TD models of reward predictive responses in dopamine neurons

Journal Article

Author(s):
Suri, Roland Erwin

Publication date:
2002

Permanent link:
https://doi.org/10.3929/ethz-a-007587140

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
Neural networks : the official journal of the International Neural Network Society, European Neural Network Society, Japanese Neural Network Society 15(4-6), https://doi.org/10.1016/S0893-6080(02)00046-1

PERGAMON

Neural Networks xx (0000) xxx–xxx

Neural Networks

www.elsevier.com/locate/neunet

TD models of reward predictive responses in dopamine neurons

Roland E. Suri*

Computational Neurobiology Laboratory, The Salk Institute, San Diego, CA 92186, USA

Received 10 October 2001; revised 16 April 2002; accepted 16 April 2002

Abstract

This article focuses on recent modeling studies of dopamine neuron activity and their influence on behavior. Activity of midbrain dopamine neurons is phasically increased by stimuli that increase the animal's reward expectation and is decreased below baseline levels when the reward fails to occur. These characteristics resemble the reward prediction error signal of the temporal difference (TD) model, which is a model of reinforcement learning. Computational modeling studies show that such a dopamine-like reward prediction error can serve as a powerful teaching signal for learning with delayed reinforcement, in particular for learning of motor sequences.

Several lines of evidence suggest that dopamine is also involved in 'cognitive' processes that are not addressed by standard TD models. I propose the hypothesis that dopamine neuron activity is crucial for planning processes, also referred to as 'goal-directed behavior', which select actions by evaluating predictions about their motivational outcomes. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Temporal difference; Reinforcement; Neuromodulation; Sensorimotor; Prediction; Planning

1. Introduction

In a famous experiment Pavlov (1927) trained a dog with the ringing of a bell (stimulus) that was 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 suggesting that the salivation response elicited by the bell ring reflects anticipation of food delivery. A large body of experimental evidence led to the hypothesis that such Pavlovian learning is dependent upon the degree of the unpredictability of the reinforcer (Rescorla & Wagner, 1972; Dickinson, 1994). 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 a reinforcement prediction error signal to learn a reinforcement prediction signal. The reinforcement prediction error signal progressively decreases when the reinforcement prediction signal becomes similar to the desired reinforcement prediction signal. Characteristics of the reinforcement prediction signal are comparable to those of anticipatory responses such as salivation in Pavlov's experiment and may guide approach behavior (Montague, Dayan, Person, & Sejnowski, 1995).

In Pavlov's experiment, the salivation response of the dog does not influence the food delivery. Consequently, the TD model computes predictive signals but does not select optimal actions. In contrast, instrumental learning paradigms, such as learning to press a lever for food delivery, demonstrate that animals are able to learn to perform actions that optimize reinforcement. To model sensorimotor learning in such paradigms, a model component called the Actor is taught by the reward prediction error signal of the TD model. In such architectures, the TD model is also called the Critic. This approach is consistent with animal learning theory (Dickinson, 1994) and was successfully applied to machine learning studies (Sutton & Barto, 1998).

The reinforcement prediction error signal of the TD model remained a purely hypothetical signal until researchers discovered that the activity of midbrain dopamine neurons in substantia nigra and ventral tegmental area is strikingly similar to the reward prediction error of the TD model (Montague, Dayan, & Sejnowski, 1996; Schultz, 1998; Suri & Schultz, 1999). Midbrain dopamine neurons project to striatum and cortex and are characterized by rather uniform

* Address: Intelligent Optical Systems (IOS), 2520 W 237th Street, Torrance, CA 90505-5217, USA. Tel.: +1-310-530-71-30x108; fax: +1-310-530-74-17.

E-mail address: rsuri@intopsys.com (R.E. Suri).

0893-6080/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0893-6080(02)00046-1

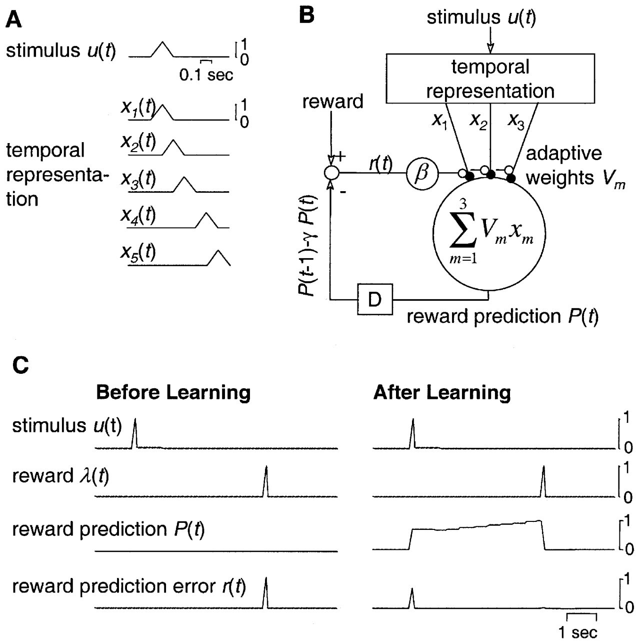

Fig. 1. (A) Temporal stimulus representation. A stimulus $u(t)$ is represented as a signal that is one during presentation of this stimulus and zero otherwise. The temporal stimulus representation of this stimulus $u(t)$ consists of a series of phasic signals $x_1(t)$, $x_2(t)$, $x_3(t)$,… that cover trial duration (only three components are shown). Each component of this temporal representation peaks with amplitude one and is zero otherwise. (B) Scheme of TD model for one stimulus followed by a reward (scheme adapted from Suri & Schultz, 2001). For the stimulus $u(t)$ the temporal stimulus representation $x_1(t)$, $x_2(t)$, $x_3(t)$,… 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 TDs from this prediction signal (discounted with factor γ). The reward prediction error $r(t)$ 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 $r(t)$ and to the learning rate β. (C) Signals of the TD model for a stimulus followed by a reward. Left Before learning, all weights V_m initialized with the value zero. As the reward prediction signal (line 3) is zero, the reward prediction error (line 4) is increased to the value of one when the reward is presented. Right After learning (20 stimulus–reward pairings). The reward prediction signal already increases when the stimulus is presented (line 1) and then progressively increases until the occurrence of the reward (line 2). The slope of the progressive increase is determined by the discount factor γ. Since its value is set to 0.99, the reward prediction increases with a rate of 1% per 100 ms. The reward prediction error is already phasically increased when the stimulus occurs and at baseline levels when the reward is presented.

responses throughout the whole neuron population of midbrain dopamine neurons. Comparison of the Actor–Critic architecture to biological structures suggests that the Critic may correspond to pathways from limbic cortex via limbic striatum to dopamine neurons, whereas the Actor may correspond to pathways from neocortex via sensorimotor striatum to basal ganglia output nuclei.

The Actor–Critic model with the standard TD model as the Critic mimics learning of sensorimotor associations or habits. Since this standard Actor–Critic model is not able to solve tasks that require planning, animal learning and machine learning theorists extended the Critic to an internal model approach (Balleine & Dickinson, 1998; Dickinson, 1994; Sutton & Barto, 1998). Several lines of evidence suggest that dopamine neuron activity may be reproduced more accurately by using such an extended TD model as the Critic than by using the standard TD model (Suri, 2001). This hypothesis is consistent with experimental evidence suggesting that dopamine neuron activity may not only be involved in sensorimotor learning but also in planning (Lange et al., 1992).

2. Temporal difference (TD) model

The TD algorithm is popular in machine learning studies and was proven to converge to the optimal solution (Dayan & Sejnowski, 1994). Despite these successes, their development was strongly influenced by studies of animal learning (Sutton & Barto, 1990, 1998). Since animals often learn to estimate the time of the reward occurrence in Pavlovian learning paradigms, the TD model uses a time estimation mechanism (Sutton & Barto, 1990). This time estimation mechanism is implemented using a temporal stimulus representation, which consists of a large number of signals $x_m(t)$ for each stimulus. Each of these signals $x_m(t)$ has a value of one for one time point and is zero for all other times. Exactly one signal of the temporal stimulus representation $x_m(t)$ peaks for each time step of the period between the stimulus and the trial end (Fig. 1(A)). Similar hypothetical temporal stimulus representations have also been referred to as 'complete serial compound stimulus' (Sutton & Barto, 1990) or 'spectral timing mechanism' (Brown, Bullock, & Grossberg, 1999). A temporal stimulus representation is necessary to reproduce the depression of dopamine activity below baseline levels at the time when an expected reward is omitted, since this reflects a timing mechanism (Montague et al., 1996; Schultz, 1998). Its physiological correlate may include metabotropic glutamate receptor-mediated Ca2 + spikes occurring with different delays in striosomal cells of the striatum (Brown et al., 1999). The shape of these signals is not important for the algorithm, but the number of signals has to be sufficiently large to cover the duration of the intratrial interval ($m = 1, 2, \ldots, 50$ for 5 s interstimulus interval with time steps of 100 ms). The reward prediction $P(t)$ is computed as the weighted sum over the temporal stimulus representation signals $x_m(t)$ with

$$P(t) = \sum_{m=1}^{50} V_m(t) x_m(t).$$

The algorithm is designed to learn a 'desired' prediction signal that increases successively from one time step to the next by a factor $1/\gamma$ until the reward $\lambda(t)$ occurs and decreases to the baseline value of zero after the reward

Table 1
List of symbols

Symbol	Comments
Time t	Discretized in 100 ms time steps
Reward prediction error $r(t)$	Resembles dopamine neuron activity
Reward $\lambda(t)$	Signal is one when reward is present and zero when reward is absent
Temporal discount factor γ	$= 0.99/100$ ms estimated for dopamine neuron activity
Prediction $P(t)$	Resembles anticipatory behavior and anticipatory neural activity in cortex and striatum
Adaptive weights $V_m(t)$	Long-term memory storage
Component $x_m(t)$	Component of temporal stimulus representation
Learning rate β	Small constant
Stimulus $u(t)$	Signal is one when stimulus is present and zero when stimulus is absent

presentation. The prediction error signal is computed with

$$r(t) = \lambda(t) + \gamma P(t) - P(t-1)$$

and is zero as long as the prediction signal is equal to the desired prediction signal and nonzero otherwise. Since one time step corresponds to 100 ms, $t-1$ is a short hand for $t-100$ ms. The value of a discount factor γ is set between zero and one (Table 1).

The adaptive weights $V_m(t)$ are initialized with the value zero and adapted according to the learning rule

$$V_m(t) = V_m(t-1) + \beta r(t) x_m(t-1),$$

with a small learning rate constant β (Table 1). The TD model can be represented with a neuron-like element whose weights $V_m(t)$ correspond to synaptic conductances (Fig. 1(B)).

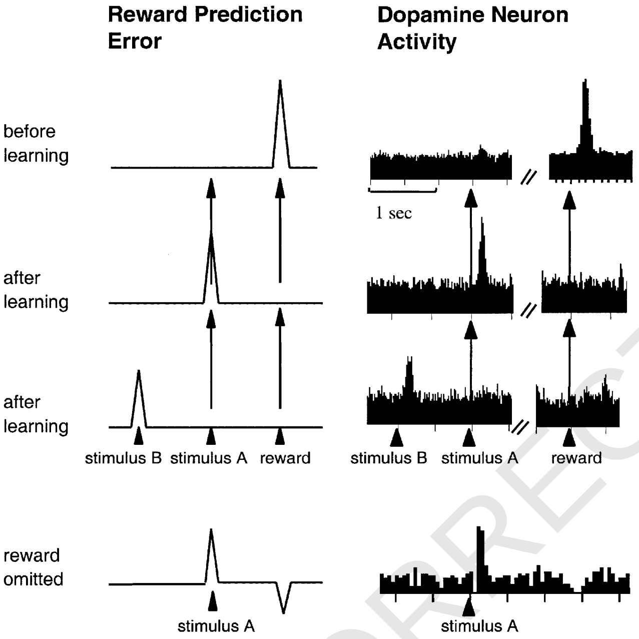

Fig. 2. Prediction error signal of the TD model (left) similar to dopamine neuron activity (right) (figure adapted from Suri & Schultz, 1998; discount factor $\gamma = 0.98$). If a neutral stimulus A is paired with reward, prediction error signal and dopamine activity respond to the reward (line 1) (activities reconstructed from Ljungberg et al., 1992; Mirenowicz & Schultz, 1994). After repeated pairings, the prediction error signal and dopamine activity are already increased by stimulus A and on baseline levels at the time of the reward (line 2). After training with an additional stimulus B, which precedes stimulus A, prediction error signal and dopamine activity are increased by stimulus B and neither affected by stimulus A nor by the reward (line 3). If the stimulus A is conditioned to a reward but is occasionally presented without reward, the prediction error signal and dopamine activity are decreased below baseline levels at the predicted time of reward (line 4). (Activities lines 2–4 reconstructed from Schultz, Apicella, & Ljungberg, 1993).

When the stimulus is followed by the reward for the first time, the reward prediction is zero and the reward prediction error is phasically increased at the time of the reward (Fig. 1(C)). After repeated presentations of the stimulus followed by the reward, the reward prediction increases before the anticipated reward. Characteristics of this reward prediction signal resemble those of reward anticipatory behaviors of animals (Sutton & Barto, 1990). The rate of this gradual increase is determined by the constant γ, which is referred to as the temporal discount factor. We use the value $\gamma = 0.99$ per 100 ms, which leads to an increase in the prediction signal of 1% for each 100 ms. The reward prediction error signal is at the time of the stimulus equal to the change in the reward prediction. Since dopamine responses decrease proportionally to the learned duration of the interval between the stimulus and the reward, dopamine neuron activity was used to estimate the value of the discount factor (Suri & Schultz, 1999). At the time of the reward, the reward prediction error is zero because the change in the prediction signal cancels out the reward signal.

3. TD error resembles dopamine neuron activity

The prediction error signal of the TD model is strikingly similar to activities of midbrain dopamine neurons (Montague et al., 1996; Schultz, 1998; Suri & Schultz, 1999). The prediction error signal is phasically increased by unpredicted reward and by the earliest reward-predicting stimulus, and it is negative when a predicted reward is omitted (Fig. 2, left). This signal closely resembles dopamine responses (Fig. 2, right). The depression in dopamine activity below baseline levels at the time of the predicted but omitted reward reflects a central timing mechanism because no stimulus is present at the time of the omitted reward.

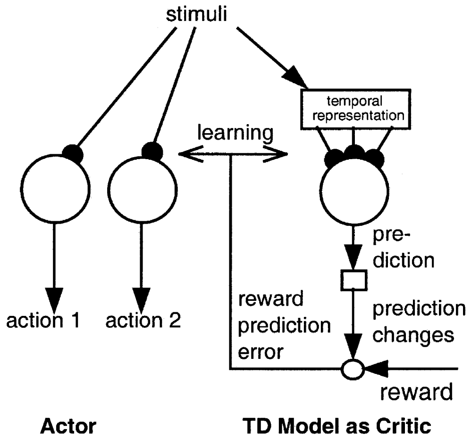

Fig. 3. Model architecture consisting of the Actor (left) and the Critic (right). The dopamine-like reward prediction error signal serves to modify the synaptic weights of the Critic itself and the synaptic weights of the Actor (heavy dots). Actor (left side). The Actor learns with the prediction error signal to associate stimuli with actions. Every Actor neuron (large circles) represents a specific action. Critic (right side). The dopamine-like reward prediction error is computed by the TD model (shown in Fig. 1) and serves as a teaching signal for the Actor.

The reward prediction error signal of the TD model by Suri & Schultz (1999) reproduces dopamine neuron activity in the situations: (1) upon presentation of unpredicted rewards, (2) before, during, and after learning that a stimulus precedes a reward, (3) when two stimuli precede a reward with fixed time intervals, (4) when the interval between the two stimuli are varied, (5) in the case of unexpectedly omitted reward, (6) delayed reward, (7) reward earlier than expected (Hollerman & Schultz, 1998), (8) in the case of unexpectedly omitted reward-predictive stimulus, (9) in the case of a novel, physically salient stimulus that has never been associated with reward (see allocation of attention, below), (10) and for the blocking paradigm (Waelti, Dickinson, & Schultz., 2001). To reach this close correspondence, three constants of the TD model were tuned to characteristics of dopamine neuron activity (learning rate, decay of eligibility trace, and temporal discount factor), some weights were initialized with positive values to achieve (9), and some ad hoc changes of the TD algorithm were introduced to reproduce (7) (see Discussion).

4. Actor–Critic architecture

To learn the actions that optimize the reward, the reward prediction error signal of the TD model teaches sensorimotor associations to the Actor (Fig. 3). A major computational benefit of learning with the dopamine-like reward prediction error signal as compared to learning with the reward signal is that the reward prediction error signal reports earlier about the task outcome than the reward signal. Indeed, machine learning studies demonstrate that TD algorithms serve as powerful approaches to solve reinforcement learning problems with delayed reinforcement (Barto, Sutton, & Anderson, 1983; Sutton & Barto, 1998; Tesauro, 1994). Examples for tasks with delayed rewards are board games such as backgammon. In such games the TD reward prediction signal codes for the chance to win and serves as the value of the board situation. A nonzero reward prediction error codes for surprising changes in the value of the board situation. If a player would learn only at the end of the game, corresponding to reinforcement learning with unconditional reinforcement, it would be unclear which sensorimotor associations between board situation and action should be adapted. However, if learning uses a TD prediction error signal, prediction errors of the estimated outcome can be used for learning: learning occurs during the game whenever the predicted outcome changes. Indeed, TD learning studies demonstrate that this strategy can be used to learn backgammon (Tesauro, 1994). For such machine learning applications, an Actor network is not necessary since the number of legal moves for each board situation is small. Instead, the algorithm computes the TD reward predictions for the board situations that would occur after all legal half-moves and executes the half-move that leads to the situation with the highest reward prediction. However, for applications with a large numbers of actions (or half-moves, respectively), it is advantageous to use an Actor network that is taught by the prediction error signal of the TD Critic (Barto et al., 1983). Simulations with the latter variant show that dopamine-like prediction error signals can serve as powerful teaching signals for acquiring behavioral tasks (Friston, Tononi, Reeke, Sporns, & Edelman, 1994; Montague et al., 1996; Nakahara, Doya, & Hikosaka, 2001; Suri & Schultz, 1998).

5. Learning of sequences

Disorders of dopamine transmission typically impair serially ordered movements in human patients (Phillips, Bradshaw, Iansek, & Chiu, 1993). Since TD learning with Actor–Critic architectures is particularly powerful for learning action sequences (Sutton & Barto, 1998), this finding is consistent with the hypothesis that dopamine neuron activity serves as a predictive teaching signal in a biological architecture resembling the Actor–Critic architecture. To demonstrate the capability of Actor–Critic architectures to learn sequences with a dopamine-like reinforcement signal, an Actor–Critic model is trained to learn a sequence of seven actions. Since only one action out of seven actions is correct, only one out of $7^7 = 823,543$ sequences is rewarded. The Actor consists of seven neuron-like elements. After each correct action, a stimulus is presented and the Actor–Critic model has to select the next correct action. The model is trained in seven phases, with 100 trials each phase. Training starts with the stimulus–action pair closest to the reward and then the sequence length is increased in every training phase by one

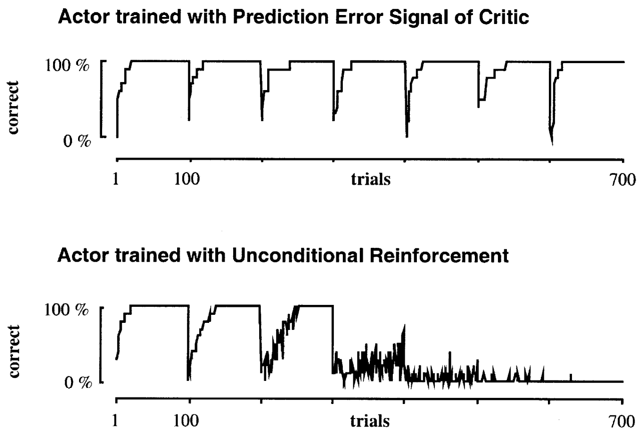

Fig. 4. Learning curves for training a sequence of seven stimulus–action associations (figure adapted from Suri & Schultz, 1998). Every 100 trials a novel additional stimulus–action pair is added to the sequence. Mean proportions of correct trials for learning 10 sequences are presented. (Top) Training with the prediction error signal results in a minimum number of incorrect trials. (Bottom) When trained with unconditional reinforcement signal only three stimulus - action associations are learned.

stimulus–action pair. Correct actions are followed by the presentation of the sequence learned in the previous phase. Incorrect actions terminate the trial.

Learning with the TD prediction error signal is compared to learning with the reward signal (learning rate $\beta = 0$ in TD Critic). With the adaptive prediction error signal, the sequence of seven actions is quickly learned (Fig. 4, top). In contrast, with the reward signal serving as the reinforcement signal only the first three actions of the sequence are learned (Fig. 4, bottom), demonstrating the advantage of learning with a dopamine-like reinforcement signal (Suri & Schultz, 1998).

If the reward signal serves as the reinforcement signal, learning does not occur without reward, and therefore once learned actions are repeated even if they are not rewarded any longer. With such an unconditional reinforcement signal, there is no mechanism for unlearning previously learned actions when the reward is omitted. In contrast, if a dopamine-like reward prediction error is used for learning, the probability of actions that have once been rewarded but are not rewarded any longer progressively decreases. This extinction of a previously learned action happens due to the depression of dopamine neuron activity at the time of the omitted reward (Suri & Schultz, 1999). This suggests that decreased adaptation of dopamine activity could lead to perseveration. Indeed, perseveration is a cognitive symptom of Parkinsonian patients (Lees & Smith, 1983). In addition, the influence of the reward prediction error on the Actor is investigated by setting this signal to a constant value below zero. This leads to extinction of previously learned actions, which resembles the extinction of previously learned lever-pressing in animals after being systemically injected with the dopamine receptor-blocking agent pimozide (Mason, Beninger, Fibiger, & Phillips, 1980) and may mimic the bradykinesia (slow movements) of Parkinsonian patients (Phillips et al., 1993).

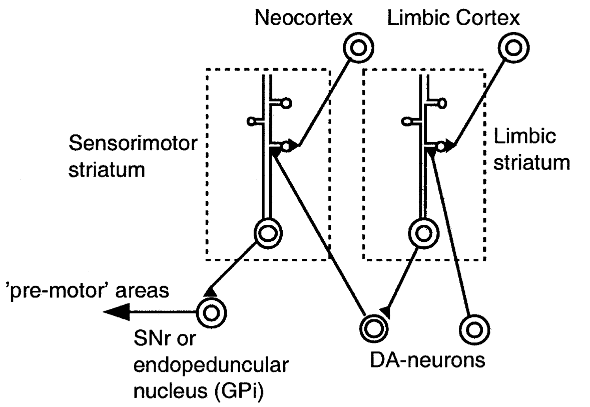

Fig. 5. Anatomical circuits that link the striatum with midbrain dopamine neurons (figure adapted from Smith & Bolam, 1990). The limbic striatum (presumably striosomes) may gate the flow of information through the sensorimotor striatum (presumably matrisomes) via midbrain dopamine neurons. These circuits closely resemble the Actor–Critic architecture (Fig. 3). Stimuli may be represented in cortical areas, the Actor may correspond to the sensorimotor striatum and motor output structures, and the Critic may correspond to the limbic striatum and dopamine neurons. The prediction signal of the TD model resembles the activity of a subset of neurons in the limbic striatum and the prediction error signal resembles dopamine neuron activity.

6. Biological correlates of the Actor–Critic model

Several characteristics of Actor–Critic architecture (see Fig. 3) resemble those of anatomical circuits (Fig. 5). (1) The neural activity of subgroups of neurons in the striatum resemble the reward prediction signal of the TD model (see Section 7). The reward prediction may be learned in the limbic striatum, which receives projections from dopamine neurons. (2) Convergence of information from extended representations to compute the reward prediction error is advantageous for the TD model. Convergence from extended sensory representations to a smaller number of actions is also typical for Actor networks (Barto et al., 1983). Indeed, there is a strong convergence from striatum to basal ganglia output nuclei. (3) The Critic component emits the reward prediction error to all the Actor units and to its own prediction unit, similar to the divergent projection from midbrain dopamine neurons to a several hundredfold higher number of striatal neurons (Schultz, 1998). (4) Dopamine neuron activity seems to induce long-term changes in corticostriatal transmission (Reynolds, Hyland, & Wickens, 2001: Schultz, 1998). Dopamine neurotransmission would be in the anatomical position to decisively influence corticostriatal transmission, as the dendritic spines of striatal medium spiny neurons are commonly contacted by cortical and dopamine afferents (Smith & Bolam, 1990). Such dopamine-dependent plasticity could provide a

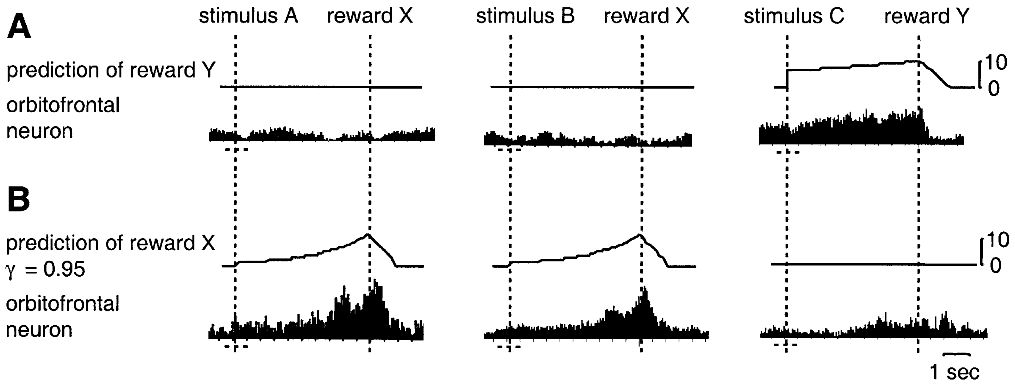

Fig. 6. Comparable time courses of TD prediction signals and orbitofrontal activities after learning pairings between different stimuli and rewards (figure adapted from Suri, 2001). Simulated stimuli are compared with the instruction stimuli in a delayed response task. (Histograms reconstructed from Schultz et al., 2000; Tremblay & Schultz, 1999, 2000) (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; discount factor $\gamma = 0.99$). Prediction of reward Y was comparable to the activity of a subset of orbitofrontal neurons that appear to anticipate reward Y but not reward X (bottom). (In the histogram at top, right, neural activity before the task was larger than after the task, because the previous task predicted already reward Y.) (B) Prediction of reward X was learned with a discount factor $\gamma = 0.95$ per 100 ms. 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 a subset of orbitofrontal neurons that appear to anticipate reward X but not reward Y (bottom). 9% of orbitofrontal neurons seem to reflect reward anticipation, as they are active during delay periods before specific rewards as shown in (A) and (B) (Schultz et al., 2000; Tremblay & Schultz, 1999, 2000).

biological basis for the postulated learning mechanisms in the Actor–Critic architecture.

Dopamine neurons not only project to the striatum but also to most cortical areas and may play in the cortex similar roles as in the striatum. According to this view, dopamine-dependent learning of sensorimotor associations as well as dopamine-dependent learning of prediction activities may also occur in the cortex.

7. Prediction activity in striatum and cortex

Anatomical considerations suggest that the reward prediction signal of the TD model may correspond to anticipatory firing rates of a subset of striatal and cortical neurons. How can we distinguish neural activity that serves as a reward prediction signal from other sustained activity? A crucial feature of the reward prediction signal in the TD model is that it is an anticipatory signal that may correspond to anticipatory neural activity. Anticipatory neural 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. Therefore, anticipatory activity precedes a future event irrespective of the physical features of the antedating events that make this future event predictable. Tonic delay period activity of several hundred milliseconds duration that anticipates stimuli, rewards or the animal's own actions was termed 'anticipatory', 'preparatory', or 'predictive' and has been reported in the striatum, supplementary motor area, prefrontal cortex, orbitofrontal cortex, premotor cortex, and primary motor cortex (Schultz, 2000; Suri & Schultz, 2001). The characteristics of reward-anticipatory neural activity in frontal cortices resemble those in the striatum (Hassani, Cromwell, & Schultz, 2001).

We compare reward prediction signals simulated with the TD model with reward-specific anticipatory activity recorded in orbitofrontal cortex (Fig. 6). Before recording started, monkeys had been trained in a delayed response task with instruction stimuli A and B followed by reward X and instruction stimulus C followed by reward Y (Schultz, Tremblay, & Hollerman, 2000; Tremblay & Schultz, 1999, 2000). The TD model is trained with the corresponding pairs of events. In trials without occurrence of reward Y, prediction of reward Y is not affected (Fig. 6(A), top, left and middle). In trials with occurrence of reward Y, this prediction signal is activated when stimulus C was presented and then progressively increased until reward Y (Fig. 6(A), top, right), because reward Y is completely predicted by stimulus C. Prediction of reward Y is comparable to reward-specific activity of a subset of orbitofrontal neurons that appears to anticipate reward Y but not reward X (Fig. 6(A), bottom).

The model is trained with the same pairs of events, but the value of 0.95 per 100 ms was used for the temporal discount factor γ. Therefore, after learning prediction signals increased more rapidly according to a rate of about 5% for each 100 ms (Fig. 6(B), top). Prediction of reward X was only slightly increased at the onset of stimuli A and B and then increased rapidly until reward X (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 (top, right side). Prediction of reward X was comparable to the activity of a subset of orbitofrontal neurons with activity that appears to anticipate reward X (Fig. 6(B), bottom).

Note that the temporal discount factor γ that correctly reproduces dopamine neuron activity is usually 0.98–0.99, corresponding to 1–2% increase in the prediction signal per 100 ms, and therefore 1–2% decrease of the dopamine response amplitude for each 100 ms between stimulus and reward (Suri & Schultz, 1999). For anticipatory neural activity, the correct value of the temporal discount factor γ was between 0.99 (Fig. 6(A)) and 0.95 (Fig. 6(B)), as the steepness of the progressive increase in the anticipatory neural activity varied. I suggest that the values of the temporal discount factors vary because predictions are computed over varying time scales, which can be crucial for some tasks (Precup & Sutton, 1998).

An alternative interpretation of neural activities as those proposed here is that these activities may code for sustained memory activity. However, this alternative explanation does not explain why these activities progressively increase before and decrease after the reward, as memory activity would be expected to progressively decrease after the stimulus presentation. Furthermore, this alternative explanation does not explain the occurrence of sustained activity following the stimuli A and B but not following C (Fig. 6(B)), although all three stimuli A, B, and C were physically different stimuli.

For these reasons, it was suggested that the neural activities shown in Fig. 6(A) and (B) anticipate specific rewards as do the simulated prediction signals (Schultz et al., 2000; Tremblay & Schultz, 1999, 2000). Although such reward-specific prediction signals can be used to compute a dopamine-like prediction error (Suri & Schultz, 2001), from a computational viewpoint it seems unnecessary that they are reward-specific. Why are anticipatory neural activities in cortex and striatum specific for rewards and do they influence dopamine neuron activity? In the remainder of this article I describe more advanced TD algorithms that indeed compute event-specific prediction signals and argue that these computationally more advanced TD algorithms may reproduce dopamine neuron activity more accurately than the standard TD model.

8. Internal model approaches

The standard TD algorithm (Fig. 1) can be used to learn to play board games by computing for each situation a prediction of the chance to win. This win prediction is the value of the board situation. The prediction error can be used to teach the optimal moves to an Actor network (Fig. 3). However, this approach is limited to well-trained board situations. To achieve world-class performance in board games like backgammon (Tesauro, 1994), it is necessary to calculate several half-moves ahead and to evaluate the values of hypothetical future board situations to select the best half-move. This can be achieved by extending the standard TD algorithm (Section 2) to an internal model approach. Such an internal model approach uses a computational unit that is able to simulate future board situations. This computational unit is also called internal model, world model (Sutton & Barto, 1981), forward model (Garcia, Prett, & Morari, 1989), or predictor and is defined as an algorithm that is able to learn to emulate the real experience. Pavlovian responses in animal experiments can be used to determine whether an animal has access to an internal model in a specific situation. An animal uses an internal model if it forms novel associative chains in the sensory preconditioning paradigm[1] (Sutton & Barto, 1981). Since the internal model can simulate experience, it can replace the Actor's real experience. Actor–Critic models that use internal model approaches for the Critic simulate future moves and use hypothetical future outcomes to select the best move. To make predictions, internal models typically simulate the evolution of the game within much shorter time periods than the real evolution of the game. Since they can simulate a sequence of future moves, internal model approaches form novel associative chains and are able to select the best move even in novel situations. This capability is usually called planning in machine learning studies (Sutton & Barto, 1998, see Dyna architecture) and goal-directed behavior or goal-directed instrumental action in animal learning studies (Balleine & Dickinson, 1998). Planning capabilities were demonstrated in many animal experiments (Balleine & Dickinson, 1998; Morris, Garrud, Rawlins, & O'Keefe, 1982; Thistlethwaite, 1951).[2]

For most motor actions of animals, it is not known whether they are achieved by planning or by sensorimotor learning, as the results of necessary control experiments are not known. When a monkey learns to press a lever it usually does not simply learn a pattern of muscle activation, since

[1] The sensory preconditioning paradigm is a simple experiment that demonstrates latent learning and formation of novel associative chains (Mackintosh 1974; Dickinson 1980). This paradigm is composed of three phases: In the first phase, a neutral stimulus A precedes a neutral stimulus B; in the second phase, stimulus B precedes a reinforcer; and in the third phase, stimulus A is presented alone. Animals show an acquired behavioral response to stimulus A in the third phase that resembles the response to the reinforcer. The similarity between this conditioned response to stimulus A and the unconditioned response to the reinforcer suggests that animals anticipate the occurrence of the reinforcer. This conclusion implies that animals internally form the novel associative chain "stimulus A is followed by stimulus B and stimulus B is followed by the reinforcer".

[2] Planning was demonstrated for rats in T-maze experiments (Thistlethwaite, 1951). This experiment consists of three phases: In the exploration phase, the rat is repeatedly placed in the start box where it can go left or right without seeing the two goal boxes at the end of the maze. When the rat turns to the left it reaches the red goal box, and if it turns to the right it reaches the green goal box. No rewards are presented in this exploration phase. In the rewarded phase, the rat is fed in the green goal box. In the test phase, the rat is returned to the start of the T-maze. In the first trial of to the test phase, the majority of the rats turn right. Note that neither the act of turning right nor the act of turning left is ever temporally associated with reward. It was concluded that the rat forms a novel associative chain between its own act, the color of the box, and the reward. Moreover, the rat selects its act in relation to the outcome predicted by this novel associative chain. Thus, the rat demonstrates in this first test phase trial its capability for planning.

even after changes in the monkey's position its hand does not miss the lever. This indicates that either (1) the monkey learns a sensorimotor association between a representation of the leaver and the press of its hand, or (2) the monkey presses the lever because it associates the pressed lever with reward delivery by using an internal model. Only sophisticated experiments that change the motivational value associated to the pressed lever without requiring the animal to press the lever can distinguish between both possibilities (Balleine & Dickinson, 1998; Suri, Bargas, & Arbib, 2001).

Does dopamine neuron activity reflect the processing of the standard TD model or rather that of a TD model extended to an internal model approach? To answer this question, dopamine neuron activity would have to be recorded in situations that test formation of novel associative chains, as does the sensory preconditioning paradigm, to investigate if a change in the motivational value of the outcome of a situation influences dopamine neuron activity. Since I am not aware of such a study, I rely on indirect evidence that supports that dopamine neuron activity may reflect the processing of an internal model. First, as striatal dopamine concentration is influenced by the formation of novel associative chains in the sensory preconditioning experiment (Young, Ahier, Upton, Joseph, & Gray, 1998), dopamine concentration reflects the use of an internal model, suggesting that dopamine neuron activity may be the output of an internal model. Second, since Parkinsonian patients seem to be impaired in planning tasks, dopamine may be involved in planning (Lange et al., 1992; Wallesch et al., 1990). Third, reward-specific and event-specific anticipatory neural activities in cortex and striatum represent the outcome of their actions already at the time of the behavior towards the outcome, which is typical for internal model approaches and not required for the standard TD model (Hassani et al., 2001; Schultz, 2000). For these reasons, I propose to model dopamine neuron activity and anticipatory neural activity in striatum and cortex with an internal model approach (Suri, 2001) and to use the dopamine-like signal of this internal model to select the correct actions in the Actor network (Suri et al., 2001). This approach implies that the rapid actions of dopamine on target neurons (Gonon, 1997), presumably in striatal matrisomes, select correct actions in situations that require planning (Suri et al., 2001). According to this model, preparatory activity for reward-promising actions is enhanced by increases in dopamine neuron activity. Activation of dopamine neurons occurring slightly before a saccadic eye movement to a visual stimulus, presumably due to neural activity anticipating the retinal consequences of the intended saccade (Duhamel, Colby, & Goldberg, 1992), may help to trigger intentional saccades. Using such planning processes, dopamine may attribute salience to reward-related stimuli and thereby trigger the animal's visual and internal attention to such targets (Redgrave, Prescott, & Gurney, 1999; Salamone, Cousins, & Snyder, 1997).

The physiological correlate of the internal model that seems to influence striatal dopamine concentration is not completely known. Recently, it has been speculated that certain cerebellar functions that can be mimicked with internal model approaches influence dopamine neuron activity (Doya, 1999). However, the term 'internal model' in the context of the cerebellum is defined differently than in the current paper (Kawato & Gomi, 1992). Whereas cerebellar pathways seem to compute an estimate of the current sensory experience, the internal model described here computes an estimate of future sensory experience by emulating the animal's environment. Since the internal model approach described here is an extension of the standard TD model, this internal model is likely to correspond to similar anatomical circuits as the Critic (Figs. 3 and 5). Cortical areas may be involved in learning associations between contingent events and in the formation of novel associative chains (Balleine & Dickinson, 1998). Event-specific anticipatory activities in cortex and striatum may correspond to prediction signals of the internal model. It is unclear how these structures may represent sensory events on a compressed time scale, which is a salient feature of internal models, but representations of such time compression occur in hippocampal place cells of mice running in known environments (Skaggs, McNaughton, Wilson, & Barnes, 1996). Within each theta cycle of about 100 ms duration, firing of place cell neurons reflects a tenfold temporal compression of the sensory experience. Therefore, if the spike timing is evaluated with respect to the local theta cycle, the reconstructed apparent path oscillates during each theta cycle with an amplitude of about 0.1 m around the physical path (Tsodyks, Skaggs, Sejnowski, & McNaughton, 1996). I speculate that in similar manner anticipatory neural activities in cortex and striatum may code for time compression mechanisms of internal models.

9. Conclusions

The finding that the TD model reproduces dopamine neuron activity in a variety of task situations is a great success for our understanding of brain functions in computational terms. Dopamine neuron activity appears to code a reward prediction error that is derived from reward prediction activities in the striatum and cortex. The comparison with Actor–Critic architectures suggest that dopamine neuron activity serves as an internal reward signal, or teaching signal, that helps to acquire motor habits in tasks with delayed reinforcement. Such a signal is crucial to learn movement sequences, since they are typically rewarded at the end of the sequence. Although Actor–Critc models that use the standard TD model as the Critic are successful for sensorimotor learning of habits, several lines of evidence suggest that the Critic should be extended to an internal model approach to reproduce dopamine neuron activity in tasks that require planning (Suri, 2001; Suri et al., 2001). Internal model approaches are computationally powerful (Garcia et al., 1989; Tesauro, 1994), can be effectively combined with TD algorithms (Sutton & Barto, 1998; see Dyna architecture), and their processing somewhat resembles aspects of ‘rehearsal’, ‘dreaming’, or ‘imagination’. The use of internal models by animals can be tested with the sensory preconditioning paradigm, whereas planning can be tested with the paradigm explained in the section above. The use of single cell recordings for both paradigms would reveal which neurons are involved in these cognitive processes. I suggest the novel hypothesis that the spike times of anticipatory neural activities in cortex and striatum relative to local rhythms may underlie the processing of internal models. This hypothesis can be tested using methods described by Skaggs et al. (1996) in the sensory preconditioning paradigm.

Since TD models only reproduce the phasic dopamine activities that are accessible in neuron recording experiments, it cannot be assumed that these models reproduce slow changes in the base line firing rates of dopamine neurons or in the dopamine concentrations in target areas. Furthermore, striatal dopamine concentration does not seem to be closely correlated with dopamine neuron activity as dopamine concentration is often enhanced in response to aversive stimuli (Horvitz, 2000; Young et al., 1998) whereas dopamine neuron activity is usually depressed (Mirenowicz & Schultz, 1994). Nevertheless, TD models may improve our understanding of addiction. According to the proposed Actor–Critic architecture, phasic increases in dopamine neuron activity reinforce previous behaviors and increase the probability that the reinforced behavior will be repeated in the same situation. Electrical self-stimulation and addictive drugs seem to elevate dopamine concentrations at forebrain dopamine terminals (Robinson & Berridge, 1993; White & Milner, 1992; Wise, 1996) and indeed lead to addictive behavior. In further agreement with the TD model, stimuli predicting the administration of heroin, cocaine (Kiyatkin, 1995), or food increase dopamine levels (Bassareo & Chiara, 1997). Note that TD Actor–Critic architectures do not imply that the subject experiences subjectively pleasurable feelings when dopamine neuron activity is increased but rather an urge to repeat previously reinforced habits. It has been hypothesized that separate systems are responsible for wanting (the urge to repeat habits) as compared to liking (pleasurable feelings) (Robinson & Berridge, 1993).

In addition to the responses to rewards and to reward prediction stimuli described earlier, dopamine neurons biphasically respond to physically salient stimuli that are not necessarily associated to reward. These responses are characterized as phasic increases of firing rates (about 100 ms duration) that are immediately followed by a depression in firing below baseline levels (100–300 ms duration) as if they coded for a brief reward expectation that is frustrated after 100 ms (Ljungberg, Apicella, & Schultz, 1992). These responses are consistent with the TD model because their occurrence and their habituation characteristics are consistent with those modeled by the standard TD model if certain adaptive weights are initialized with positive values (Kakade & Dayan, 2000; Suri & Schultz, 1999, 2002 (current issue of Neural Networks)). Since positive initial weights serve as a novelty bonus in TD algorithms and are used to stimulate exploration (Sutton & Barto, 1998), dopamine novelty responses may influence saccadic eye movements and other orienting responses to salient stimuli by rapid effects of dopamine neuron activity on target neurons (Gonon, 1997; Suri et al., 2001). There seems to be an interesting exception to the otherwise close correspondence between the reward prediction error signal of the standard TD model and the reported responses of midbrain dopamine neurons. It was reported for one dopamine neuron that its activity was not consistent with that predicted by the TD model if the reward was delivered earlier than usual (Hollerman & Schultz, 1998). The early reward delivery may reset internal states, similar to attention shifts that happen to us when a salient and surprising event interrupts our concentration. Although the TD model was adapted to correctly model this situation (Suri & Schultz, 1999), this extension requires some ad hoc assumptions that are hard to justify from a mathematical viewpoint. A mathematically convincing approach would probably require computational methods that resemble the updating of the states of the internal model by a Kalman filter approach (Dayan & Kakade, 2000).

10. Uncited reference

Kakade and Dayan, 2001.

Acknowledgments

I thank Srinivas Turaga and Martina Wicklein for comments on earlier versions of the manuscript. This study was supported by fellowships of the EMDO Stiftung Zurich and of the Swiss National Science Foundation (823A-056731).

References

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4–5), 407–419.

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

Bassareo, V., & Chiara, G. D. (1997). Differential influence of associative and nonassociative learning mechanisms on the responsiveness of prefrontal and accumbal dopamine transmission to food stimuli in rats fed ad libitum. *Journal of Neuroscience*, *17*, 851–861.

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.

Dayan, P., & Kakade, S. (2000). *Explaining away in weight space*. *NIPS*, pp. 451–457.

Dayan, P., & Sejnowski, T. J. (1994). TD(λ) converges with probability 1. *Machine Learning*, *14*, 295–301.

Dickinson, A. (1994). *Instrumental conditioning*. *Animal learning and cognition*, San Diego: Academic Press, pp. 45–78.

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

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.

Friston, K. J., Tononi, G., Reeke, G. N., Jr., Sporns, O., & Edelman, G. M. (1994). Value-dependant selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, *59*(2), 229–243.

Garcia, C. E., Prett, D. M., & Morari, M. (1989). Model predictive control: Theory and practice—A survey. *Automatica*, *25*, 335–348.

Gonon, F. (1997). Prolonged and extrasynaptic excitatory action of dopamine mediated by D1 receptors in the rat striatum in vivo. *Journal of Neuroscience*, *17*(15), 5972–5978.

Hassani, O. K., Cromwell, H. C., & Schultz, W. (2001). Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. *Journal of Neurophysiology*, *85*(6), 2477–2489.

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.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656.

Kakade, S., & Dayan, P. (2000). Dopamine bonuses. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), (pp. 131–137). *NIPS 2000*.

Kakade, S., & Dayan, P (2001). *Dopamine: Generalization and bonuses*. Current Issue of Neural Networks.

Kawato, M., & Gomi, H. (1992). The cerebellum and VOR/OKR learning models. *Trends in Neurosciences*, *15*(11), 445–453.

Kiyatkin, E. A. (1995). Functional significance of mesolimbic dopamine. *Neuroscience and Biobehavioral Reviews*, *19*, 573–598.

Lange, K. W., Robbins, T. W., Marsden, C. D., James, M., Owen, A. M., & Paul, G. M. (1992). L-dopa withdrawal in Parkinson's disease selectively impairs cognitive performance in tests sensitive to frontal lobe dysfunction. *Psychopharmacology (Berlin)*, *107*, 394–404.

Lees, A. J., & Smith, E. (1983). Cognitive deficits in the early stages of Parkinson's disease. *Brain*, *106*, 257–270.

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

Mason, S. T., Beninger, R. J., Fibiger, H. C., & Phillips, A. G. (1980). Pimozide-induced suppression of responding: Evidence against a block of food reward. *Pharmacology, Biochemistry, and Behavior*, *12*, 917–923.

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

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*(6551), 725–728.

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

Morris, R. G., Garrud, P., Rawlins, J. N., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*(5868), 681–683.

Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences—A computational approach. *Journal of Cognitive Neurosciences*, *13*(5), 626–647.

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

Phillips, J. G., Bradshaw, J. L., Iansek, R., & Chiu, E. (1993). Motor functions of the basal ganglia. *Psychological Research*, *55*, 175–181.

Precup, D., & Sutton, R. S. (1998). *Multi-time models for temporally abstract planning* (*11*). *Advances in neural information processing systems 11*, Cambridge, MA: MIT Press.

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

Reynolds, J. N. J., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67–70.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug creaving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*, 247–291.

Salamone, J. D., Cousins, M. S., & Snyder, B. J. (1997). Behavioral functions of nucleus accumbens dopamine: Empirical and conceptual problems with the anhedonia hypothesis. *Neuroscience and Biobehavioral Reviews*, *21*, 341–359.

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

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, *1*(3), 199–207.

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

Skaggs, W. E., McNaughton, B. L., Wilson, M. A., & Barnes, C. A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, *6*(2), 149–172.

Smith, A. D., & Bolam, J. P. (1990). The neural network of the basal ganglia as revealed by the study of synaptic connections of identified neurons. *Trends in Neurosciences*, *13*(7), 259–265.

Suri, R. E. (2001). Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Experimental Brain Research*, *140*(2), 234–240.

Suri, R. E., & Schultz, W. (1998). Dopamine-like reinforcement signal improves learning of sequential movements by neural network. *Experimental Brain Research*, *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.

Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Computation*, *13*(4), 841–862.

Suri, R. E., Bargas, J., & Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, *103*, 65–85.

Sutton, R. S., & Barto, A. G. (1981). An adaptive network that constructs and uses an internal model of its world. *Cognition and Brain Theory*, *4*(3), 217–246.

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, Bradford Books, Available at www-anw.cs.umass.edu/ ~ rich/book/the-book.html.

Tesauro, G. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, *6*, 215–219.

Thistlethwaite, D. (1951). A critical review of latent learning and related experiments. *Psychological Bulletin*, *48*, 97–129.

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. *Journal of Neurophysiology*, *83*(4), 1864–1876.

Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., & McNaughton, B. L. (1996). Population dynamics and theta rhythm phase precession of hippocampal place cell firing: A spiking neuron model. *Hippocampus*, *6*(3), 271–280.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature 5*, *412*(6842), 43–48.

Wallesch, C. W., Karnath, H. O., Papagno, C., Zimmermann, P., Deuschl, G., & Lucking, C. H. (1990). Parkinson's disease patient's behaviour in a covered maze learning task. *Neuropsychologia*, *28*, 839–849.

White, N. M., & Milner, P. M. (1992). The psychobiology of reinforcers. *Annual Reviews in Psychology*, *43*, 443–471.

Wise, R. A. (1996). Addictive drugs and brain stimulation reward. *Annual Reviews in Neuroscience*, *19*, 319–340.

Young, A. M., Ahier, R. G., Upton, R. L., Joseph, M. H., & Gray, J. A. (1998). Increased extracellular dopamine in the nucleus accumbens of the rat during associative learning of neutral stimuli. *Neuroscience*, *83*(4), 1175–1183.