & Joseph, 1995; Watanabe, 1996).

Tonic anticipatory activity was reported to precede stimuli, reinforcers, and movements in delayed-response tasks (Apicella et al., 1992; Schultz & Romo, 1992; Schultz, Apicella, Scarnati, & Ljungberg, 1992; Tremblay et al., 1998; Hollerman et al., 1998). Each correct trial of this task consists of an instruction stimulus, a delay period, a trigger stimulus, a behavior, and a reward. Animals have to remember the instruction stimulus to react correctly to the trigger stimulus. Apicella and collaborators (1992) reported anticipatory neural activity in the monkey striatum after training a go-nogo version of this task. From the 1173 studied striatal neurons, 615 showed some change in activity during task performance. The activity of 193 task-related neurons increased in advance of at least one task component: the instruction stimulus (16 neurons), the trigger stimulus (15 neurons), the animal's movement (56 neurons), or the reward delivery (87 neurons) (see Figure 1B). These neurons with anticipatory activity were found in dorsal and anterior parts of caudate and putamen and were slightly more frequent in the proximity of the internal capsule.

Tremblay and collaborators (Tremblay & Schultz, 1999, 2000; Schultz et al., 2000) trained monkeys in delayed-response tasks in which each instruction stimulus preceded presentation of a specific reward (two liquids with different taste). Instruction stimulus A was followed by reward X, instruction stimulus B was followed by the same reward X, and instruction stimulus C was followed by reward Y. Neural activity was recorded in six-layered parts of orbitofrontal areas 11 and 14 and rostral area 13. These neurons showed three principal types of activation: responses to instructions (15% of 1095 tested neurons), responses following reward (8%), and sustained activations preceding reward (9%). Unrewarded control trials demonstrated that all three types of activation were influenced by the expected reward.


A
event A
event X
event B
event X
event C
event Y
activity
time
B
Expectation of instruction
Preparation of movement
Expectation of trigger
Expectation of reward
imp/s
5
0
-2
-1
0
1
2
3
4
5
6 s
Instruction
Trigger
Reward
C
L U R
R L U
U R L
L R U
U L R
R U L
spikes/sec
30
1 sec

The prereward activations began several seconds before the reward and subsided less than 1 sec after reward delivery.

Tonic anticipatory activities can be specific for anticipated stimuli (Hikosaka et al., 1989; Alexander & Crutcher, 1990b; Apicella et al., 1992; Kermadi & Joseph, 1995; Tremblay et al., 1998; Hollerman et al., 1998). Kermadi & Joseph (1995) analyzed the neural activity of 2100 neurons in the caudate nucleus. During the instruction phase, a sequence of three visual targets was presented, and the monkey was required to fixate on a central fixation point. In the subsequent behavioral phase, the monkey had to press

Figure 1: *Facing page*. (A) Illustration providing a criterion for the expression "activity anticipating event X." The three well-trained trial types "event A followed by event X," "event B followed by event X," and "event C followed by event Y" are assumed to be separated by sufficiently long intertrial intervals and are presented randomly intermixed. Presentations of event A (left) and event B (middle), which both precede presentation of event X, increase the activity. Event C, which precedes event Y, does not influence the activity (right side). This last control trial shows that the anticipatory activity is specific for event X. Furthermore, this control trial indicates that the activity is not related to a common physical feature of the events A, B, and C and therefore does not reflect memorization of these preceding events. The responses following events X and event Y are not shown because they are not relevant for this criterion. Events A and B were termed "predictive" events and event X the "predicted" or "anticipated" event. (B) Population activity of expectation- and preparation-related striatal neurons (figure from Apicella et al., 1992). (Top) Activation of 16 neurons preceding instruction onset. The intertrial interval was 4–7 seconds. (Middle) Activation of 44 neurons preceding the trigger stimulus in go trials. The histogram is split because the intervals between instruction and trigger varied from 2.5 to 3.5 sec. Neurons responding to the trigger stimulus, activated during movement, or activated before instruction or reward, are excluded. (Bottom) Activation of 68 neurons preceding reward in no-go trials. The activation began to a modest extent before trigger onset, gained increasingly in amplitude after trigger onset, and reached its peak when the reward was delivered. In each display, histograms for each neuron normalized for trial number are added, and the resulting sum is divided by the number of neurons. (C) Activity of this neuron in caudate anticipated presentation of stimulus L only if stimulus L occurred in the sequence ULR (figure from Kermadi & Joseph, 1995). During the shown instruction phase of the task, the monkey withheld movements and fixated on a central fixation point. The three visual stimuli—L (left target), U (upper target), and R (right target)—were presented in the six sequences: LUR, RLU, URL, LRU, ULR, and RUL (top of each subfigure). Each cell discharge is indicated by a dot, and the 6 to 9 successive trials per neuron are shown on successive lines (middle of each subfigure). The histogram shows the sum of the individual discharges (bottom of each subfigure).

three levers in the order indicated by the instruction phase. Six different sequences of the three targets were presented in the instruction phase. Because these six sequences were always presented in the same order, the three target stimuli in each trial were completely predictable. From 125 neurons responding in the instruction phase, the activity of 81 neurons preceded the presented stimuli. The activity of 46 neurons anticipated the offset of the central fixation point, the activity of 7 neurons anticipated the illumination of any target, the activity of 17 neurons anticipated the illumination of the first target, and the activity of 11 neurons anticipated the onset of specific targets. In a majority (35 neurons), the responses to specific targets were modulated by the rank of the target in the sequence or by complex relationships with other targets. Anticipatory activity started increasing about 1 second before stimulus onset. Then this activity progressively increased until it reached the maximum at the onset of the anticipated stimulus. A neuron with activity anticipating the specific target L in the specific sequence ULR is shown in Figure 1C.

2.3 Anticipatory Activity in Paradigms Without Delay Period. Although we do not intend to reproduce anticipatory neural activity in paradigms without delay period, we briefly mention such findings here. Activity of the head-direction cells in the anterior thalamus anticipates the future head direction by a neuron-specific duration between 0 and 50 msec (Blair, Lipscomb, & Sharp, 1997). Event-specific anticipatory activity can also depend on the future behavior of the animal. Activity that anticipates the retinal consequences of intended eye movements by about 100 msec was reported in frontal eye fields (Goldberg & Bruce, 1990; Umeno & Goldberg, 1997), superior colliculus (Walker, Fitzgibbon, & Goldberg, 1995), parietal cortex (Duhamel, Colby, & Goldberg, 1992), and striate cortex (Nakamura & Colby, 1999).

3 Description of the TD Model

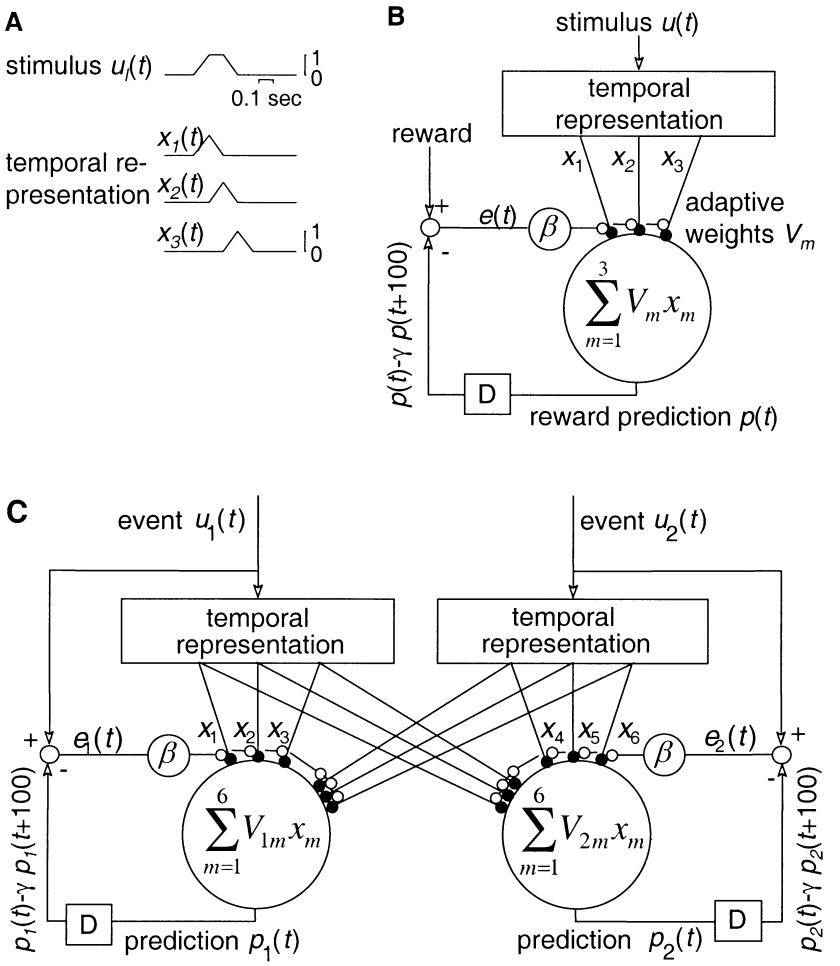
In Pavlovian learning paradigms, animals often learn to estimate the time of reward occurrence (Gallistel, 1990). Therefore, the TD model of Pavlovian learning (Sutton & Barto, 1990) proposes a time-estimation mechanism. The same time-estimation mechanism was also used to reproduce the finding that dopamine neuron activity is decreased below baseline levels when an expected reward is omitted (Montague et al., 1996; Schultz, 1998). This time-estimation mechanism is implemented by assuming that each stimulus is represented with a series of short components following stimulus onset. This is achieved by mapping each stimulus to a fixed temporal pattern of phasic signals $x_1(t), x_2(t), \ldots$ that follow stimulus onset with varying delays. This temporal pattern is referred to as a complete serial compound stimulus or temporal stimulus representation (see Figure 2A). Note that the choice of these signals is rather arbitrary. The single components have

been proposed to be phasic (Sutton & Barto, 1990), sustained (Desmond & Moore, 1988, 1991), or phasic immediately after the stimulus onset and then progressively more sustained (Grossberg & Schmajuk, 1989; Brown, Bullock, & Grossberg, 1999; Suri & Schultz, 1999). In some models the representation of a stimulus depends on successive events (Dominey, Arbib, & Joseph, 1995; Suri & Schultz, 1999). Although the shapes of the components differ among these models, the learned prediction and prediction error signals are usually not affected by this choice.

The temporal stimulus representation is used to compute the reward prediction signal with the adaptive weights $V_m(t)$ (see equation A.2; Sutton & Barto, 1990; Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1998). A representation of the TD model using a neuron-like element is shown in Figure 2B. According to the TD model, the reward prediction develops during learning in a similar way as the animal's anticipatory behavior. The reward prediction increases gradually before an anticipated reward if this reward is completely predicted. The rate of this gradual increase is determined by the constant γ, which is referred to as the temporal discount factor. The value of the discount factor γ was estimated from the time course of measured anticipatory neural activity. We usually used the standard value $\gamma = 0.99$ per time step (1 time step = 100 msec), which led to an increase in the prediction signal of 1% each 100 msec (see section 6). Previously, $\gamma = 0.98$ per 100 msec had been estimated from dopamine neuron activity (Suri & Schultz, 1999).

The TD model learns the reward prediction signal from stimuli antedating reward occurrence using a signal that reflects "errors" in the reward prediction. The TD model uses the difference between the actual occurrence and the prediction of the reward as this reward prediction error. Thus, the reward prediction error $e(t)$ is computed from discounted temporal differences in the prediction signal $p(t)$ and from the reward signal (see equation A.3). In order to minimize these prediction errors, the elements of the weight matrix $V_m(t)$ are incrementally adapted according to the product of the prediction error with eligibility traces of the temporal stimulus representation (see equation A.4). These traces are defined as slowly decaying versions of the representation components (see equation A.5). Such stimulus traces were originally introduced to explain learning for situations with a delay between the stimulus and the reinforcer, as they bridge the time interval between the predictive stimulus and the reinforcer (Hull, 1943). Although TD models with complete temporal stimulus representation learn without representation traces (Montague et al., 1996), the proposed model uses traces to accelerate learning (Sutton & Barto, 1998).

3.1 One TD Model for Each Event. Since rewards are usually accompanied by certain sensory stimuli, it was proposed to represent a reward for the TD model as a composite of a reward and a stimulus (Suri & Schultz, 1999). With this approach, the reward prediction error signal is phasically



activated at reward onset after repeated reward presentations. We use the same approach in this study and provide a copy of the reward signal as an additional stimulus to the TD model.

The TD model processes two types of input: input that the model learns to predict (usually the reward) and input that serves as information for these predictions (usually the stimuli). The first event in a trial (usually a stimulus) serves as the information to learn prediction signals for the second event (usually the reward). Since behaviorally important stimuli are preceded by anticipatory neural activity (see Figure 1C), we want to compute prediction signals not only for rewards but also for stimuli. Therefore, we propose a model that does not distinguish between stimuli and rewards. We use

several TD models, each computing predictions for each event (stimulus or reward) that occurs in the simulated paradigm. The TD models receive all events as information input in order to compute optimal prediction signals (see Figure 2C). Since these TD models are independent, each of them is mathematically equal to the standard TD model, and all the simulation results (except Figure 5C) could be computed with the standard TD model.

4 Model Simulations

4.1 Pretraining with Rewards Alone. For the experimental situations, animals were typically familiar with the single rewards occurring in the experiments. Therefore, the model was always pretrained with 20 presentations of the single rewards alone. These single rewards were presented during 1 second. Reward presentations were separated by intervals that were long enough to prevent learning of associations between rewards.

4.2 Delayed-Response Task. We did not intend to reproduce anticipatory neural activity for all events of the delayed-response task but simulated a trial with a presentation of a stimulus followed after a delay by presentation of a reward. The duration of the instruction stimulus (1 sec) and the duration of the intratrial interval (5 sec) corresponded to similar durations in the monkey experiments. Trigger stimulus and the animal's movements were not modeled. Simulated intertrial intervals were long enough to avoid

Figure 2: *Facing page*. (A) Temporal stimulus representation. Each stimulus $u_l(t)$ is followed by a series of phasic signals $x_1(t), x_2(t), x_3(t), \ldots$ that cover trial duration. The first component of this temporal representation peaks with amplitude one (line 2), the second with amplitude δ (line 3), the third with amplitude δ^2 (line 4), and so on. Representation computed with the standard value $\delta = 1$ is shown (without decay of the temporal representation). (B) TD model for one stimulus and one reward (Sutton & Barto, 1990). For the stimulus $u(t)$ the temporal stimulus representation $x_1(t), x_2(t), x_3(t), \ldots$ is computed. Each component $x_m(t)$ is multiplied with an adaptive weight $V_m(t)$ (filled dots). The reward prediction $p(t)$ is the sum of the weighted representation components of all stimuli. The difference operator D takes temporal differences from this prediction signal (discounted with factor γ). The reward prediction error $e(t)$ reports deviations to the desired prediction signals. This error is minimized by incrementally adapting the elements of the weights $V_m(t)$ proportionally to the prediction error signal $e(t)$ and to the learning rate β. (C) Two TD models for two events $u_1(t)$ and $u_2(t)$. Each event signal $u_l(t)$ reports about a stimulus or a reward that occurs in the experimental paradigm. All events are modeled as a composite of a stimulus and a reward. Each temporal representation component $x_m(t)$ is multiplied with an adaptive weight V_{lm} (filled dots). The event prediction $p_l(t)$ is computed from the sum of the weighted components.

associations between trials. If only one reward type was delivered in the animal experiment, the model was trained with 20 trials in which stimulus A (instruction) was followed by reward B. If three different instruction stimuli preceded delivery of two different rewards, the model was trained with the corresponding three pairs of events (stimulus A $\rightarrow$ reward X, stimulus B $\rightarrow$ reward X, and stimulus C $\rightarrow$ reward Y). Each pair was presented 20 times. Tonic and phasic anticipatory activity was compared with prediction signals and prediction error signals, respectively. The temporal discount factor γ was chosen to approximate the time course of the anticipatory neural activity.

The chosen stimulus representation covered equally the whole interval between stimuli and rewards without "forgetting" the stimulus presentation. However, not all neurons may have access to such a complete temporal stimulus representation. We therefore examined the influence of an incomplete stimulus representation that decayed rapidly after presentation of stimuli. This was achieved by setting the value of the decay rate δ to 0.8 per 100 msec (see the legend to Figure 2A). Using this parameter value, the peaks of the stimulus representation components decreased 20% for each additional 100 msec stimulus-peak interval. This model with incomplete stimulus representation was trained according to the schedule with two rewards.

5 Results

During the pretraining with repeated presentations of the reward (see section 4), the time of reward onset was unpredictable, but the reward duration remained constant. After pretraining, the prediction signals correctly decreased during reward presentation, reflecting the remaining reward duration, and the prediction error signals increased phasically at the reward onset (see Figure 3A).

The model was trained with a stimulus A followed by a reward B (see Figure 3B). When the discount factor γ was set to the standard value of 0.99 per 100 msec (left), the prediction signals increased at onset of stimulus A and then progressively increased with a rate similar to the desired rate of 1% per 100 msec (see left, line 3). When the model was trained with the discount factor $\gamma = 0.85$, the prediction signals increased with a rate similar to the desired rate of 15% per 100 msec see (see right, line 3). For $\gamma = 0.99$, the interstimulus interval was too short to learn the desired prediction signals for times before presentation of stimulus A. Therefore, the prediction error was phasically increased at the onset of stimulus A (see bottom, left side). For $\gamma = 0.85$, the interstimulus interval was long enough to learn the desired prediction signals, which led to a very small prediction error signal (see bottom, right side).

Simulated reward prediction signals of the proposed model were comparable with anticipatory neural activity measured in the putamen (part of

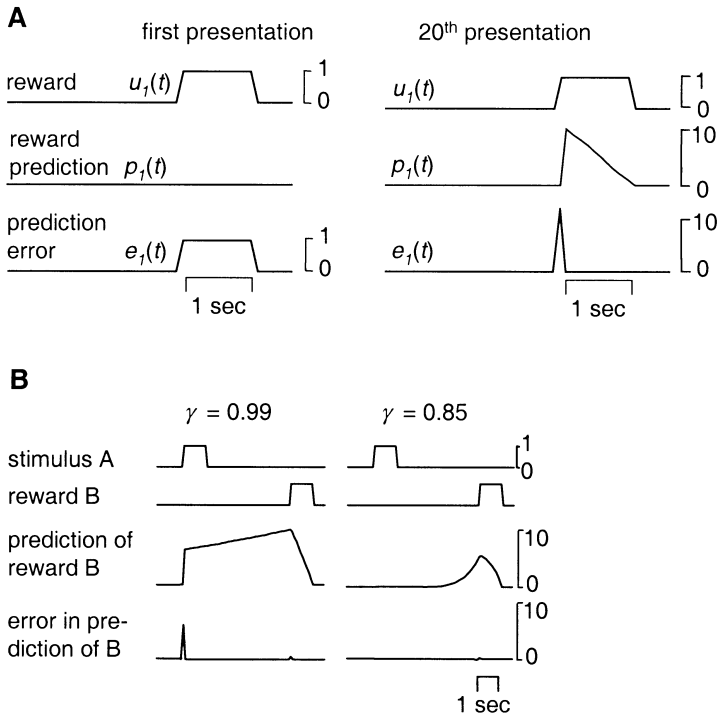


Figure 3: (A) Pretraining with a single reward. For the first presentation of a novel reward $u_1(t)$ of 1 sec duration (left side, line 1), the prediction signal $p_1(t)$ (left, line 2) was zero, as the weight matrix V_1m was initialized with zeros (see equation A.2). The prediction error signal $e_1(t)$ (left, line 3) was equal to the reward signal $u_1(t)$ (see equation A.3). After 20 presentations of this reward (right side), learning was completed and the duration of the reward $u_1(t)$ was correctly predicted (right side, line 2). The prediction signal decreased during the reward presentation, as it correctly reflected the remaining future reward duration. The prediction error signal (right side, line 3) increased phasically at the reward onset, as the reward onset was unpredictable. The discount factor was set to the value of $\gamma = 0.99$ per time step (1 time step = 100 msec). (B) After 20 presentations of stimulus A (line 1) followed by reward B (line 2). The model was trained with the standard value of 0.99 for the discount factor γ (left side) and the value of 0.85 (right side). Both prediction signals were learned correctly. The increases in the prediction signals were equal to the desired rates of 1% per 100 msec (left side) and 15% per 100 msec (right side). For $\gamma = 0.99$ (left side), the prediction error signals increased phasically at onset of stimulus A, as onset of stimulus A was unpredictable. For $\gamma = 0.85$ (right side), the prediction error signal was almost zero for all time steps, because the simulated prediction signals were close to the desired prediction signals. (This figure was computed without representation decay ($\delta = 1$).)

striatum), and reward prediction error signals were comparable with activity of midbrain dopamine neurons (see Figure 4). When the stimulus and the reward were presented in temporal succession for the first time (see Figure 4A), the reward prediction signal was not affected by presentation of the stimulus, as the stimulus was not associated with reward. At the time of the unpredicted reward, the reward prediction signal was increased (see Figure 4A, line 3). Similar to this simulated signal, activity of a subset of putamen neurons was not affected by presentation of the stimulus and was increased by the reward (see line 4). The simulated reward prediction error was phasically increased at the reward onset as the reward was unpredicted (see line 5). This signal was comparable to the activity of midbrain dopamine neurons (bottom).

Simulated signals were then compared with neural activities after learning (see Figure 4B). The simulated reward prediction signal was already increased at stimulus onset and progressively increased until reward onset (see line 3) as the stimulus predicted the reward. This signal correctly increased between the stimulus and the reward about 1% for each 100 msec, because the discount factor γ was set to 0.99 per 100 msec (standard value). The simulated reward prediction signal was comparable to reward anticipatory activity of a subset of striatal neurons (see line 4). The signal representing the reward prediction error was phasically increased by the unpredicted stimulus but not affected by the predicted reward (see line 5). This response to the stimulus was smaller than the response to the unpredicted reward before learning (compare Figure 4A, line 5) as the prediction signal (see Figure 4B, line 3) progressively increased according to the discount factor. The reward prediction error was comparable to dopamine neuron activity (see Figure 4B, bottom).

Simulated prediction signals were also comparable with reward-specific anticipatory activity recorded in orbitofrontal cortex (see Figure 5). Monkeys had been trained in a delayed-response task with three instruction stimuli, A, B, and C, followed by two different rewards, X and Y (see section 2). The model had been trained with the corresponding pairs of events (see section 4). In trials without occurrence of reward Y, prediction of reward Y was not affected (see Figure 5A, top, left, and middle). In trials with occurrence of reward Y, this prediction signal was activated when stimulus C was presented and then progressively increased until reward Y (see Figure 5A, top, right), because reward Y was completely predicted by stimulus C. Prediction of reward Y was comparable to reward-specific activity of a subset of orbitofrontal neurons anticipating reward Y but not reward X (see Figure 5A, bottom).

The model was trained with the same pairs of events, but the value of 0.95 per 100 msec was used for the temporal discount factor γ. Therefore, prediction signals increased more rapidly according to a correct rate of about 5% per 100 msec (see Figure 5B, top). Prediction of reward X was only slightly increased at the onset of stimuli A and B and then increased rapidly un-

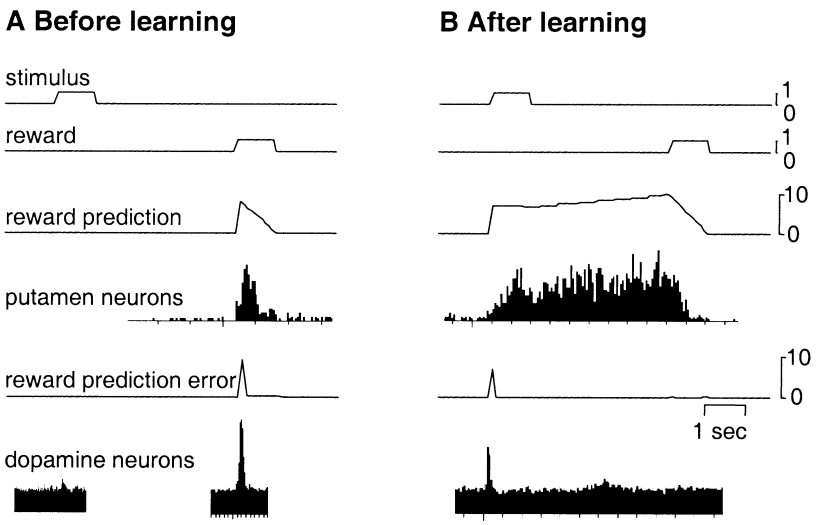
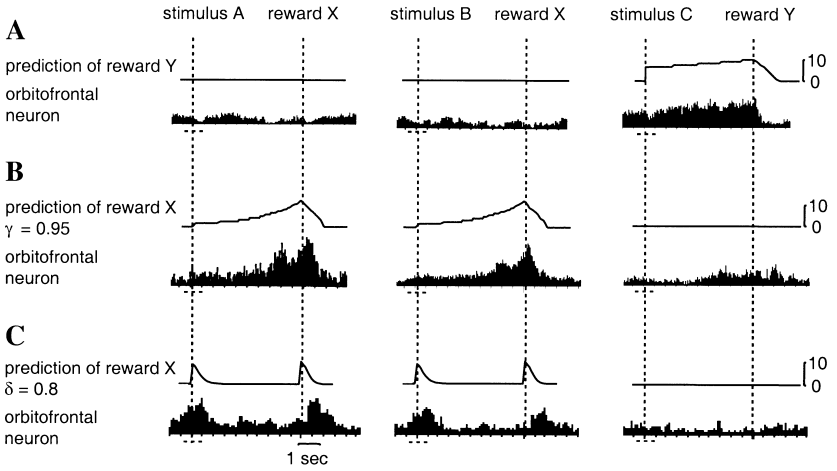


Figure 4: Comparable time courses of predictive signals and activity histograms for a stimulus (line 1) preceding a reward (line 2). The same timescale applies to simulated trajectories and histograms of neural activity. Activity histograms were selected from delayed-response tasks with the simulated stimulus corresponding to the instruction stimulus. Simulated signals were selected from the above simulations (see Figure 3A, right, and Figure 3B, left). (A) Before learning. A typical reward prediction signal was increased when the reward was presented (line 3). This signal was comparable to the activity histogram of a set of putamen neurons aligned to an unpredictable reward (line 4) (from Schultz, Apicella, Ljungberg, Romo, & Scarnati, 1993). A typical signal reflecting reward prediction errors was phasically increased at onset of the reward (line 5). This signal was comparable to dopamine neuron activity (bottom). These neurons do not respond to a small instruction light (from Ljungberg et al., 1992) but rather to an unpredictable drop of liquid reward delivered to the mouth of the animal (from Mirenowicz & Schultz, 1994). (B) After learning (20 stimulus-reward pairings). A typical reward prediction signal had already increased when the stimulus was presented and then progressively increased until occurrence of the reward (line 3). This reward prediction signal was comparable to anticipatory activity of a putamen neuron (line 4; from Apicella et al., 1992). This neural activity seems to anticipate the future reward, because it was also increased before the reward regardless of the instruction stimulus that indicated reward delivery. Of 1173 studied neurons, 6 striatal neurons showed similar sustained reward anticipatory activity lasting over the task duration (compare Figure 1B). A typical signal reflecting reward prediction errors was already phasically increased at the stimulus onset and on baseline level when the reward was presented (line 5). This signal was comparable to the activity of dopamine neurons, which respond after learning to a small instruction light but not to a predictable drop of liquid reward (bottom) (from Ljungberg et al., 1992).



til reward X (see top, left, and middle), because reward X was completely predicted by the stimuli A and B. Prediction of reward X was not affected in trials without reward X (see top, right side). Prediction of reward X was comparable to the activity of a subset of orbitofrontal neurons with activity anticipating reward X (see Figure 5B, bottom). Figures 5A and 5B demonstrate that simulated prediction signals and anticipatory neural activities discriminate between specific predicted rewards.

We studied the influence of a rapidly decaying stimulus representation on the prediction signal. After 20 presentations of the stimulus-reward pairs, prediction of reward X was activated by the stimuli A and B and by the reward X (see Figure 5C, top). Since learning was not completed, prediction of reward X was associated with the correct predictive stimuli A and B but did not bridge the time between the stimuli and the reward. The responses to the predictive stimuli A and B were learned with the representation traces, because the traces still bridged the time gap between the stimuli and the rewards. Prediction of reward X was comparable to the activity of a subset of orbitofrontal neurons anticipating reward X but not reward Y (see Figure 5C, bottom).

6 Discussion

This study demonstrates for the first time that an adaptive model can reproduce characteristics of anticipatory delay period activity. Each of the TD models learned a tonic reward-specific prediction signal using a phasic reward-specific prediction error signal. Since the reward prediction error signal was computed as in previous TD models (Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1999), characteristics of the reward prediction error signal were comparable to phasic activities of midbrain

dopamine neurons (see Figure 4). In addition, simulated reward-specific prediction signals were comparable to reward-specific anticipatory activity of subsets of cortical and striatal neurons before and after learning the contingency between a stimulus and a reward (see Figures 4 and 5). Variation of two model parameters, the discount factor and the decay rate of temporal stimulus representation reproduced variations in time courses of anticipatory neural activity (see Figure 5). Although the shown model simulations do not include learning between more than two events, the model can be

Figure 5: *Facing page*. Comparable time courses of prediction signals and orbitofrontal activities after learning pairings between different stimuli and rewards. Simulated stimuli are compared with the instruction stimuli in a delayed-response task. (Histograms reconstructed from Tremblay & Schultz, 1999, 2000; Schultz et al., 2000; see section 2. Durations of simulated stimuli, rewards, and intratrial interval as in Figure 4.) (A) Prediction of reward Y. In trials without reward Y, all signals reflecting prediction of reward Y were zero (top, left, and middle). When stimulus C preceded reward Y, the signal-reflecting prediction of reward Y was activated when stimulus C was presented and then progressively increased until reward Y (top, right side). Prediction of reward Y was comparable to the activity of an orbitofrontal neuron anticipating reward Y but not reward X (bottom). (In the histogram at bottom, right, neural activity before the task was larger than after the task, because the previous task already predicted reward Y.) (B) Prediction of reward X was learned with a discount factor $\gamma = 0.95$ per 100 msec. This signal slightly increased when stimuli A or B were presented and then increased rapidly until reward X (top, left, and middle). This signal was zero in trials without reward X (top, right). The prediction of reward X was comparable to the activity of an orbitofrontal neuron anticipating reward X but not reward Y (bottom). Nine percent of orbitofrontal neurons are active during delay periods before specific rewards, as shown in *A* or *B* (Tremblay & Schultz, 1999, 2000; Schultz et al. 2000). (C) Prediction of reward X with representation decay ($\delta = 0.8$ per 100 msec). In trials with presentations of reward X, a typical signal reflecting the prediction of reward X increased when the stimuli A and B or the reward X were presented (top, left and middle). Prediction of reward X did not bridge the intratrial interval, as the temporal event representation decayed rapidly. In the trial without presentation of reward X, prediction of reward X was zero (right). Prediction of reward X was comparable to the activity of an orbitofrontal neuron. This neuron responded to the stimuli A and B and to the reward X (bottom, left and middle). The activity did not bridge the intratrial interval. Activity was at baseline levels in trials without reward X (bottom, right). This neuron belongs to the subset with reward-specific responses to the instruction (15% of tested neurons) and to the subset with reward-specific responses following the reward (8% of tested neurons) (unpublished data from Tremblay & Schultz, 1999, 2000, and Schultz et al. 2000). (Activities were aligned to rewards. Horizontal broken lines below histograms indicate stimulus onsets.)

applied to experiments with more events. Anticipatory neural activity is usually influenced by only two events: a stimulus, which elicits the activity, and a reward, which terminates the activity. Therefore, striatal anticipatory neural activity in the delayed-response task could be reproduced with separate models that separately learn the pairs reward-instruction (see Figure 1B, top), instruction-trigger (see Figure 1B, middle), instruction-reward (see Figure 1B, bottom), and trigger-reward (see Figure 1B, bottom). This assumption of separate networks is plausible, as networks of biological neurons are usually partially connected. For the sequence reproduction task (Kermadi & Joseph, 1995), an elaborated and biologically plausible stimulus representation has been proposed that reproduces order- and stimulus-dependent neural activities (Dominey et al., 1995). As the temporal characteristics of anticipatory activities resemble those of the simulated prediction signals (compare Figure 1C with Figure 3B, right), these anticipatory neural activities could probably be reproduced by training a TD model with an elaborated internal representation.

Associative weights involved in the computation of tonic prediction signals were adapted according to phasic signals reporting prediction errors. Therefore, the model suggests that phasic anticipatory activities induce long-term adaptations of neurons with tonic anticipatory activities. Consistent with evidence for dopamine-dependent long-term adaptation of corticostriatal transmission (Calabresi, Pisani, Mercuri, & Bernardi, 1992; Calabresi et al., 1997; Wickens, Begg, & Arbuthnott, 1996), the model suggests that activity of midbrain dopamine neurons leads to long-term adaptations of cortical or striatal neurons with tonic reward-anticipating activity. Furthermore, the model postulates the existence of a category of neurons that are phasically active, report errors in predictions of specific rewards, and induce long-term adaptations of neurons with tonic reward-specific anticipatory activity. Although phasic context-dependent neural activities in striatum and prefrontal cortex have been reported (see Schultz & Romo, 1992), it has not been investigated if these activities anticipate rewards.

When the model was trained using an incomplete temporal stimulus representation, the prediction signal did not progressively increase before the predicted reward but instead decreased to zero in the intratrial interval. This time course was similar to some time courses of anticipatory activity (see Figure 5C). This finding suggests that neurons with anticipatory neural activity that is on baseline levels during the intratrial interval do not have access to the complete temporal stimulus representation. If a series of distinguishable stimuli were presented during the trial, these stimuli would serve as a complete temporal stimulus representation, and the model would learn the progressively increasing prediction signals. This suggests that anticipatory neural activity would reveal their optimal time course when measured in an experiment with a series of stimuli that precede the anticipated reward. Such an experiment would test our basic model assumptions.

The reward prediction error signal of the TD model reproduces dopamine neuron activity in many situations (see section 1). Unfortunately, the TD model fails to reproduce dopamine neuron activity when a reward is delivered earlier than expected (Hollerman & Schultz, 1998). This inconsistency has been corrected with a temporal event representation that is influenced by subsequent events (Suri & Schultz, 1999). For situations with delayed or omitted rewards, prediction signals of the proposed TD model decrease to zero when the reward is expected. These prediction signals resemble some, but not all, tonic anticipatory activity for delayed reward presentation (Hikosaka et al., 1989). These subtle inconsistencies between simulated signals and measured anticipatory activity suggest that some components of the temporal event representation are influenced by subsequent events (Suri & Schultz, 1999).

The proposed model can be partially related to networks of biological neurons. The model learns to predict stimuli and rewards. It has been suggested that reward predictions are learned in limbic parts of pathways from cortex via striatum to midbrain dopamine neurons (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Brown et al., 1999; Suri & Schultz, 1999). In contrast, stimulus predictions may be learned predominantly in the cortex. For visual stimuli, it has been proposed that pyramidal neurons in higher cortical areas learn to predict neural activity of pyramidal neurons in lower cortical areas (Rao & Ballard, 1997, 1999). These studies suggest that the feedforward connections to higher cortical areas carry the prediction errors, whereas the feedback connections carry the prediction signals.

Anticipatory neural activity may influence the behavior of the animal by several mechanisms. Neural activity preceding specific reinforcers may lead to anticipatory responses in Pavlovian paradigms. Reward anticipatory activity was suggested to be used for learning (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Suri & Schultz, 1999) or planning (Suri, Bargas, & Arbib, submitted), whereas stimulus anticipation activity may be used as a predictive representation to learn prediction-reward associations (Dayan, 1993) or to shorten the reaction time to anticipated situations (Goldberg & Bruce, 1990).

Appendix A: TD Model

Since the model does not distinguish between stimuli and rewards, both are here referred to as events. We give the model equations for a paradigm with L different events and therefore L independent TD models ($L = 2$ for Figure 2C). The event signal $u_l(t)$ reports the presence ($u_l(t) = 1$) and absence ($u_l(t) = 0$) of the event with number l ($l = 1, \ldots, L$). The desired prediction signal for this event $p_{desired,l}(t)$ is defined as the discounted sum of this future event $u_l(t)$

$$p_{desired,l}(t) = u_l(t) + \gamma u_l(t+1) + \gamma^2 u_l(t+2) + \cdots \tag{A.1}$$

(standard value of discount factor $\gamma = 0.99$ per time step, 1 time step = 100 msec). The proposed model consists of the equations below (equations A.2–A.5), which allow estimating the desired prediction signals $p_{desired,l}(t)$.

Events $u_l(t)$ are represented with the temporal representation $x_m(t)$ as shown in Figure 2A. The first event $u_1(t)$ is represented in the representation components $x_1(t), x_2(t), \ldots, x_N(t)$, the second event in the representation components $x_{N+1}(t), x_{N+2}(t), \ldots, x_{2N}(t)$, and so on. In order to cover trial durations of 7 seconds with the event representation, each event is represented with 70 phasic representation components ($N = 70$; $N = 3$ in Figure 2 is only for illustration). We estimate a weight matrix V_{lm} (for L events $l = 1, \ldots, L$, and $m = 1, \ldots, N \times L$) that computes the prediction signal $p_l(t+1)$ from the temporal representation with

$$p_l(t+1) = \sum_{m=1}^{N \times L} V_{lm}(t) x_m(t+1). \tag{A.2}$$

As it follows from equation A.1 that $p_{desired,l}(t) = u_l(t) + \gamma p_{desired,l}(t+1)$, the error $e_l(t)$ between the estimated prediction $p_l(t)$ and the desired prediction $p_{desired,l}(t)$ is computed from discounted temporal differences between successive predictions (difference operator D in Figures 2B and 2C) and from the event $u_l(t)$ with

$$e_l(t) = u_l(t) + \gamma p_l(t+1) - p_l(t). \tag{A.3}$$

The weight matrix V_{lm} is initiated with zeros and then adapted with the two-factor learning rule

$$V_{lm}(t+1) = V_{lm}(t) + \beta e_l(t) x_m^T(t), \tag{A.4}$$

where the learning rate β was set to the value of 50. This value is larger than that of usual learning rates since the desired prediction signals are much larger than one. The trace $x_m^T(t)$ is a slowly decaying version of the temporal representation component $x_m(t)$,

$$x_m^T(t) = \lambda x_m^T(t-1) + (1-\lambda) x_m(t), \text{ with } x_m^T(0) = 0. \tag{A.5}$$

The traces decrease 0.3% each time step (1 time step = 100 msec, $\lambda = 0.997$), as this produces fast learning. The intertrial interval was long enough for all eligibility traces to decrease to zero, which was simulated by setting the traces to zero.

In contrast to the proposed algorithm, the TD model computes only one prediction signal ($L = 1$) but still uses all stimuli in the trial to compute this signal (the sum in equation A.2 is over $m = 1, 2, \ldots, N \times$ *number of stimuli*).

Acknowledgments

We thank Leon Tremblay for permission to use neural activities he recorded in orbitofrontal cortex (Tremblay & Schultz, 1999, 2000; Schultz et al., 2000). This study was supported by the James S. McDonnell Foundation grant 94-39.

References

Alexander, G. E., & Crutcher, M. D. (1990a). Preparation for movement: Neural representations of intended direction in three motor areas of the monkey. *J. Neurophysiol., 64*, 133–163.

Alexander, G. E., & Crutcher, M. D. (1990b). Neural representation of the target (goal) of visually guided arm movements in three motor areas of the monkey. *J. Neurophysiol., 64*, 164–178.

Apicella, P., Scarnati, E., Ljungberg, T., & Schultz, W. (1992). Neuronal activity in monkey striatum related to the expectation of predictable environmental events. *J. Neurophysiol., 68*, 945–960.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Blair, H. T., Lipscomb, B. W., & Sharp, P. E. (1997). Anticipatory time intervals of head-direction cells in the anterior thalamus of the rat: Implications for path integration in the head-direction circuit. *J. Neurophysiol.*, 78(1), 145–159.

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. *J. Neurosci., 19*(23), 10502–10511.

Calabresi, P., Pisani, A., Mercuri, N. B., & Bernardi, G. (1992). Long-term potentiation in the striatum is unmasked by removing the voltage-dependent magnesium block of NMDA receptor channels. *Europ. J. Neurosci., 4*, 929–935.

Calabresi, P., Saiardi, A., Pisani, A., Baik, J. H., Centonze, D., Mercuri, N. B., Bernardi, G., & Borrelli, E. (1997). Abnormal synaptic plasticity in the striatum of mice lacking dopamine D2 receptors. *J. Neurosci., 17*(12), 4536–4544.

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

Desmond, J. E., & Moore, J. W. (1988). Adaptive timing in neural networks: The conditioned response. *Biol Cybern, 58*(6), 405–415.

Desmond, J. E., & Moore J. W. (1991). Altering the synchrony of stimulus trace processes: Tests of a neural-network model. *Biol Cybern., 65*(3), 161–169.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge: Cambridge University Press.

Dominey, P., Arbib, M., & Joseph, J.-P. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *J. Cognitive Neurosci., 7*(3), 311–336.

Duhamel, J. R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science, 255*(5040), 90–92.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Goldberg, M. E., & Bruce, C. J. (1990). Primate frontal eye fields. III. Maintenance of a spatially accurate saccade signal. *J. Neurophysiol., 64*(2), 489–508.

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

Hikosaka, O., Sakamoto, M., & Usui, S. (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J. Neurophysiol., 4*(4), 814–832.

Hollerman, J.R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1*(4), 304–309.

Hollerman, J. R., Tremblay, L., & Schultz, W. (1998). Influence of reward expectation on behavior-related neuronal activity in primate striatum. *J. Neurophysiol., 80*(2), 947–963.

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

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

Kermadi, I., & Joseph, J. P. (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *J. Neurophysiol., 74*, 911–933.

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

Mackintosh, N. M. (1974). *The psychology of animal learning*. London: Academic Press.

Mauritz, K. H., & Wise, S. P. (1986). Premotor cortex of the rhesus monkey: Neuronal activity in anticipation of predictable environmental events. *Exp. Brain Res., 61*(2), 229–244.

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

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci., 16*(5), 1936–1947.

Nakamura, K., & Colby, C. L. (1999). Updating of the visual representation in monkey striate and extrastriate cortex during saccades. *Soc. Neurosci. Abstr., 25*(1), 1163.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

Rao, R. P., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput., 9*(4), 721–763.

Rao, R.P., & Ballard, D.H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci., 2*(1), 79–87.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts.

Romo, R., & Schultz, W. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements. III. Neuronal activity in the supplementary motor area. *Exp. Brain Res., 91*, 396–407.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol., 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. *J. Neurosci., 13*, 900–913.

Schultz, W., Apicella, P., Ljungberg, T., Romo, R., & Scarnati, E. (1993). Reward-related activity in the monkey striatum and substantia nigra. *Progress in Brain Research, 99*, 227.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience, 12*(12), 4595–4610.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Schultz, W., & Romo, R. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements: I. Preparatory activity in the anterior striatum. *Exp. Brain Res., 91*, 363–384.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cereb. Cortex, 10*(3), 272–284.

Suri, R. E., Bargas, J., & Arbib, M. A. Modeling functions of striatal dopamine modulation in learning and planning. Submitted.

Suri, R. E., & Schultz, W. (1998). Dopamine-like reinforcement signal improves learning of sequential movements by neural network. *Exp. Brain Res., 121*, 350–354.

Suri, R. E., & Schultz, W. (1999). A neural network learns a spatial delayed response task with a dopamine-like reinforcement signal. *Neuroscience, 91*(3), 871–890.

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. 539–602). Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press. Available online at: http://www-anw.cs.umass.edu/~rich/book/the-book.html.

Tremblay, L., Hollerman, J. R., & Schultz, W. (1998). Modifications of reward expectation-related neuronal activity during learning in primate striatum. *J. Neurophysiol., 80*(2), 964–977.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature, 398*(6729), 704–708.

Tremblay, L., & Schultz, W. (2000). Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *J. Neurophysiol., 83*(4), 1864–1876.

Umeno, M. M., & Goldberg, M. E. (1997). Spatial processing in the monkey frontal eye field. I. Predictive visual responses. *J. Neurophysiol., 78*(3), 1373–1383.

Wagner, A. R. (1978). Expectancies and the priming of STM. In S. H. Hulse, W. Fowler, & W. K. Honig (Eds.), *Cognitive processes in animal behavior* (pp. 177–210). Hillsdale, NJ: Erlbaum.

Walker, M. F., Fitzgibbon, E. J., & Goldberg, M. E. (1995). Neurons in the monkey superior colliculus predict the visual result of impending saccadic eye movements. *J. Neurophysiol., 73*(5), 1988–2003.

Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature, 382*(6592), 629–632.

Wickens, J. R., Begg, A. J., & Arbuthnott, G. W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex *in vitro*. *Neuroscience, 70*(3), 1–5.

Received September 9, 1998; accepted July 15, 2000.