MODELING FUNCTIONS OF STRIATAL DOPAMINE MODULATION IN LEARNING AND PLANNING

Suri R. E.[1,*], Bargas J.[2], and Arbib M.A. [1]
[1]USC Brain Project, Los Angeles CA 90089-2520
[2]Departamento de Biofisica, Instituto de Fisiologia Cellular, Universidad Nacional Autonoma de Mexico, Mexico City DF
Version from 24 February, 2000

The activity of midbrain dopamine neurons is strikingly similar to the reward prediction error of TD reinforcement learning models. Experimental evidence and simulation studies suggest that dopamine neuron activity serves as an effective reinforcement signal for learning of sensorimotor associations in striatal matrisomes.

In the current study, we simulate dopamine neuron activity with the Extended TD model (Suri and Schultz, submitted) and examine the influence of this signal on medium spiny neurons in striatal matrisomes. This model includes transient membrane effects of dopamine, dopamine-dependent long-term adaptations of corticostriatal transmission, and rhythmic fluctuations of the membrane potential between an elevated "up-state" and a hyperpolarized "down-state." The most dominant activity in the striatal matrisomes elicits behaviors via projections from the basal ganglia to the thalamus and the cortex.

This "standard model" performs successfully when tested for sensorimotor learning and goal-directed behavior (planning). To investigate the contributions of these model assumptions to learning and planning, we test the performance of several model variants that lack one of these mechanisms. These simulations show that the adaptation of the dopamine-like signal is necessary for planning and for sensorimotor learning. Lack of dopamine-like novelty responses decreases the number of exploratory acts, which deteriorates planning capabilities. Sensorimotor learning requires dopamine-dependent long-term adaptation of corticostriatal transmission. The model loses its planning capabilities if the dopamine-like signal is simulated with the original TD model. The capability for planning is improved by transient dopamine membrane effects, dopamine-dependent long-term effects on corticostriatal transmission, dopamine- and input-dependent influences on the durations of membrane potential fluctuations, and manipulations that prolong the reaction time of the model. These simulation results suggest that striatal dopamine is important for sensorimotor learning, exploration, and planning.

INTRODUCTION

Midbrain dopamine neurons are phasically activated by unpredicted rewards or by the first sensory event that allows the animal to predict the reward but do not respond to predicted rewards. When a predicted reward is omitted, their activity is depressed at the time when the reward fails to occur (Schultz, 1998). The reward prediction error of temporal difference models (TD models) reproduces these features of dopamine neuron activity (Sutton and Barto, 1990; Montague *et al.*, 1996; Schultz *et al.*, 1997; Suri and Schultz, 1999, submitted). In addition, dopamine neurons respond to novel, physically salient stimuli (Schultz, 1998). Such stimuli elicit action potential bursts followed by activity decreases below baseline

[*] Present Address: Roland Suri, Computational Neurobiology Laboratory,
Salk Institute, Post Office Box 85800, San Diego CA 92186-5800; suri@salk.edu,
http://www.cnl.salk.edu/~suri

levels. These biphasic novelty responses diminish with repeated stimulus presentation. TD models reproduce these characteristics of dopamine novelty responses if the associative weights of stimulus onsets are initialized with positive values (Suri and Schultz, 1999).

Simulation studies with TD models demonstrate that a dopamine-like reward prediction error can serve as a powerful effective reinforcement signal for sensorimotor learning (Houk *et al*. 1995; Suri and Schultz, 1998, 1999). In such models, the TD model is termed "Critic" and the model component that learns sensorimotor associations is termed "Actor." The Critic was related to pathways from cortex via striatal striosomes to midbrain dopamine neurons and the Actor to pathways from cortex via striatal matrisomes, basal ganglia output nuclei, and thalamus to motor cortices (Fig. 1A) (Houk *et al*. 1995; Montague *et al*., 1996; Schultz *et al*., 1997; Suri and Schultz, 1998, 1999).

Behaviors of humans and animals are often influenced by expectations about task outcomes. Such goal-directed behavior requires planning (see section "Planning and Sensorimotor Learning"). Planning was simulated with Actor-Critic models for which the action selection of the Actor is guided by transient influences of the Critic (Sutton and Barto, 1981). Since dopamine bursts transiently influence striatal activity (Gonon, 1997) and dopamine is involved in planning tasks (Wallesch *et al.*, 1990; Salamone, 1992; Lange *et al.*, 1992; Talor and Saint-Cyr, 1995), dopamine may guide action selection in such planning tasks. To test this hypothesis, we simulate interactions between the anatomical structures shown in Fig. 1A in a simulation experiment that assesses sensorimotor learning and planning.

To model the transient influences of dopamine on membrane properties of medium spiny neurons in striatal matrisomes, we simulate the *in vitro* finding that activation of D1 class dopamine receptors decreases firing evoked from resting potentials but increases firing evoked from elevated holding potentials (Herndandez-Lopez *et al*., 1997; see section "Striatal Membrane Effects of Dopamine D1 Agonists *In Vitro*"). Likewise, dopamine long-term effects on corticostriatal transmission depend on the postsynaptic membrane potential (Cepeda and Levine, 1998). The membrane potential is influenced by synaptic inputs, by dopamine levels, and by rhythmic fluctuations of about 1 Hz between a depolarized up-state and a hyperpolarized down-state (Stern *et al.*, 1997). We propose a model for striatal medium spiny neurons that mimics membrane potential fluctuations as well as dopamine effects on membrane properties and on corticostriatal transmission (section "Striatal Dopamine Modulation *In Vivo*").

Then we model the influence of dopamine neuron activity on medium spiny neurons in striatal matrisomes in concert with the sensorimotor components of the basal ganglia-thalamocortical system (Fig. 1B; section "Basal Ganglia-Thalamus-Cortex"). The dopamine-like signal is computed with the Extended TD model (Suri and Schultz, submitted; section "Extended TD Model"). Since one out of two acts can be selected in the simulated experiment, we model dopamine-dependent influences on two neuron populations in striatal matrisomes. Each neuron population is thought to correspond to a small population of medium spiny neurons with highly correlated activations that is able to elicit one of both acts. We assume that the simulated signals in the basal ganglia-thalamocortical pathway are carried by similar populations of neurons. For simplicity, we call these neuron populations "(simulated) neurons". Following the proposal of Berns and Sejnowski (1996), we assume that only the predominant striatal firing rate is represented in the basal ganglia output nuclei globus pallidus interior (GPi) and substantia nigra pars reticulata (SNr) and projected via thalamus to cortical areas. Strong and persistent depressions of firing rates in these basal ganglia output nuclei elicit acts via thalamocortical projections.

In addition to testing the performance of this "standard model" (Fig. 1B) in sensorimotor learning and planning, we are interested in the performance of simpler model variants. Therefore, we test model variants without dopamine-like novelty responses, without projections from striatal matrisomes to the Extended TD model (salience $a = 0$, see Fig. 1B), without synaptic long-term effects, without transient dopamine membrane effects, with constant up- and down-state durations, and with the original TD model instead of the Extended TD model.

The model is implemented in time steps of 100 *m*sec to reduce computation time (about 200 hours for the shown results on a Sun Ultra 1). The documented NSLJ code can be assessed and executed with

standard web browsers (Suri, Marmol, and Arbib, in preparation) and a Matlab code at http://www.cnl.salk.edu/~suri. A study proposal was presented in abstract form (Suri and Arbib, 1998).

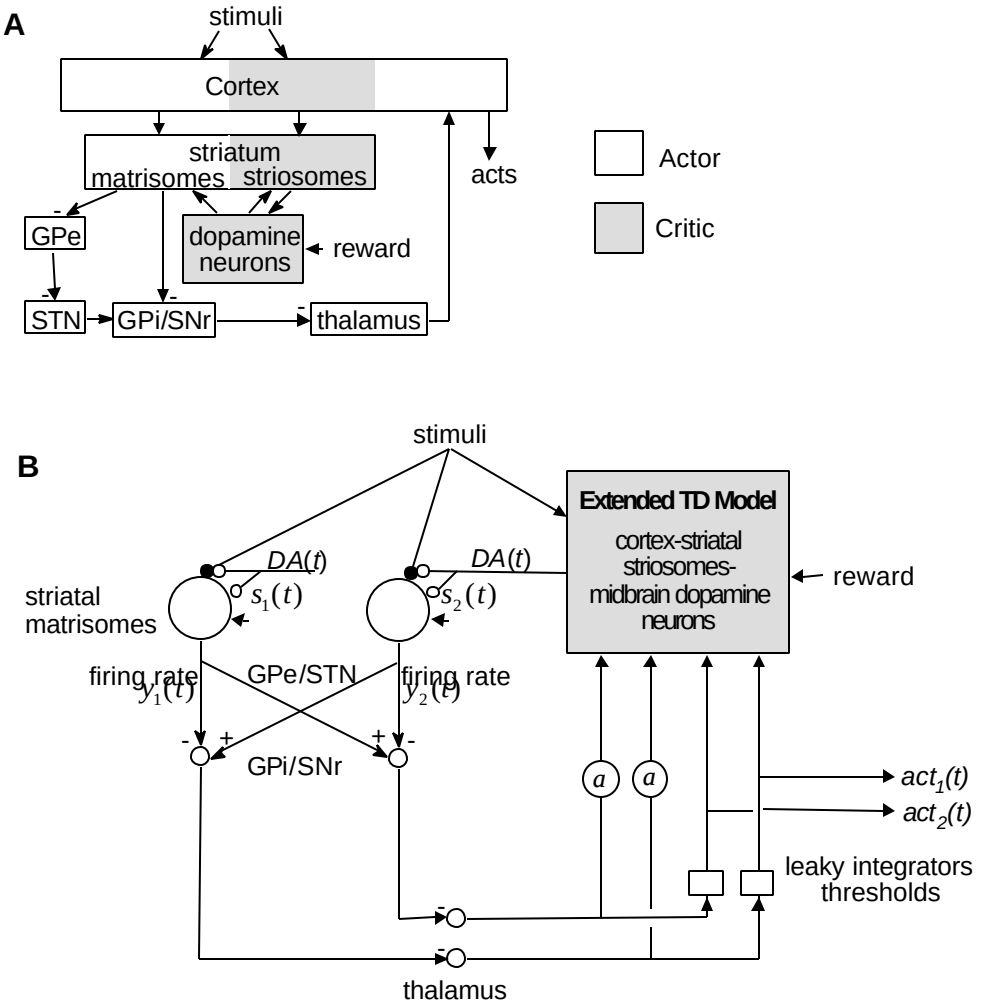

Fig. 1 (A) Interactions between cortex, basal ganglia, and midbrain dopamine neurons mimicked by the model. Cortical pyramidal neurons project to the striatum, which can be divided in striosomes (patches) and matrisomes (matrix) (Graybiel, 1990). Prefrontal and insular cortices project chiefly to striosomes, whereas sensory and motor cortices project chiefly to matrisomes (Graybiel, 1990). Midbrain dopamine neurons are contacted by medium spiny neurons in striosomes and project to both striatal compartments (Graybiel, 1990; Smith and Bolam, 1990). Striatal matrisomes directly inhibit the basal ganglia output nuclei globus pallidus interior (GPi) and substantia nigra pars reticulata (SNr), whereas they indirectly disinhibit these output nuclei via globus pallidus exterior (GPe) and subthalamic nucleus (STN) (Albin *et al*., 1989; Alexander and Crutcher, 1990). The basal ganglia output nuclei project via thalamic nuclei to motor, occulomotor, prefrontal, and limbic cortical areas (Alexander and Crutcher, 1990). The structures shown as grey boxes correspond to the Critic and those shown as white boxes to the Actor. **(B)** Model architecture. The Extended TD model serves as the Critic component (grey box), and the Actor component (remaining architecture) elicits acts. *Actor:* Sensory stimuli influence the membrane potentials of two medium spiny projection neurons in striatal matrisomes (large circles). The membrane potentials of these neurons are also influenced by fluctuations between an elevated up-state and a hyperpolarised down-state simulated with the functions $s_1(t)$ and $s_2(t)$. Adaptations in corticostriatal weights (filled dots) and dopamine membrane effects are influenced by the membrane potential and the dopamine-like signal $DA(t)$ (open dots). The firing rates $y_1(t)$ and $y_2(t)$ of both striatal neurons inhibit the basal ganglia output nuclei substantia nigra pars reticulata (SNr) and globus pallidus interior (GPi). An indirect disinhibitory pathway from striatum to GPi/SNr suppresses insignificant inhibitions in the basal ganglia output nuclei (Berns and Sejnowski, 1996). The winning inhibition disinhibits the thalamus. These signals in the thalamus lead only to acts, coded by the signals $act_1(t)$ and $act_2(t)$, if they are sufficiently strong and persistent. This is accomplished by integrating the cortical signal and eliciting acts when it reaches a threshold. *Critic:* The Critic and computes the dopamine-like reward prediction error $DA(t)$ from the sensory stimuli, the reward signal, the thalamic signals (multiplied with the salience a), and the act signals $act_1(t)$ and $act_2(t)$.

PLANNING AND SENSORIMOTOR LEARNING

Animal Learning

In a broad spectrum of situations, animals select acts based on formation of novel associative chains. Animals learn act outcomes and select their acts based on the motivational value of these outcomes (reviews in Thistlethwaite, 1951; MacCorquodale and Meehl, 1954; Mackintosh 1974; Dennett 1978; Dickinson 1980; Dickinson 1994; Dickinson and Balleine 1994; Balleine and Dickinson 1998). Such goal-directed behavior is termed “planning” in reinforcement learning studies, or “cognition” in animal learning studies (Craik 1943; Sutton and Barto, 1981, 1998; Sutton and Pinette 1985; Dickinson 1994). Planning and sensorimotor learning were demonstrated for rats in *T*-maze experiments (Fig. 2A) (reviews in Thistlethwaite, 1951; MacCorquodale and Meehl, 1954). The 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. 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 turns 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 dependent on the outcome predicted by this novel associative chain. Thus, the rat demonstrates its capability to plan in this first test phase trial.

In test phase trials, the rat is fed in the green goal box but not in the red goal box. The more test phase trials are presented, the higher is the probability that the rat turns right. Since the rat is fed after right turns, this progressive performance improvement is interpreted as a result of sensorimotor learning.

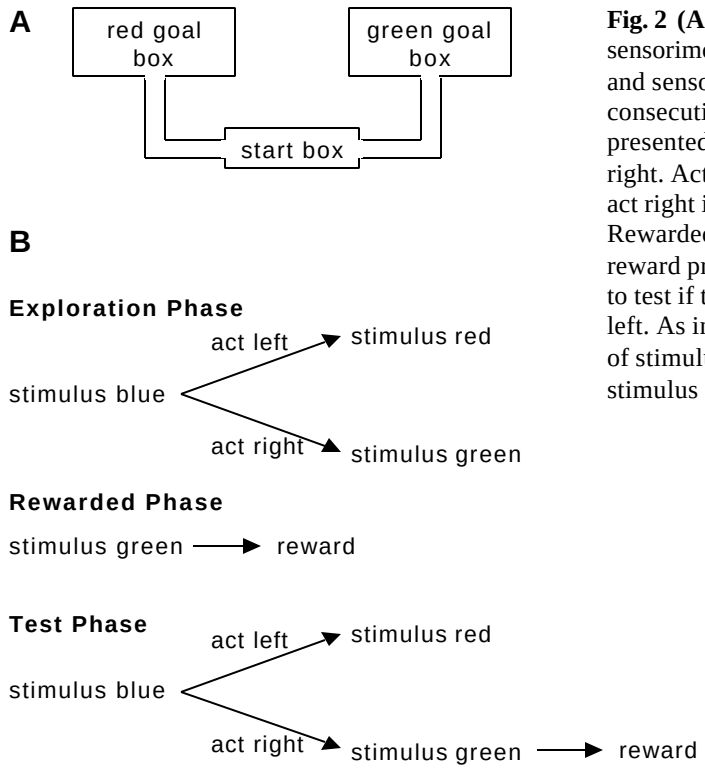

Fig. 2 (A) Configuration of *T*-maze to test planning and sensorimotor learning in rats. **(B)** Simulated task to test planning and sensorimotor learning. The task is composed of three consecutive phases. *Top*: Exploration phase. When stimulus blue is presented, the model selects with equal chance the act left or the act right. Act left is followed by presentation of stimulus red, whereas act right is followed by presentation of stimulus green. *Middle*: Rewarded phase. Presentation of stimulus green is followed by reward presentation. *Bottom*: Test phase. Stimulus blue is presented to test if the model elicits the correct act right or the incorrect act left. As in the exploration phase, act left is followed by presentation of stimulus red, whereas act right is followed by presentation of stimulus green and by that of the reward.

Animal Learning Models

Based on such experimental findings, animal learning theorists suggest that animals form an internal model of their environment that allows them to predict the sensory consequences of their acts and to form novel associative chains. Furthermore, animals seem to use these predictions to select their acts (Sutton and Pinette, 1985; Sutton and Barto 1981; Dickinson and Balleine 1994; Balleine and Dickinson 1998). This insight led to the implementation of neural network architectures in which an internal model, serving as the Critic, transiently influences the Actor to elicit acts (Sutton and Barto, 1981). In such Actor-Critic architectures, the Actor computes small random variations in act preparation signals and executes acts when these preparation signals reach a threshold. The Critic component learns associations between sensory stimuli, rewards, and act preparation signals and uses these associations to form novel associative chains. The output of the Critic is a signal that reflects the value of the predicted outcome and reinforces or attenuates the act preparation signals. In this manner, the effective reinforcement signal of the Critic selects the act that predicts the optimal outcome. This animal learning model resembles the model that will be proposed in the current study.

Task Simulation

In the current study, we test model performance in a task analogous to the *T*-maze task. In the exploration phase (Fig. 2B, top), each trial starts with presentation of a stimulus called "blue" (stimulus blue) that represents sensory features of the start box. When either the act right or the act left is executed during presentation of stimulus blue, the stimulus is extinguished and either stimulus green or stimulus red is presented, respectively. The act right and the act left represent the rat's right and left turn, respectively, whereas the stimuli green and red correspond to the colors of the goal boxes. If no act is selected, stimulus blue is extinguished after 600 *m*sec. This exploration phase is simulated for a time span corresponding to 80 sec (exclusive of intertrial intervals), during which stimulus blue is presented about 100 times. The subsequent rewarded phase consists of only one trial (Fig. 2B, middle), in which presentation of stimulus green is followed by reward presentation. The beginning of the test phase (Fig. 2B, bottom) is equal to the exploration phase. Stimulus green but not stimulus red is followed by reward presentation. In all three phases, the stimuli green and red are presented for 300 *m*sec and the reward for 100 *m*sec. Planning is assessed in the first trial of the test phase and the progress of sensorimotor learning in subsequent trials.

STRIATAL MEMBRANE EFFECTS OF DOPAMINE D1 AGONISTS *IN VITRO*

The resting membrane potential of medium spiny neurons *in vitro* is about -80 *m*V. If firing is evoked from such polarized potentials, dopamine D1 class receptor activation attenuates the firing rate. This reduction in firing rate has been attributed to subthreshold K^+ channels (Pacheco-Cano *et al.* 1996), to the modulation of Na^+ channels (Calabresi *et al.* 1987; Surmeier *et al.* 1992; Cepeda *et al.* 1995), and to channels participating in the afterhyperpolarization (Rutherford et al. 1988; Hernandez-Lopez *et al.* 1996). In contrast, if firing is evoked from elevated holding potentials, dopamine D1 class receptor activation enhances firing via a G-protein dependent potentiation of an L-type calcium current (Surmeier *et al*., 1995). Hernandez-Lopez and collaborators (1997) demonstrated that both effects occur in the same medium spiny neuron. In this study, the effects of three dopamine D1 class agonists were investigated for firing evoked by 200-300 *m*sec current steps at a frequency of 0.1-0.2 Hz. Bath application of dopamine D1 class agonists attenuated firing when evoked from the resting membrane potential. To test the effect of D1 class agonists for elevated membrane potentials, a sustained subthreshold current was injected to hold the membrane potential on an elevated value just below firing threshold. Suprathreshold current steps were superimposed on this sustained holding current. Dopamine D1 class agonists enhanced firing rate when evoked from this elevated holding potential (Fig. 3).

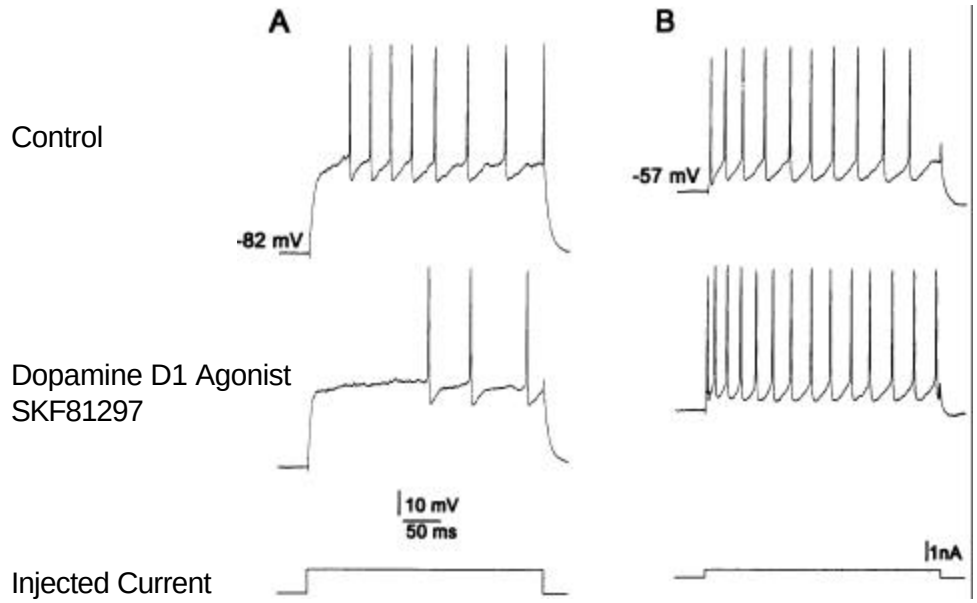

Fig. 3 Dopamine D1 class receptor agonist SKF 81297 enhances or attenuates evoked firing depending on the holding potential (Figure adapted with permission from Hernandez-Lopez *et al*., 1997). **(A)** Firing was evoked with a current step from the resting potential of -82 *m*V (top, eight action potentials). 1 *m*M of D1 receptor agonist SKF 81297 attenuated evoked firing (middle, three action potentials). Injected current was maintained for both conditions (bottom). **(B)** For the same neuron, firing was evoked from a holding potential of -57 *m*V (top, 10 action potentials). 1 *m*M of D1 receptor agonist SKF81297 increased evoked firing (middle, 14 action potentials). Injected current was again maintained for both conditions (bottom).

Model

Effects of neuromodulators on neuronal membrane properties have been simulated with various modeling techniques (reviewed in Fellous and Linster, 1998). To simulate the findings of Hernandez-Lopez *et al.* (1997), we propose a phenomenological model for membrane effects mediated by dopamine D1 class receptors (Fig. 4A). Since dopamine D1 class receptor activation enhances or attenuates evoked firing depending on the holding potential, we introduce a reverse potential (Table 1). This reverse potential corresponds to the hypothetical holding potential between resting potential and firing threshold for which the effect of D1 activation on evoked firing vanishes as it reverses its sign. The term reverse potential should not be confused with the biophysically defined term reversal potential that is defined by a current or voltage reversal. The influence of the holding potential on the firing rate during the current step is modeled using a slowly varying parameter $W_{mem}(t)$ that is initialized with a value of zero and then adapted with

$$W_{mem}(t) = d\ W_{mem}(t\text{-}100) + hDA(t\text{-}100)[E(t\text{-}100)\text{-}reverse\_potential]. \qquad \text{(eq. 1)}$$

The time t is given in units of *m*sec throughout this paper. A constant d denotes the decay rate of the D1 effects. Since the D1 agonist effects decay to a value of 40% over 10 minutes (see figure 4B in Hernandez-Lopez et al., 1997), we estimate the value of the decay rate d of the dopamine membrane effects to be 0.99985, which corresponds to 0.015 % decrease for each 100 *m*sec. Note that this value depends on the mode of the D1 agonist application, since the D1 agonist effects decay faster if the agonists are applied directly on isolated cells (Surmeier *et al*., 1995) and much faster if dopamine neurons are activated *in vivo* (Gonon, 1997). Therefore, we use the value d =0.99985 to reproduce the experiment of Hernandez-Lopez et al. (1997) but will adjust this value for the *in vivo* model (see next section). The signal $DA(t)$ corresponds to the dopamine D1 agonist concentration and the parameter h is a scaling factor. A signal $E(t)$ denotes the membrane potential in *m*V and is defined below. The absolute value of

the dopamine membrane effects $W_{mem}(t)$ is limited to $W_{mem,max} = 9$, as this value is appropriate to reproduce the maximal amplitude of the D1 agonist effects.

The subthreshold membrane potential $E_{sub}(t)$ is computed from the resting membrane potential E_{rest}, from the resistance R, and from the injected current $I(t)$ according to Ohm's law with

$$E_{sub}(t) = E_{rest} + R\, I(t). \qquad \text{(eq. 2)}$$

For the neuron shown in Fig. 3, the value of the resting potential is estimated to E_{rest} = -82 *m*V and the value of the resistance R to 27 MOhm. The latter value does not correspond to a biophysical property of the neuron but depends on the used electrode. Since the seal between the used intracellular sharp electrode and the neural membrane is far from being complete, a major part of the injected current $I(t)$ does not reach the inside of the neuron. For membrane potentials below firing threshold, $E_{sub}(t)$ approximates the membrane potential. For values above firing threshold, the membrane potential $E_{sub}(t)$ does not have a direct biological correspondence, as it is only defined for time steps of 100 *m*sec and therefore cannot reflect the quickly varying time course of the membrane potential for the spiking neuron. Since we assume that the firing rate increases with increasing values of the membrane potential $E_{sub}(t)$ and of the adaptive parameter $W_{mem}(t)$, we compute the firing rate $y(t)$ of the striatal neuron with

$$y(t) = y_{max} \times tanh\{a \times [E_{sub}(t) + W_{mem}(t) - firing\_threshold]/\, y_{max}\}. \qquad \text{(eq. 3)}$$

From Fig. 3, the firing threshold is estimated to be −56 *m*V. This value does not correspond to the average firing threshold *in vivo* and will be adjusted for the *in vivo* model. The hyperbolic tangent *tanh* is a sigmoid function, and the function $y_{max} \times tanh\{\,./\,y_{max}\,\}$ smoothly limits the firing rate $y(t)$ to values below the maximal firing rate y_{max} of medium spiny neurons. Medium spiny neurons fire with a maximal firing rate y_{max} of about 6 Spikes per 100 *m*sec (Apicella *et al.*, 1992; Pineda *et al.*, 1992; Nisenbaum *et al.* 1994). A factor a = 0.3 Spikes/(100 *m*sec * *m*V) is used to scale the firing rate to experimental data and is estimated from firing rates for constant current injections in the absence of dopamine agonists (estimated from figure 1 in Nisenbaum *et al.* 1994; similar in figure 3 in Pineda *et al.*, 1992). Eq. 3 does not take into account that firing rate adaptations occur during a few hundreds of milliseconds after current step injections (Fig. 3; Pineda *et al.*, 1992). The dopamine-dependent signal $W_{mem}(t)$ influences the firing rate (eq. 3) rather than the subthreshold membrane potential $E_{sub}(t)$ (eq. 2), as the membrane potential before and after current step injections is not substantially influenced by D1 agonist application (Hernandez-Lopez *et al.*, 1997).

Eq. 2 computes the subthreshold membrane potential $E_{sub}(t)$ but not the membrane potential for the firing neuron. Since the 100 *m*sec step size of our implementation is too long to simulate single spikes, we approximate the membrane potential $E(t)$ for a certain firing rate with the average membrane potential of real neurons with this firing rate. This is achieved by computing the contribution of measured action potentials to the average membrane potential. From intracellular voltage recordings of spontaneously active striatal medium spiny neurons (figure 3F in Wickens and Wilson, 1998), we estimate for a single spike an area of 6 *m*V × 100 *m*sec between the firing threshold and the membrane potential. Using this value, the membrane potential is computed with

$$E(t) = \begin{cases} firing\_threshold + y(t) \times 6\ mV, \text{ if } E_{sub}(t) > firing\_threshold \\ E_{sub}(t), \text{ else.} \end{cases} \qquad \text{(eq. 4)}$$

For subthreshold membrane potentials, this equation sets the membrane potential $E(t)$ to $E_{sub}(t)$, whereas for the firing neuron, $E(t)$ is set to the average membrane potential of real neurons with this firing rate.

Simulation

To test the proposed model, we simulate the experimental conditions of Hernandez-Lopez *et al.* (1997) (Fig. 3). To mimic the conditions for the result shown in Fig. 3A, we simulate current step injections of 1.3 *n*A amplitude and 300 *m*sec duration with a frequency of 0.1 Hz. To reproduce Fig. 3B, the holding potential of -57 *m*V is induced with a sustained current $I(t)$ of 0.9 *n*A. Superimposed on this holding

current, current steps of 0.4 *n*A amplitude and 300 *m*sec duration are injected with a frequency of 0.1 Hz. Therefore, these two simulated current injection protocols differ only in the holding current and not in the total amount of current injection during current steps. For the slice experiment, this was at least approximately the case (Fig. 3). The effective D1 agonist concentration $h \times DA(t)$ is set to zero for the condition without D1 agonist application and to the value of 0.1 during D1 agonist application (the symbol h is here only introduced to make the notation consistent throughout this paper). This value is chosen to reproduce the extent of the effect measured by Hernandez-Lopez *et al.* (1997). To simulate the condition with D1 agonist application, current steps are repeated until the firing response is stable, which is the case after 30 sec of simulated physical time.

To reproduce dopamine modulation for the recorded neuron (Fig. 3), the value of the reverse potential is set to the value of –58 *m*V. Values for the reverse potential between about –70 *m*V and about –57 *m*V lead to results that are consistent with the experimental findings (results not shown).

Results

The finding that D1 agonists can be inhibitory or excitatory for medium spiny neurons depending on the holding potential (see Fig. 3) is reproduced with the model (Fig. 4). Since the term $h \cdot DA(t)$ is set to zero to simulate the control condition without D1 agonist application, the membrane effect signal $W_{mem}(t)$ remains on the initial value of zero (not shown, follows from eq.1). Therefore, the evoked firing rate (Fig. 4B and 4C, top) does not depend on the holding potential but only on the injected current (Fig. 4B and 4C, bottom line; eqs. 2 and 3). Although this model feature does not hold exactly for medium spiny neurons (Pineda *et al*., 1992), the simulated subthreshold membrane potentials and firing rates reproduce approximately those for the medium spiny neuron shown on top of Fig. 3.

Since the term $h \cdot DA(t)$ is positive to simulate D1 agonist application and the resting membrane potential of -82 *m*V is below the reverse potential, the dopamine membrane effect signal $W_{mem}(t)$ is negative (Fig. 4B, line 3; eq. 1). In contrast, $W_{mem}(t)$ is positive if the holding membrane potential of -57 *m*V is above the reverse potential (Fig. 4C, line 3). Since the firing rate depends on the dopamine membrane effect signal $W_{mem}(t)$ (eq. 3), D1 agonist application attenuates firing evoked from the resting potential of –82 *m*V but enhances firing evoked from the holding potential of -57 *m*V (Fig. 4B and 4C, line 2). These simulated D1 effects reproduce approximately those for the medium spiny neuron shown in Fig. 3.

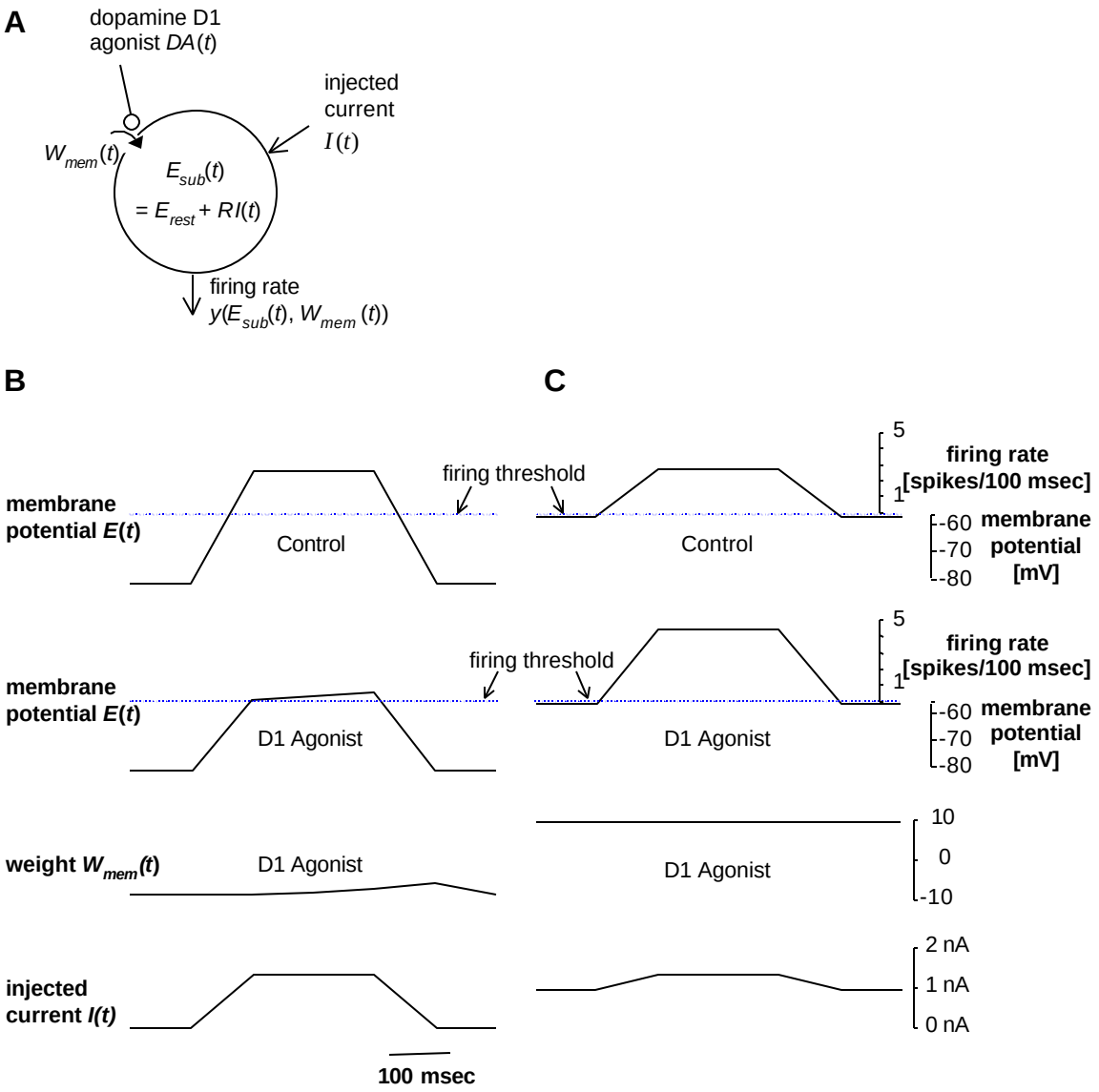

Fig. 4 (A) Model for effects of dopamine D1 class receptor activation on the firing rate of a medium spiny neuron *in vitro*. The subthreshold membrane potential $E_{sub}(t)$ depends on the constant resting membrane potential E_{rest} and on the product of the injected current $I(t)$ with a resistance R. The subthreshold membrane potential $E_{sub}(t)$ and dopamine D1 agonist concentration $DA(t)$ influence the value of the signal $W_{mem}(t)$. The firing rate $y(t)$ is a monotonically increasing function of the subthreshold membrane potential $E_{sub}(t)$ and the signal $W_{mem}(t)$. **(B,C)** Simulation of the experimental result shown in Fig. 3. Note that for the four lines on top (B and C, line 1 and line 2), the signal $E(t)$ [*m*V] denotes the membrane potential averaged over the 100 *m*sec step size of the model. Above firing threshold, values of $E(t)$ also correspond to firing rates [spikes/100 *m*sec]. **(B)** Current injection of 1.3 *n*A for 300 *m*sec (bottom line). Current injection without D1 agonist application (line 1, $h' DA(t) = 0$) leads to a firing rate of about 3 spikes/100 *m*sec. The signal coding for the dopamine membrane effects $W_{mem}(t)$ remains on the initial value of zero (not shown, follows from eq. 1). With dopamine D1 agonist application (line 2, $h' DA(t) = 0.1$), evoked firing is attenuated to less than 1 spike/100 *m*sec because the value of the dopamine membrane effect signal $W_{mem}(t)$ is negative (line 3). **(C)** Current injection of 1.3 *n*A for 300 *m*sec from a sustained holding current of 0.9 *n*A (bottom line). Without dopamine D1 agonist application (line 1), the rate of evoked firing does not depend on the holding current (line 1 in B) because the dopamine membrane effect signal $W_{mem}(t)$ remains on the value of zero (not shown). With dopamine D1 agonist application (line 2, $h' DA(t) = 0.1$), evoked firing is increased to 4.5 spikes/100 *m*sec because the dopamine membrane effect signal $W_{mem}(t)$ is positive (line 3).

STRIATAL DOPAMINE MODULATION *IN VIVO*

In this section, we describe how the proposed model for membrane effects of dopamine D1 class receptor activation *in vitro* (eqs. 1-4) is adapted to simulate dopamine membrane effects *in vivo.* In addition, the model is extended to simulate dopamine-dependent long-term plasticity of corticostriatal transmission (Fig. 5A).

Dopamine Membrane Effects

In vivo, dopamine released by bursts of action potentials influences firing rates of striatal neurons primarily by dopamine D1 class receptor activation (Gonon, 1997). Indeed, the proposed model (eqs. 1-4) for dopamine D1 class receptor effects reproduces short-term dopamine membrane effects on medium spiny neurons reported in *in vivo* studies. Similar to the simulated D1 class effects, dopamine enhances or attenuate firing of striatal medium spiny neurons (reviews by Cepeda and Levine, 1998; Schultz, 1998). Moreover, striatal dopamine application *in vivo* increases the ratio between high firing rates and low firing rates, which is interpreted as enhancement of signal-to-noise ratio (Rolls, 1984; reviewed by Servan-Schreiber *et al.*, 1998a). The following argument shows that the proposed model is consistent this finding (simulation results not shown). When low firing rates are simulated by infrequent injection of brief suprathreshold current steps, the term $E(t\text{-}100)$-*reverse_potential* in eq. 1 is negative most of the time. Therefore, the membrane effect signal $W_{mem}(t)$ decreases below zero and firing decreases (as in Fig. 4B). When high firing rates are simulated with sustained suprathreshold current injection, the term $E(t\text{-}100)$-*reverse_potential* in eq. 1 is always positive, and therefore the membrane effect signal $W_{mem}(t)$ increases above zero and firing increases (as in Fig. 4C). Thus, similar to dopamine application *in vivo,* the simulated dopamine D1 class receptor activation enhances activity of neurons with high firing rates and attenuates activity of neurons with low firing rates.

For these reasons, we adapt the proposed model for membrane effects of dopamine mediated by D1 class receptors *in vitro* (eq. 1, eq. 3, and eq. 4) to simulate the membrane effects of dopamine *in vivo.* The values of the maximal membrane effect $W_{mem,max}$, maximal firing rate y_{max}, and scaling factor a for the *in vitro* condition are also used to simulate the *in vivo* condition (Table 1). Some parameter values are adapted to the *in vivo* condition. Action potential bursts of midbrain dopamine neurons *in vivo* have a much shorter influence on striatal firing than dopamine bath applications *in vitro* (Williams and Millar 1990; Gonon 1997; Hernandez-Lopez *et al.*, 1997). *In vivo,* effects of dopamine bursts on firing rates of striatal neurons decay with a rate of about 20% each 100 *m*sec (Gonon, 1997). This decay rate is reproduced by setting the value of the decay rate d of the dopamine membrane effects to 0.8. The firing threshold is set to the average value for medium spiny neurons (Table 1) (Wickens and Wilson, 1998). As for the *in vivo* simulation, the reverse potential is set to a value just below firing threshold (Table 1).

Dopamine Modulation of Corticostriatal Transmission

Activation of neocortical afferents evokes responses in neostriatal medium spiny neurons that are mediated by NMDA, AMPA and kainate glutamate receptors (Cherubini *et al.* 1988, Cepeda and Levine 1998). Synaptic responses mediated by NMDA receptors are typically small at resting potentials, but become significantly larger when the magnesium block is removed due to membrane depolarization (Nisenbaum *et al.* 1993; Kita 1996; Levine *et al.* 1996). Dopamine D1 class receptor activation seems to potentiate NMDA responses (Blank et al. 1997; Cepeda *et al.* 1998; Fienberg et al., 1998; but see Calabresi *et al.* 1997a or Nicola *et al.* 1998). In contrast, several studies reported that dopamine attenuates responses of striatal medium spiny neurons mediated by nonNMDA receptors (Cepeda *et al.* 1993; Levine *et al.* 1996; Cepeda and Levine 1998; but see Nicola *et al.* 1998). These dopamine effects may play a role in long-term plasticity of corticostriatal transmission, for instance via back propagating action potentials (Cepeda and Levine, 1998). Indeed, *in vitro* studies reported dopamine-dependent posttetanic long-term adaptations of corticostriatal transmission that depends on calcium influx throught Ca(L) channels (Calabresi *et al.*, 1992; Wickens *et al.*, 1996; Calabresi *et al.*, 1997b; Cepeda and Levine 1998). Removal

of the magnesium block was reported to reverse postetanic long-term depression to D1 receptor-dependent long-term potentiation (Calabresi *et al.* 1992; Calabresi *et al.* 1997b). Thus, long-term potentiation of corticostriatal transmission may results from dopamine D1 class receptor activation during depolarized postsynaptic membrane potentials, whereas long-term depression may results from dopamine receptor activation during hyperpolarized membrane potentials (Cepeda and Levine, 1998).

In addition to the postsynaptic membrane potential, dopamine concentration seems to influence the direction of long term adaptation in corticostriatal transmission. Tetanic stimulation of corticostriatal fibers in slices that lack dopamine agonists in the bath produces long-term depression of excitatory postsynaptic potentials, which is reversed by simultaneous pulsative dopamine application (Wickens *et al.*, 1996; Calabresi *et al.* 1997b).

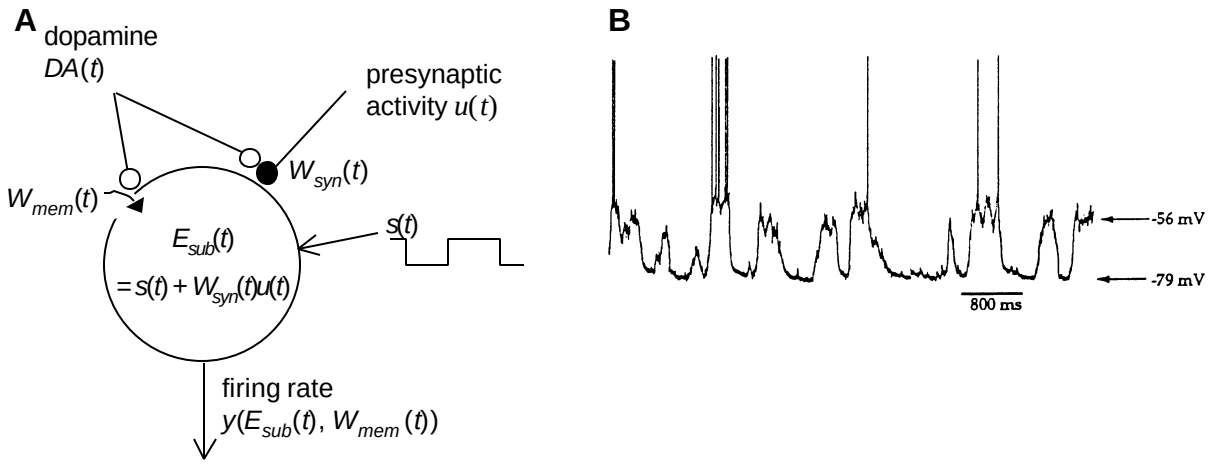

Fig. 5 (A) Model for dopamine membrane effects and synaptic effects for a medium spiny neuron *in vivo*. As in the model for the *in vivo* findings, the membrane potential-dependent effect of dopamine on D1 class receptor activation is mimicked with the dopamine membrane effect signal $W_{mem}(t)$. The corticostriatal weight $W_{syn}(t)$ is adapted according to dopamine concentration, membrane potential, and presynaptic activity. Membrane potential fluctuations are simulated with a rhythmically fluctuating signal $s(t)$. The firing rate $y(t)$ is a monotonously increasing function of the subthreshold membrane potential $E_{sub}(t)$ and the signal $W_{mem}(t)$. **(B)** *In vivo* intracellular recording of striatal medium spiny projection neuron in anesthetized rat (adapted from Stern et. al. 1997). The membrane potential fluctuates between the elevated up-state of -56 *m*V and the hyperpolarized down-state of -79 *m*V.

Model. As in previous studies, we express long-term effects of dopamine on corticostriatal transmission in the dynamics of adaptive weights (Montague *et al*., 1996; Schultz *et al*., 1997; Suri and Schultz, 1998, 1999). Since the direction of dopamine long-term effects depends both on the membrane potential of the postsynaptic medium spiny neuron and on dopamine concentration, we include these two factors in the adaptation rule. Thus, we model the long-term influence of dopamine on corticostriatal transmission with the corticostriatal weight

$$W_{syn}(t) = W_{syn}(t\text{-}100) + e\ DA(t\text{-}100)[E(t\text{-}100) - synaptic\_reverse\_potential]u(t\text{-}100). \quad \text{(eq. 5)}$$

A parameter *e* denotes the adaptation rate of the corticostriatal weight (Table 1). The signal $DA(t)$ represents the normalized dopamine concentration. This signal is zero for average physiological concentrations, negative for lower and positive for higher concentrations. As long-term adaptation in corticostriatal transmission presumably requires presynaptic activity, the adaptation of the corticostriatal weight $W_{syn}(t)$ is proportional to the presynaptic firing rate $u(t\text{-}100)$. $E(t)$ denotes the membrane potential of the striatal medium spiny neuron (eq. 4). The synaptic reverse potential is the membrane potential for which synaptic adaptation switches its sign and should not be confused with the biophysically defined synaptic reversal potential. Since dopamine membrane effects may mediate long-term adaptations in corticostriatal transmission, eq. 5 is somewhat similar to eq. 1, and the value of the synaptic reverse

potential is set to a similar value as the reverse potential for dopamine membrane effects (Table 1). In contrast to eq. 1, we assume in eq. 5 that the synaptic adaptations do not decay during the simulated task, and therefore we do not introduce a decay rate (the decay rate is one).

Striatal Membrane Potential Fluctuations

The effect of dopamine on striatal medium spiny neurons crucially depends on their membrane potential, which fluctuates *in vivo* with a frequency of about 1.2 Hz between the hyperpolarized down-state and the depolarized up-state (Fig. 5B) (Wilson and Kawaguchi 1996; Stern *et al*. 1997; Wickens and Wilson 1998). Membrane potential fluctuations were usually recorded in anesthetized animals but also occur in awake animals with a less regular frequency (Wilson and Groves, 1981; Wilson, 1993). Increased activity of corticostriatal neurons seems to induce transitions to the up-state and to maintain the up-state (Wilson and Kawaguchi 1996; Stern *et al.* 1997). *In vivo* stimulation of the cortex resets the phase of the oscillation (Katayama et al., 1980). Membrane effects of dopamine D1 agonists seem to prolong down-state duration (Cooper and White 1998) by influencing an outward potassium current (Wilson 1992; Surmeier and Kitai 1993; Kitai and Surmeier 1993) and to prolong up-state duration as they increase the firing rate (Hernandez-Lopez *et al.*, 1997). These findings suggest that up-state durations are prolonged and down-state durations are shortened by influences that increase the membrane potential or the firing rate. Furthermore, up-state durations seem to be shortened and down-state durations prolonged by influences that decrease the membrane potential or the firing rate.

Model. When dopamine membrane effects and corticostriatal weights are at baseline levels ($W_{mem}(t) = 0$ and $W_{syn}(t)u(t) = 0$), the simulated medium spiny neurons switch their state each 400 *m*sec. As suggested by experimental evidence, excitatory dopamine-dependent effects ($W_{mem}(t) > 0$ or $W_{syn}(t)u(t) > 0$) prolong the duration of the up-state and shorten the duration of the down-state, whereas inhibitory dopamine-dependent effects ($W_{mem}(t) < 0$, $W_{syn}(t)u(t) < 0$) have the opposite effects on state durations. This is accomplished with the fluctuating function

$$s(t) = \begin{cases} E_{up}, \text{ during } 400 \text{ msec} + [W_{mem}(t) + W_{syn}(t)u(t)]\,j \\ E_{down}, \text{ during } 400 \text{ msec} - [W_{mem}(t) + W_{syn}(t)u(t)]\,j\,. \end{cases} \qquad \text{(eq. 6)}$$

Parameters E_{up} and E_{down} denote the average membrane potentials in the up-state and in the down-state, respectively (Table 1). A parameter j scales the influence of dopamine membrane effects $W_{mem}(t)$ and corticostriatal activation $W_{syn}(t)u(t)$ on the typical state duration of 400 *m*sec (Table 1).

Membrane Potential of Medium Spiny Neurons

Below firing threshold, we assume that the subthreshold membrane potential $E_{sub}(t)$ is determined by membrane potential fluctuations $s(t)$, the presynaptic activity $u(t)$, and the corticostriatal weight $W_{syn}(t)$ with

$$E_{sub}(t) = s(t) + W_{syn}(t)\,u(t). \qquad \text{(eq. 7)}$$

To avoid decreasing the simulated membrane potential $E_{sub}(t)$ below physiological values, it is limited to values above the constant E_{min} (Table 1). For simulations of *in vivo* conditions, eq. 7 replaces eq. 2, since eq. 2 describes the current injection *in vitro*.

BASAL GANGLIA-THALAMUS-CORTEX

In this section we present the model for the influence of sensory stimuli on activity in striatal matrisomes and, via basal ganglia-thalamocortical pathways, on motor acts (Fig. 1B). Since only two acts are important for the simulated task, two striatal medium spiny neurons are simulated, and each of them can elicit one act. Dopamine effects on corticostriatal transmission and dopamine membrane effects on the firing rates $y_1(t)$ and $y_2(t)$ of the two simulated medium spiny neurons are simulated by using for each neuron the eqs. 1, and 3-7. Dopamine concentration $DA(t)$ (eqs. 1 and 5) is simulated with the Extended TD model which will be presented in the next section.

Since the striatal striosomes receive projections from somatosensory cortices that report about sensory stimuli (Alexander *et al.*, 1986), we assume that the firing rates of corticostriatal neurons are proportional to delayed versions of the visual stimuli. Thus, we replace the presynaptic activity $u(t)$ in eqs. 5-7 with a delayed representation of the visual input (see next section "Parameter Values and Initial Conditions")..

To simplify the model equations, the firing rates in GPi/SNr, thalamus, and cortex are simulated as signals that are proportional to the difference of the firing rate from the baseline. Therefore, tonic firing rates of the real neurons are ignored, and the baseline and the amplitude of the simulated firing rates are chosen arbitrarily.

The model does not account for direct interactions between medium spiny neurons, as the physiological effects of collaterals between medium spiny neurons was suggested to be weak (Jaeger *et al.*, 1994). The two simulated striatal neurons inhibit the simulated firing rates $GPi\_SNr_1(t)$ and $GPi\_SNr_2(t)$ of the two neurons in the basal ganglia output nuclei globus pallidus interior and substantia nigra pars compacta. Since there is strong convergence from cortex via striatal matrisomes to the basal ganglia output nuclei (Alexander *et al.*, 1986), only the predominant striatal activity may be represented in the firing rates of the output nuclei. We implement a selection mechanism according to the proposal of Berns and Sejnowski (1996), who demonstrated in simulation experiments that this selection may be caused by the projections of the indirect pathway via the globus pallidus exterior (GPe) and the subthalamic nucleus (STN). Their results suggest that the indirect pathway selects the predominant representations by disinhibiting minor suppressions of neural activities in the basal ganglia output nuclei. For the current study, we assume that firing rates $GPi\_SNr_1(t)$ and $GPi\_SNr_2(t)$ of two simulated neurons in the basal ganglia output nuclei are inhibited by striatal firing rates $y_1(t)$ and $y_2(t)$, but at baseline levels when both striatal neurons fire ($y_1(t) > 0$ and $y_2(t) > 0$):

$$GPi\_SNr_n(t) = \begin{cases} 0, \text{ if } y_1(t) > 0 \text{ and } y_2(t) > 0 \\ -y_n(t), \text{ else.} \end{cases} \quad \text{(for } n = 1, 2\text{).} \quad \text{(eq. 8)}$$

These neurons inhibit two neurons in the thalamus:

$$thalamus_n(t) = -GPi\_SNr_n(t) \quad \text{(for } n = 1, 2\text{).} \quad \text{(eq. 9)}$$

The thalamus elicits cortical firing rates $u_5(t)$ and $u_6(t)$ that are proportional to the salience parameter a (Table 1):

$$u_5(t) = a * thalamus_1(t),\ u_6(t) = a * thalamus_2(t). \quad \text{(eq. 10)}$$

These cortical firing rates serve as inputs for the Extended TD model.

An act is elicited when the thalamic activation of motor cortical areas is substantial and persistent enough. To take the duration and the amount of the thalamic activity into account, cortical activity is computed from the thalamic signal with the leaky integrator

$$cortex_{integr,n}(t) = l_{act} \times cortex_{integr,n}(t\text{-}100) + thalamus_n(t), \quad \text{(eq. 11)}$$

where the initial values of the cortical activations equal zero ($cortex_{integr,n}(t = 0)$, for $n = 1$ or 2) and l_{act} denotes an integration constant (Table 1). As in previous model equations, this equation is expressed using a step size of 100 msec.

The two acts are coded by the binary signals $act_1(t)$ (act left) and $act_2(t)$ (act right). We say that the model is executing an act when the corresponding act signal equals one. The act with number n is executed when the signal $cortex_{integr,n}(t)$ is above the threshold parameter act_{thres}:

$$act_n(t+100) = \begin{cases} 1, \text{ if } cortex_{integr,n}(t) > act_{thres} \\ 0, \text{ else.} \end{cases} \quad \text{(for } n = 1, 2\text{)} \quad \text{(eq. 12)}$$

If both signals $cortex_{integr,1}(t)$ and $cortex_{integr,2}(t)$ are above act_{thres}, then the act corresponding to the larger signal is selected. Since we assume that a selected act persists for 200 *m*sec, $act_n(t)$ is set to the value of one for 200 *m*sec.

Parameter Values and Initial Conditions

Since some parameter values cannot be determined from experimental findings, they are set to values that are suitable to solve the task (Table 1). The fluctuating function $s(t)$ (eq. 6) is randomly set to the up-state value E_{up} or the down-state value E_{down} at the beginning of each trial for each neuron, since membrane potential fluctuations are only weakly periodic and intertrial intervals are assumed to be sufficiently long.

In the simulated task, the model should learn sensorimotor associations between stimulus blue and the two acts. Stimulus blue is coded by the signal $u_1(t)$ that is one when stimulus blue is present and zero when it is absent. We assume that the firing rate of some corticostriatal neurons is proportional to a delayed version of $u_1(t)$. Thus, we substitute the presynaptic activity $u(t)$ in eqs. 5-7 with the delayed stimulus blue ($u(t) = u_1(t\text{-}100)$). Since associations of the stimuli green and red with the acts are not required, the simulated neurons do not receive inputs reporting about these stimuli.

The simulated task requires that the presentation of stimulus blue is often followed by act left or act right during the exploration phase. To induce execution of these exploratory acts, the two corticostriatal weights $W_{syn,left}(t=0)$ and $W_{syn,right}(t=0)$ that associate stimulus blue with the firing rate of the two striatal neurons are initialized with a positive value (Table 1).

Table 1. Standard model parameters for striatal medium spiny neurons *in vivo* and basal ganglia-thalamocortical pathway

	Name	Symbol	Value	Choice of Value
Membrane Effects	decay of dopamine membrane effects	d	0.8 / 100 *m*sec	Gonon (1997)
	maximal absolute value of the membrane effect $W_{mem}(t)$	$W_{mem,max}$	9	from *in vitro* simulation
	initial value of the membrane effects $W_{mem,1}(t)$ and $W_{mem,2}(t)$	$W_{mem,1}(t=0)$ $W_{mem,2}(t=0)$	0	from *in vitro* simulation
	maximal firing rate	y_{max}	6 spikes / 100 *m*sec	Apicella *et al*., (1992), Nisenbaum *et al*. (1994)
	scaling factor	a	0.3 Spikes $\times$ (100 *m*sec * *m*V)$^{-1}$	Nisenbaum *et al*. (1994)
	average firing threshold	*firing_ threshold*	-46 *m*V	Wickens and Wilson (1998)
	reverse potential of membrane effect	*reverse_ potential*	-48 *m*V	just below firing threshold
	amplitude of dopamine membrane effects	h	180	appropriate for planning
Corticostriatal Transmission	learning rate for corticostriatal learning	e	0.45	chosen to induce sensorimotor learning in a few trials
	synaptic reverse potential	*synaptic_ reverse_ potential*	-41 *m*V	appropriate for simulated task
	initial corticostriatal weights stimulus blue – act left and stimulus blue – act right	$W_{syn,left}(0)$, $W_{syn,right}(0)$	5	to induce a few exploratory acts in exploration phase
Membrane Potential Fluctuations	average membrane potential in up-state	E_{up}	-47 *m*V	Stern *et al*. (1997)
	average membrane potential in down-state	E_{down}	-69 *m*V	Stern *et al*. (1997)
	lower limit of membrane potential	E_{min}	-82 *m*V	Stern *et al*. (1997)
	dopamine modulation of up and down-states durations	j	0.1	to get a small but significant effect on model performance as compared to $j = 0$
Thalamus-Cortex	integration constant for leaky integrator	l_{act}	0.7 / 100 *m*sec	for plausible reaction time in planning
	act threshold	act_{thres}	2.6	for plausible reaction time in planning
	salience of thalamic activity for Extended TD model	a	0.1	appropriate for successful planning

EXTENDED TD MODEL

TD Model

In Pavlovian learning paradigms, animals learn to anticipate reward delivery. In addition, they are often able to estimate the time of reward occurrence (Gallistel, 1990). In order to implement a time estimation mechanism, the TD model of Pavlovian learning (Sutton and Barto, 1990) assumes that the nervous system represents each stimulus with a series of short components following stimulus onset. This is achieved by mapping each stimulus to a fixed temporal pattern of phasic signals $x_1(t)$, $x_2(t)$, … that follow stimulus onset with varying delays. This temporal pattern is referred to as a "serial compound stimulus" or "temporal stimulus representation" (Fig. 6A). The temporal stimulus representation is used to compute the reward prediction signal with $p(t) = \sum_m v_m(t) \cdot x_m(t)$, where $v_m(t)$ are the adaptive weights (Sutton and Barto, 1990; Montague *et al*., 1996; Schultz *et al*., 1997; Suri and Schultz, 1998). A representation of the TD model using a neuron-like element is shown in Fig. 6B. The reward prediction develops during learning in a similar way as the animal's anticipatory behavior and shows a sustained increase before a predicted reward. According to the TD model, the animal's anticipatory response, and therefore also the reward prediction, increase gradually before an anticipated reward if this reward is completely predicted. The rate of this gradual increase is determined by the constant g, which is referred to as the temporal discount factor.

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 (Sutton & Barto, 1990). Thus, the TD model computes the reward prediction error $e(t)$ from discounted temporal differences in the prediction signal $p(t)$ and from the reward signal with the equation $e(t) = reward(t\text{-}100) - [p(t\text{-}100) - g\ p(t)]$ (time t in m*sec*, 100 m*sec* is the step size of the model implementation). The reward prediction error is phasically increased above base line levels of zero for rewards and reward-predicting stimuli if these events are unpredicted but remains on base line levels if these events are predicted. In addition, if a predicted reward is omitted, the reward prediction error decreases below base line levels at the time of the predicted reward when the predicted reward fails to occur. The TD model learns to predict the time of reward occurrence due to the temporal stimulus representation. These characteristics of the reward prediction error correspond to characteristics of dopamine neuron activity (Montague *et al*., 1996; Schultz *et al*., 1997; Suri and Schultz, 1998, 1999).

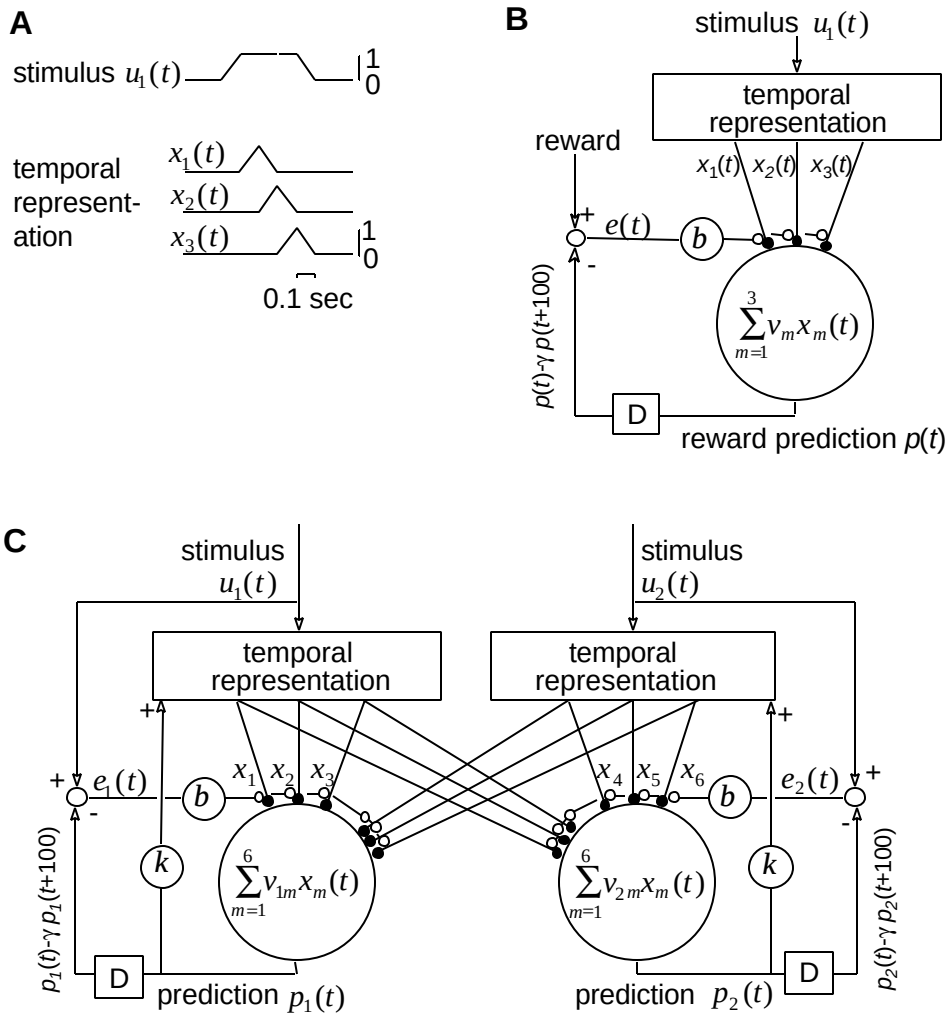

Fig. 6 Critic model. **(A)** Temporal stimulus representation $x_1(t)$, $x_2(t)$, and $x_3(t)$. Stimulus $u_1(t)$ is represented over time as a series of phasic signals $x_1(t)$, $x_2(t)$, and $x_3(t)$ that cover stimulus duration. This temporal stimulus representation is used to reproduce the finding that dopamine neuron activity is decreased when a predicted reward fails to occur. **(B)** TD model. From stimulus $u_1(t)$ the temporal stimulus representation $x_1(t)$, $x_2(t)$, and $x_3(t)$ 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. The difference operator D takes temporal differences from this prediction signal (discounted with factor g). The reward prediction error $e(t)$ is computed from these temporal differences and from the reward signal. The weights $v_m(t)$ are adapted proportionally to the prediction error signal $e(t)$ and to the learning rate b. **(C)** Extended TD model for two input events $u_1(t)$ and $u_2(t)$. The event signals $u_k(t)$ report about stimuli, rewards, thalamic activity, and acts. Each temporal representation component $x_m(t)$ is multiplied with an adaptive weight v_{km} (filled dots). Event prediction $p_k(t)$ is computed from the sum of the weighted components. Event prediction $p_k(t)$ is multiplied with a small constant k and fed back to the temporal event representation of this event $u_k(t)$. This feedback is necessary to form novel associative chains. Analogous to the TD model, the prediction error $e_k(t)$ is computed from the event $u_k(t)$ and from the temporal differences between successive predictions $p_k(t) - g\, p_k(t+100)$ (discounted with a factor g). The weights v_{km} (filled dots) are adapted as in the TD model.

Extended TD Model

Humans and animals associate sensory events (stimuli, rewards or behavioral responses) with other sensory events and use these associations to form novel associative chains (Craik 1943; Piaget 1954; Mackintosh 1974; Arbib 1972; Dickinson 1980; Wolpert *et al.* 1995). However, the TD model is limited to associations between stimuli and one type of reward and does not form novel associative chains. Therefore, the TD model has been extended to learn predictions for behaviorally relevant stimuli and for different reinforcers (Sutton and Barto, 1981; Sutton & Pinette 1985). In order to form novel associative chains, this approach learns an "internal model of its environment" that emulates the temporal development of real world processes. Since there is evidence that dopamine neuron activity is influenced by the formation of novel associative chains (Young *et al.*, 1998), dopamine neuron activity was modeled with such an internal model approach (Suri and Schultz, submitted). We adapt this internal model approach and call it here the Extended TD model. This model is shown for two input stimuli in Fig. 6C.

The Extended TD model learns to predict all its input signals: the signals $u_1(t)$, $u_2(t)$, and $u_3(t)$ coding for the stimuli green, red, and blue, respectively; the reward signal $u_4(t) = reward(t\text{-}100)$; the thalamic activities $u_5(t) = a \times thalamus_1(t)$ and $u_6(t) = a \times thalamus_2(t)$; the acts $u_7(t) = act_1(t)$ and $u_8(t) = act_2(t)$ (Table 2). Sensory events (acts, stimuli, reward) are coded as signals with a value of one when they are present and zero when they are absent, except stimulus blue in the Extended TD model. This stimulus corresponds to the start box in the *T*-maze. Since this place serves as a known context, novelty responses of dopamine neurons should be smaller than those elicited by the stimuli red and green (see Introduction). Therefore, presence of stimulus blue is coded with a signal of a small positive value, referred to as the "salience of stimulus blue" (Table 3).

As in the TD model, each stimulus is represented with a series of stimulus representation components that cover the duration of its presentation in order to learn to predict when the stimulus occurs (Fig. 6A). Since the stimuli green and red are presented for 300 *m*sec, both are represented with three temporal representation components. Stimulus green is represented with the representation components $x_1(t)$, $x_2(t)$, and $x_3(t)$ and stimulus red with the representation components $x_4(t)$, $x_5(t)$, and $x_6(t)$. Stimulus blue, which is presented for at most 600 *m*sec, is represented with the components $x_7(t)$, $x_8(t)$,…, $x_{12}(t)$ (Table 2). The temporal representation components of stimulus blue are set to the value of the salience of stimulus blue when they become active. The values of all components are set to zero when stimulus blue gets extinguished.

The stimuli green and red are coded with the binary signals $u_1(t)$ and $u_2(t)$, respectively. Temporal representations follow stimulus presentations with a delay of 100 *m*sec to account for processing delays. The reward is not represented over time, as the reward signal $reward(t)$ is only nonzero during 100 msec when the reward is presented. Also the thalamic signal is not represented over time, because this internal signal is not binary and therefore does not have clear onsets and offsets. Instead of representing these signals over time, the reward signal is copied with a delay of 100 msec to the input representation component $x_{13}(t) = reward(t\text{-}100)$, and the thalamic signals to the representation components $x_{14}(t) = a \times thalmus_1(t)$, and $x_{15}(t) = a \times thalmus_2(t)$. Since acts usually lead to sensory stimuli during their execution, they are also represented over time during their duration. As each of the two acts $act_1(t)$ and $act_2(t)$ persisted for 200 msec, $act_1(t)$ is represented with the components $x_{16}(t)$ and $x_{17}(t)$, and $act_2(t)$ with the components $x_{18}(t)$ and $x_{19}(t)$ (Table 2).

Table 2. Definitions of Critic input signals $u_k(t)$ and their temporal representation components $x_m(t)$ (see section "Extended TD Model").

	Signal	Temporal representation components
Green goal box (stimulus green)	$u_1(t)$	$x_1(t)$, $x_2(t)$, $x_3(t)$
Red goal box (stimulus red)	$u_2(t)$	$x_4(t)$, $x_5(t)$, $x_6(t)$
Start box (stimulus blue)	$u_3(t)$	$x_7(t)$, $x_8(t)$,…, $x_{12}(t)$
Reward	$u_4(t)$ = reward(t-100)	$x_{13}(t)$
Thalamic activity related to act 1	$u_5(t) = \alpha \times \text{thalamus}_1(t)$	$x_{14}(t)$
Thalamic activity related to act 2	$u_6(t) = \alpha \times \text{thalamus}_2(t)$	$x_{15}(t)$
Act left (act 1)	$u_7(t) = \text{act}_1(t)$	$x_{16}(t)$, $x_{17}(t)$
Act right (act 2)	$u_8(t) = \text{act}_2(t)$	$x_{18}(t)$, $x_{19}(t)$

To form novel associative chains, a predicted event should elicit similar prediction signals as does the experience of this event. More precisely, a predicted stimulus should produce internal representation signals that resemble the representation of the stimulus itself. Therefore, Suri and Schultz (submitted) proposed that the prediction of each stimulus is fed back to the temporal stimulus representation of this stimulus and used to estimate further prediction signals. The loop time t of this feedback is assumed to be much shorter than the usual 100 *m*sec step size of the model because the feedback is computed twice within each time step.

The prediction $p_k(t - 2t)$ for the input signal $u_k(t)$ (k = 1, 2,…, 8) is computed from the product of the adaptive weights $v_{km}(t)$ with the components of the temporal representation components $x_m(t)$:

$$p_k(t - 2t) = \sum_{m=1}^{19} v_{km}(t) \cdot x_m(t). \qquad \text{(eq. 13a)}$$

This prediction is fed back twice to the temporal stimulus representation with the two equations:

$$p_k(t - t) = \sum_{m=1}^{19} v_{km}(t) \cdot [x_m(t) + k\, s_{km}\, p_k(t - 2t)], \qquad \text{(eq. 13b)}$$

$$p_k(t) = \sum_{m=1}^{19} v_{km}(t) \cdot [x_m(t) + k\, s_{km}\, p_k(t - t)]. \qquad \text{(eq. 13c)}$$

To avoid very large absolute values of the prediction signals $p_k(t)$, these signals are limited to values between $-p_{max}$ and $+p_{max}$ (Table 3). We do not assign a value to the small time constant t, as t occurs only in eqs. 13a and 13b and prediction signals will be shown in the figures for time steps of 100 *m*sec. The feedback constant k (Table 3) determines the gain of the feedback loop and therefore the impact of a predicted stimulus on further stimulus predictions. The number of these update equations seems to correspond to the number of novel links in the associative chain the model can compute (unpublished result). We are neither aware of mathematical considerations nor of experimental evidence that would indicate which components of the temporal stimulus representation $x_m(t)$ should be influenced by this feedback. For simplicity, we assume that the feedback influences only the first component of the temporal stimulus representation. Feedbacks to further temporal stimulus representation components do not influence most simulation results but lead to slightly different time courses of prediction signals in simulations that test the formation of novel associative chains (unpublished result). Feedback to the first component of the temporal stimulus representation is accomplished by setting the factor s_{km} to one for the first component of the temporal stimulus representation of each stimulus or act. Also for the reward and the two thalamic signals s_{km} is set to 1. Otherwise, the factor s_{km} is set to zero ($s_{1,1}$ = 1, $s_{2,4}$ = 1, $s_{3,7}$ = 1, $s_{4,13}$ =1, $s_{5,14}$ = 1, $s_{6,15}$ = 1, $s_{7,16}$ = 1, $s_{8,18}$ = 1; s_{km} = 0 otherwise).

The following equations of the Extended TD model are analogous to those of the TD model. Therefore, the proposed Extended TD model with the parameter k = 0 is equivalent to a set of eight

independent TD models. For this case, the equations are for each k equivalent to those of the TD model. The prediction errors $e_k(t)$ are computed from discounted temporal differences between successive predictions (differencer D in Fig. 6C) and from the input signals $u_k(t)$ with

$$e_k(t) = u_k(t\text{-}100) - [p_k(t\text{-}100) - \gamma p_k(t)], \qquad \text{(eq. 14)}$$

where γ is the discount factor and the input signals $u_k(t)$ (k = 1, 2, …, 8) denote the three stimuli, the reward, the thalamic signals (multiplied with salience α), and the two act signals. The value of the discount factor γ is set to 0.98, because this value was estimated from dopamine neuron activity (Suri and Schultz, 1999). To minimize the prediction error signals, the weights $v_{km}(t)$ are incrementally adapted according to the product of the input prediction errors $e_k(t)$ with the eligibility traces of the temporal input representation $\bar{x}_m(t)$ with

$$v_{km}(t+100) = v_{km}(t) + \beta e_k(t) \bar{x}_m(t). \qquad \text{(eq. 15)}$$

The three weights v_{km} that associate the first component of the temporal representations of the three stimuli with the reward are initiated with the positive value v (Table 3) to reproduce dopamine novelty responses (Suri and Schultz, 1999). The other values of matrix v_{km} are initialized with zeros ($v_{4,1}(t=0) = v$, $v_{4,4}(t=0) = v$, $v_{4,7}(t=0) = v$; $v_{km}(t=0) = 0$ otherwise). For the current study, the positive value v is chosen as small as possible to reduce the number of errors due to exploration, but large enough to significantly increase the number of acts in the exploration phase.

The traces $\bar{x}_m(t)$ are slowly decaying versions of the input representation components $x_m(t)$. Such eligibility traces were introduced to explain how animals learn to associate sensory events that are separated by a delay period (Sutton and Barto, 1990). Although TD models with temporal stimulus representations learn to associate sensory events over a delay without representation traces (Montague *et al.*, 1996), traces accelerate learning (Sutton & Barto, 1998; Kearns and Singh, submitted). At the beginning of an experiment, $\bar{x}_m(t)$ is set to the initial condition $\bar{x}_m(0) = 0$. Then, the traces are computed with

$$\bar{x}_m(t) = \lambda_c \bar{x}_m(t\text{-}100) + (1\text{-}\lambda_c)\, x_m(t). \qquad \text{(eq. 16)}$$

The parameter λ_c is set to the value of 0.3, as this value guarantees fast learning. With this parameter value, the eligibility traces increase with a rate of 30% each 100 *m*sec during presentation of the event and decrease 70% each 100 *m*sec after event presentation.

The output signal of the Extended TD model is the reward prediction error $e_4(t)$ (eq. 14) that resembles the firing rate of dopamine neurons (Suri and Schultz, submitted). Since dopamine concentration in extracellular space is closely time-correlated with the firing rate of dopamine neurons (Gonon, 1997), we compute the dopamine concentration $DA(t)$ with

$$DA(t) = e_4(t). \qquad \text{(eq. 17)}$$

The simulated dopamine concentration $DA(t)$ is used to simulate the effect of dopamine on medium spiny neurons in striatal matrisomes (eqs. 1 and 5).

Table 3 Standard model parameters in Extended TD model

Parameter Name	Symbol	Value
initial weights for novelty responses	v	0.001
Critic learning rate	b	0.5
feedback constant	k	0.8
maximal value of prediction signals	p_{max}	10
discount factor	g	0.98
decay of eligibility trace	l_c	0.3
salience of stimulus blue for Critic		0.05

RESULTS

The proposed model was tested in the experiment described above. Since each trial started with a random state of the striatal membrane potentials, the model performed differently in each experiment. We show the model performance for a typical experiment and then present the statistical analysis of 1000 experiments.

In the exploration phase, the model learns to associate act left with stimulus red and act right with stimulus green. In the first trial (Fig. 7A), stimulus blue was presented and the model executed the act left (bottom line) that led to presentation of stimulus red (line 1). Since certain associative weights of the Extended TD model had been initialized with positive values, this novel stimulus phasically activated the reward prediction signal (line 2; eq. 13). This led to a biphasic response of the dopamine-like reward prediction error signal (line 3; eq. 14) resembling dopamine novelty responses. Since the salience of stimulus blue had been set to a smaller value than that of the stimuli red and green (see section "Extended TD Model" and Table 3), onset of the stimulus blue led to very small activations of the reward prediction signal and the reward prediction error signal (hardly visible). Since stimulus green was not presented, the prediction signal for stimulus green remained zero (line 4). The simulated striatal membrane potentials $E_{left}(t)$ and $E_{right}(t)$ of the two striatal medium spiny neurons in the matrisome compartment fluctuated each 400 *m*sec between the elevated up-state and the hyperpolarised down-state (line 5; eqs. 4 and 7). The membrane potentials were slightly increased during presentation of stimulus blue, since corticostriatal weights associating stimulus blue with striatal activity were set to positive initial values (compare section "Parameter Values and Initial Conditions"). As action potentials are much shorter than the 100 *m*sec time step, the averaged membrane potential is shown (eq. 4, as in Fig. 4B and 4C). The membrane potential of the striatal neuron coding act left was increased above firing threshold for 500 *m*sec. This persistent firing was integrated by two neurons in motor cortex (line 6; eq. 11). When the firing rate of the neuron coding for act left reached the act threshold act_{thres}, this act was elicited (bottom line; eq. 12). The signal coding for the act right remained on the value of zero (not shown).

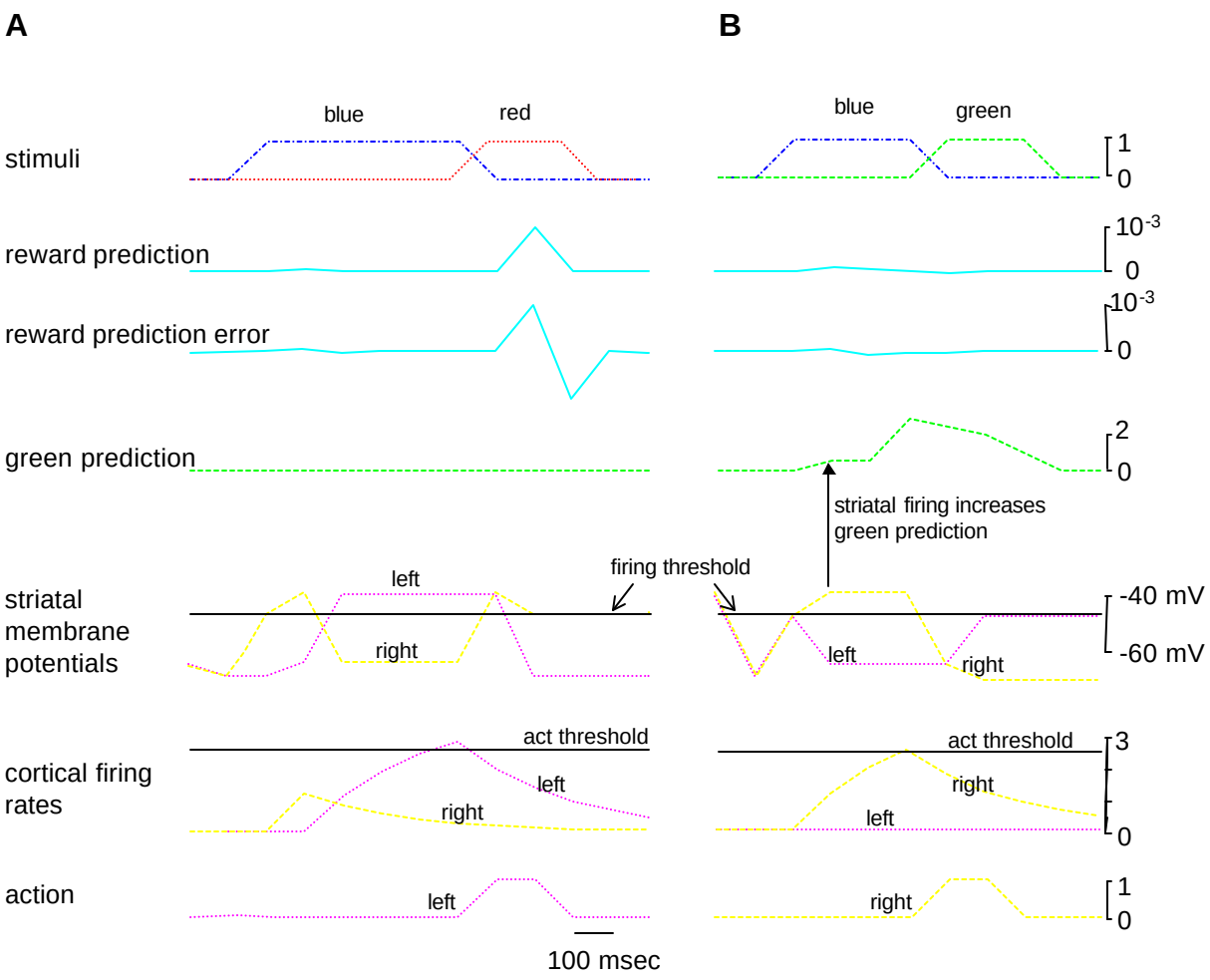

Fig. 7 Model performance during exploration phase. **(A)** First trial. When stimulus blue was presented (line 1), the model elicited the act left (bottom line) that led to presentation of stimulus red (line 1). Since stimulus red was presented for the first time, its onset phasically activated the reward prediction signal (line 2) and biphasically activated the dopamine-like reward prediction error signal (line 3). Membrane potentials of the two simulated striatal medium spiny neurons fluctuated between an elevated up-state and a hyperpolarized down-state (line 5). During presentation of stimulus blue, the simulated striatal neuron coding for act left was firing for 500 *m*sec. Neurons in motor cortex integrated this striatal firing rate over time (line 6). The act left was elicited (bottom line) when the integrated signal reached a threshold. **(B)** A trial at the end of the exploration phase. When stimulus blue was presented (line 1), the model elicited the act right (bottom line) that led to presentation of stimulus green (line 1). Since stimulus green had been presented repeatedly during the exploration phase, novelty responses were almost absent in the reward prediction signal (line 2) and in the dopamine-like reward prediction error signal (line 3). Prediction of stimulus green (line 4) was already increased when the striatal neuron coding for the act right increased its firing rate (line 5), because this had often antedated execution of act right followed by presentation of stimulus green. The striatal firing rates were integrated in cortex and the act right was elicited (bottom line) when the cortical signal coding for the act right reached a threshold (line 6).

For the next 80 presentations of stimulus blue, the model executed 11 times the act left, 14 times the act right, and 55 times no act (not shown). Trials without acts occurred when striatal membrane potentials of both neurons happened to fluctuate synchronously, as the effects of synchronous striatal firing on the cortical neurons were suppressed by the indirect pathway (eq. 8). The 81st presentation of stimulus blue was the last blue presentation in the exploration phase during which an act was executed (Fig. 7B). Since the model selected act right, stimulus green was presented (line 1). Reward prediction and reward prediction error remained on the values of zero, since dopamine-like novelty responses had

extinguished as a consequence of the Critic learning rule (lines 2 and 3, eq. 15). The green prediction signal was slightly activated when the striatal neuron coding the act right was firing (line 4, line 5), as such increased striatal firing had been followed by the corresponding act right in some, but not all, previous trials. Striatal activity influenced the green prediction signal via basal ganglia output nuclei, thalamus, and cortex (via salience a, Fig. 1B, eq. 10). Since act right had previously been followed by green presentations and an efference copy of the act signal reached the Critic, the green prediction signal was fully activated when the act right was executed (bottom line), (Fig. 1B, eq. 13a). The green prediction peaked at the correct value of three, as this value reflects the predicted future duration of stimulus green in units of 100 *m*sec. The striatal firing rates were integrated over time in cortical neurons, and the act right was executed (bottom line) when the cortical signal coding for act right reached the act threshold act_{thres} (line 6).

The rewarded phase consisted of only one trial (Fig. 8) in which presentation of stimulus green (line 1) was followed by reward presentation (line 2). The model did not execute an act during this trial, as the initial values of the corticostriatal weights disfavored acts when stimulus blue was absent (see section "Parameter Values and Initial Conditions"). The reward was unpredictable, as it was presented for the first time. Therefore, the reward prediction error (line 3) was equal to the reward signal (eq. 14). The temporal representation of stimulus green in the Critic consisted of three phasic components (line 6-8, Fig. 6A). The peak of the first component $x_1(t)$ followed presentation of green with a delay of 100 *m*sec and the peaks of the two further components with delays of 200 *m*sec and 300 *m*sec. From each of these components, eligibility traces were computed that decayed with the rate l_c to zero (lines 9-12, eq. 16). The adaptive weights $v_{41}(t)$, $v_{42}(t)$, and $v_{43}(t)$ (three lines at bottom), where the number 4 codes for the reward, associate the components $x_1(t)$, $x_2(t)$, and $x_3(t)$ of the temporal representation of stimulus green with the reward prediction signal (eqs. 13a-13c). Each of these weights was initialized with a value of zero and was adapted proportionally to the product of the trace of its component $\bar{x}_1(t)$, $\bar{x}_2(t)$, or $\bar{x}_3(t)$ with the reward signal (eq. 15). Therefore, the closer the activation of the component was to the reward, the larger was the increase in the weight associating this component to the reward. Likewise, the Critic learns the associative weights $v_{km}(t)$ between the other input events (stimuli, reward, acts, thalamic firing rate; Fig. 6C).

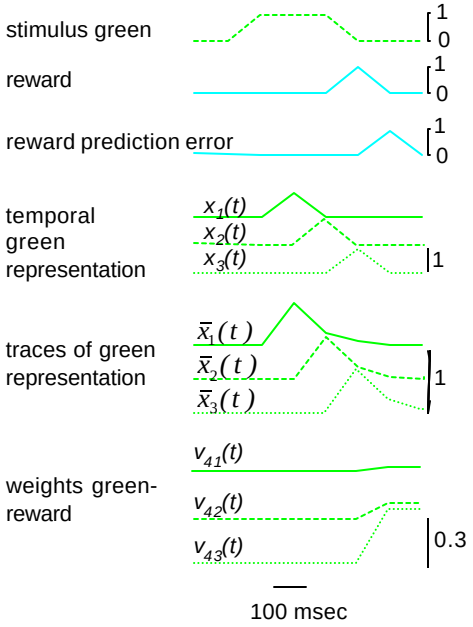

Fig. 8 Associative learning during rewarded phase. In this second phase, presentation of stimulus green (line 1) was followed by presentation of the reward (line 2) and no act was executed. Since the reward was unpredictable, the reward prediction error (line 3) was equal to the reward signal. The three components of the temporal representation of stimulus green were phasic signals with peaks following green onset with delays of 100 *m*sec, 200 *m*sec, and 300 *m*sec (lines 4-6). For each component an eligiblility trace was computed (lines 7-9) that was used to adapt the weight that associated this component with the reward (three lines at bottom). (All signals shown in this figure start with a value of zero.)

Planning was assessed in the first trial of the test phase, as this trial tested the model's ability to select the correct act right based on the formation of a novel associative chain. At the beginning of the test phase, stimulus blue was presented twice without act executions (not shown). These blue presentations were not counted as trials. In the first trial (Fig. 9A), presentation of stimulus blue (line 1) was responded to with the correct act right (bottom line). This correct act was selected, because of a positive feedback between the striatal signal for the act right (line 8) and the dopamine-like reward prediction error (line 5). In this trial, the striatal neuron coding the act right happened to stay in the elevated up-state during presentation of stimulus blue for several hundreds of milliseconds (line 8). (If instead the neuron coding for the act left had happened to stay in the up-state, this would not have increased the dopamine-like reward prediction error and no act might have been elicited.) This persistent striatal firing reached the Critic via basal ganglia output nuclei and thalamus (via salience a, Fig. 1B, eq. 10). Since in the exploration phase the model had learned to associate such striatal firing coding for the act right with presentation of stimulus green, the green prediction signal increased slightly (line 2). Because the model formed novel associative chains for positive values of the feedback parameter k, this activation in the green prediction signal served as the stimulus green itself (Fig. 6C; eqs. 13a-c). Since the model had learned in the rewarded phase to associate stimulus green with the reward, the reward prediction (line 4) and therefore the dopamine-like reward prediction error were slightly activated (first small peak in line 5, eq. 14). This activation in the reward prediction error increased the signal representing dopamine membrane effects (line 6, eq. 1) and the corticostriatal weight (line 7, eq. 5) of the striatal neuron coding for the act right. In addition, both signals for the striatal neuron coding the act left were decreased, since the membrane potential of this neuron was below both reverse potentials. These adaptations increased the firing rate of the striatal neuron coding for the act right and decreased the membrane potential of the neuron coding for act left (line 8, eqs. 3 and 7). This increase in striatal firing rate was integrated in cortical neurons (line 9), and the correct act right was executed (bottom line) when the integrated cortical signal reached the act threshold act_{thres}.

In the following 60 simulated trials of the test phase the correct act right was executed, except in the incorrect trial 3 (not shown). In trial 19 (Fig. 9B), the prediction error signals for the stimulus green (line 3) and for the reward (line 5) were phasically increased at onset of stimulus blue and the value of zero otherwise, because onset of stimulus blue was unpredictable but occurrences of the following events were predictable. Thus, all occurrences of event onsets and event offsets following onset of stimulus blue were correctly predicted. These events were completely predictable, as in the previous 10 trials the reaction times and the executed acts had been alike (not shown). The activations of the dopamine membrane effects only weakly influenced the striatal membrane potentials (line 8), while a substantial increase of the corticostriatal weight associating stimulus blue with the act right (line 7, eq. 5) strongly activated the striatal neuron coding for act right (line 8, eqs. 3, 4, and 7). The predominant influence of the corticostriatal weight demonstrates that performance is mainly controlled by sensorimotor associations. The striatal firing rates were integrated over time by cortical neurons (line 9) that elicited the correct act right (bottom line).

The previous three figures show model performance of the standard model in one typical experiment. The outcome of each experiment was influenced by the randomly selected states of the striatal membrane potentials at the beginning of each trial. Therefore, model performance was tested in 1000 experiments. In addition, we produced seven model variants by setting for each model variant one crucial parameter of the standard model to zero. As for the standard model, we simulated 1000 experiments for each of these physiologically less plausible model variants.

During the exploration phase, the model variant without dopamine-like novelty responses to novel stimuli ($v = 0$, eq. 15) executed substantially less acts than did the other model variants (Table 4). The

number of acts during the exploration phase was also substantially reduced for the model variant without transient effects of the dopamine-like signal on the membranes of striatal neurons (h = 0, eq. 1).

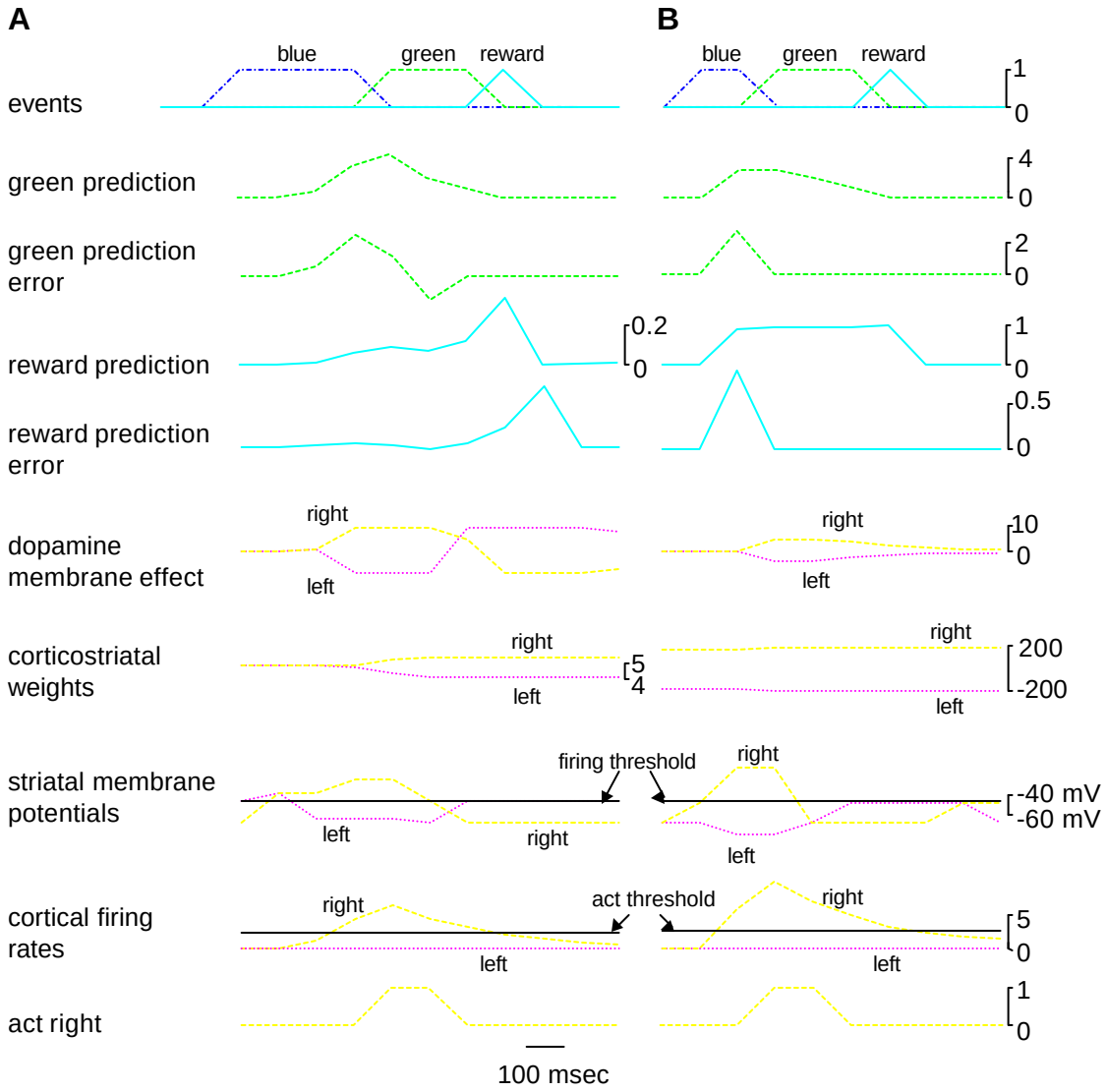

Fig. 9 Model performance in test phase. When presentation of stimulus blue (line 1) was responded to with the correct act right (bottom line), the stimulus green was presented, which was followed by the reward presentation (line 1). **(A)** Successful planning in first trial. The signal coding for prediction of stimulus green (line 2) was already slightly activated when the firing rate of the striatal neuron coding for the act right was increased (line 8). The green prediction error (line 3) first increased above zero and then decreased below zero, which reflects some uncertainty in the prediction of stimulus green. Since the green prediction was associated with the reward prediction, the reward prediction shows a first small activation (line 4). This signal shows a second higher peak when the partially predicted reward occurs. Therefore, the reward prediction was also uncertain (line 5). The first slight activation of the reward prediction error enhanced the firing rate of the striatal neuron coding for the act right (line 8), as the reward prediction error increased the corresponding dopamine membrane effect signal (line 6) and the corresponding corticostriatal weight (line 7). The cortical neurons integrated the striatal neural activity over time, and the act right was elicited (bottom line) when the cortical firing rate reached a threshold (line 9). **(B)** Successful sensorimotor association in trial 19. Since the onset of stimulus blue was unpredictable, this onset activated the prediction error signals for the stimulus green (line 3) and for the reward (line 5). These signals were otherwise on the value of zero, as the presentations of the stimulus green and of the reward were correctly predicted. The corticostriatal weights associating stimulus blue with the striatal membrane potentials (line 7) substantially increased the membrane potential of the striatal neuron coding for act right (line 8)), which triggered execution of the correct act right (bottom line).

Table 4. Number of Acts during Exploration Phase for Different Model Variants

standard	$n = 0$	$b = 0$	$k = 0$	$a = 0$	$e = 0$	$h = 0$	$f = 0$
26.3	13.6	26.3	25.8	22.7	27.1	13.8	29.5
± 0.3	± 0.1	± 0.2	± 0.1	± 0.1	± 0.2	± 0.1	± 0.1

Mean values ± standard errors were computed from 1000 experiments for each model variant.

The percentage of correct acts in the trials 1 to 19 of the test phase is shown for each model variant (Fig. 10). Significantly more than chance levels of 50% correct trials in trial 1 indicates planning capabilities, while performance improvements for successive trials indicates successful sensorimotor learning due to dopamine-dependent adaptations in corticostriatal weights (eq. 5). For the standard model (solid line with stars), in 79 % of the 1000 experiments trial 1 was correct, then performance worsened for three trials and reached 100 % correct trials with further training. For a model variant without dopamine-like novelty responses ($n = 0$, dashed line with crosses), performance in the first trial was significantly worse than that of the standard model (74 % correct, $c^2 = 7.2$, $p < 0.01$). For a model variant without adaptation of the dopamine-like signal ($b = 0$, eq. 15), the dopamine-like signal unconditionally responded to the reward and the response did not transfer to predictive stimuli according to eq. 14. Since responses in the dopamine-like signal reached the striatum too late to influence acts or associative weights, this model variant performed at chance levels in all trials (dash dotted line with triangles). If the TD model was used instead of the Extended TD model to simulate the dopamine-like signal ($k = 0$, dashed line with triangles), novel associative chains were not formed (eq. 13a-c). Since this model variant therefore did not profit from the two previous phases, performance was at chance levels in the first trial. Surprisingly, performance then decreased for some trials below chance levels and slowly improved with further training. If striatal activity could not influence the dopamine-like signal (salience $a = 0$, solid line with squares), there was no feedback between striatal activity and the Extended TD model during planning (Fig. 1B). Therefore, model performance was at chance levels for the first trial. In further trials, model performance improved progressively. For a model variant without adaptation of the corticostriatal weights ($e = 0$, dash dotted line with triangles), 79% of the first trials were correct, as planning was mostly controlled by dopamine membrane effects. Performance then decreased and kept slightly above chance levels. Since sensorimotor learning was prevented in this model variant, performance above chance levels reflects a minor influence of dopamine membrane effects. A model variant without dopamine membrane effects on the striatal neurons ($h = 0$, dash dotted line with triangles) performed significantly above chance levels in the first trial (60 % correct, $c^2 = 21$, $p < 0.01$), as dopamine-dependent synaptic adaptations led to some planning capabilities. This model variant then learned the task more rapidly than the standard model. If the dopamine-like signal did not influence the durations of up and down-states ($j = 0$, dotted line with circles), performance was significantly worse than that of the standard model (73 % correct, $c^2 = 9.9$, $p < 0.01$ for first trial).

We had expected that the model performance in the test phase would progressively increase in successive trials because sensorimotor learning would progressively dominate planning processes. Surprisingly, the learning curves of some model variants were *u*-shaped. We suspected that the presentation durations of stimuli green and red (300 *m*sec) could interfere with the durations of up and down-states (about 400 *m*sec). Therefore, the performance was tested when the stimuli green and red were presented for 800 *m*sec. Indeed, this prevented the performance of the model variant with $k = 0$ from decreasing below chance levels. However, the learning curves of other model variants still were *u*-shaped. In such incorrect trials, the dopamine-like signal was usually slightly increased although the striatal neuron coding for the incorrect act was active, which seemed to elicit the incorrect act. Therefore, we suspected that the stimulus blue became associated with the reward during the first correct test phase trials and that the activation of the dopamine-like signal elicited not only correct but also incorrect acts. This supposition was tested by decreasing the salience of stimulus blue in the Critic (from the standard value of

0.05 to 0.02). The learning curve of the model with decreased salience of stimulus blue was substantially better and less *u*-shaped than that computed with the standard parameters (improved trial 1: +2%; trial 2: +8%; trial 3: +14%). However, reaction times to stimulus blue increased (trial 1: +200 *m*sec; trial 2: +600 *m*sec; trial 3: +1000 *m*sec), presumably as the dopamine-like signal could not elicit incorrect movements. Taken together, *u*-shaped learning curves are a consequence of specific stimulus presentation durations and of the value for the salience of stimulus blue.

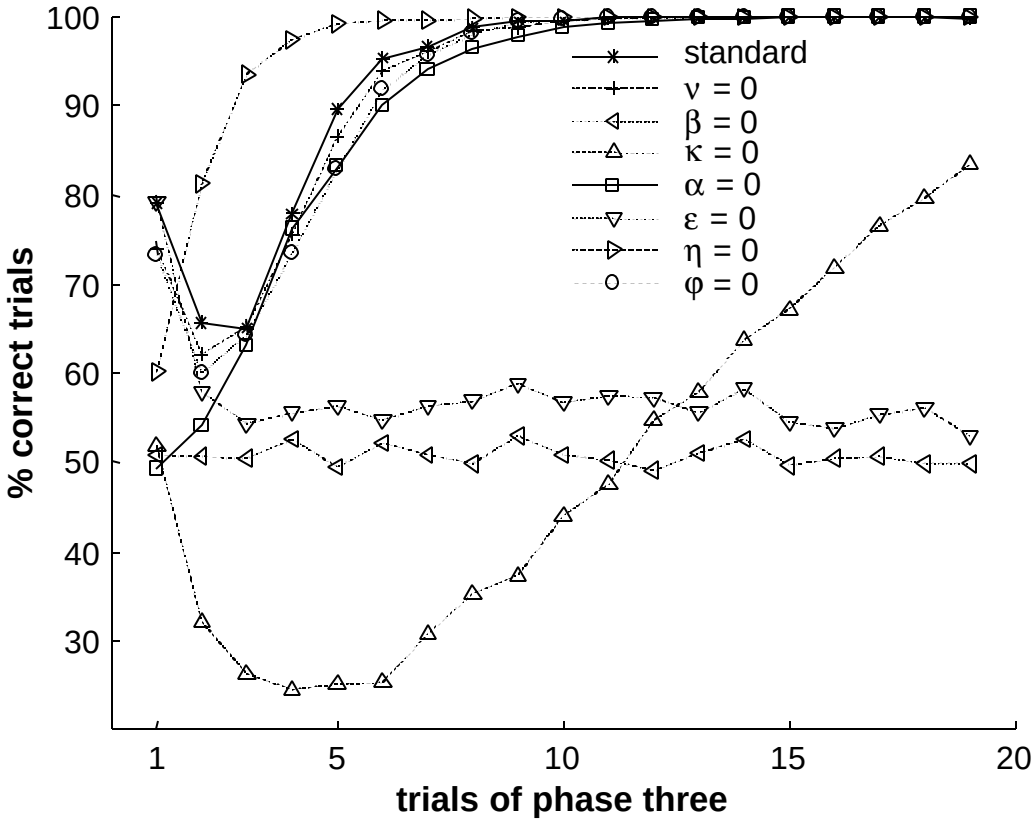

Fig. 10 Learning curves in test phase for different model variants. Each curve was computed from 1000 experiments (standard errors < 1.6 %). Trial 1 assesses planning and successive trials test the progress in sensorimotor learning. The standard model (solid line with stars) and the model variant without dopamine membrane effects (h = 0, dash dotted line with triangles) performed best. The model variant without dopamine novelty responses (n = 0, dashed line with crosses) performed in the first trial significantly worse than the standard model.

Average reaction times in test phase trials 1 to 19 were computed for each model variant (Fig. 11). The reaction time was defined as the sum of the time periods during which stimulus blue was present before act execution. Consequently, presentations of stimulus blue without acts contributed to the reaction time of the next act. Since reaction times for the performance controlled by sensorimotor associations were shorter than those controlled by planning, reaction times usually progressively decreased with repeated trial presentations. For the standard model, the reaction time in the first trial was 690 ± 10 *m*sec (mean ± standard error of the mean) and then progressively decreased to 200 *m*sec as sensorimotor learning progressively dominated planning. Reaction times were similar for the model variant without novelty responses (n = 0, dashed line with crosses). For the model variant without adaptation of the dopamine-like signal (b = 0, dash dotted line with triangles), the reaction times were increased to about 1800 *m*sec. If the Extended TD model was substituted with the TD model (k = 0, dashed line with triangles), the reaction times were increased to 2200 ± 60 *m*sec in the first trial and then progressively decreased. Similar reaction times resulted if the dopamine-like signal was not influenced by the striatal activity (salience a = 0, solid line with squares). Without sensorimotor learning (e = 0, dash dotted line

with triangles), reaction times kept at about 800 *m*sec. Without dopamine membrane effects ($h = 0$, dash dotted line with triangles), the reaction time in the first trial was substantially increased to 3000 ± 100 *m*sec and then decreased with further trials to 100 *m*sec. If the dopamine-like signal did not influence the durations of striatal up and down-states ($j = 0$, dotted line with circles), reaction times for the first trials were slightly longer than those of the standard model.

In further simulations with the standard model, the experimental paradigm was slightly changed to investigate the influence of the reaction time on planning capabilities. In these simulations, acts were prevented during the first 10 presentations of stimulus blue of the test phase. This was achieved by setting the two act signals $act_1(t)$ and $act_2(t)$ during these 10 blue presentations to zero (eq. 12). With this paradigm, performance in the first trial of the test phase significantly improved to 85.2 ± 0.01% correct trials (mean ± standard error, for 1000 experiments). The average reaction time to stimulus blue (670 *m*sec, not considering the first 10 blue presentations) was similar to that in the standard paradigm. Also performance in the next trials was better than in the control condition (73%, 85%, 94%, 99%, and 100% correct). This performance improvement was the result of improvements of the corticostriatal weights during the 10 presentations of stimulus blue. The corticostriatal weight that associated stimulus blue with the correct act right increased by an amount of +0.76 ± 0.01, whereas the weight that associated stimulus blue with the incorrect act left decreased by an amount of -0.85 ± 0.01 (mean ± standard error, for 1000 experiments). In this experimental paradigm, the model transfers information stored in Critic weights to improve sensorimotor corticostriatal weights without executing any act.

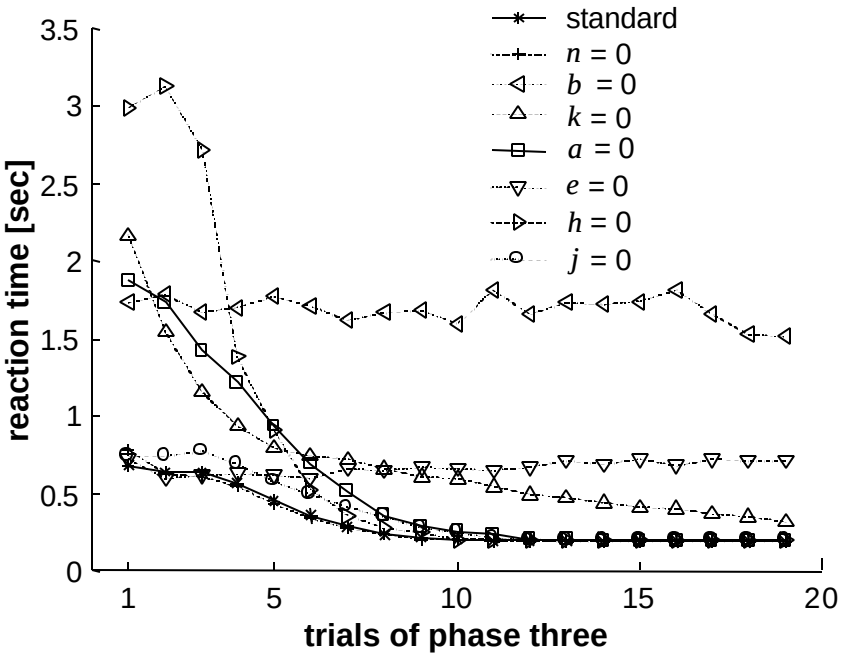

Fig. 11 Average reaction times in trials 1 to 19 of phase three for the different model variants. The reaction time for the act in the first trial, which assessed planning, was usually longer than the reaction times in successive trials, which assessed sensorimotor associations (line types and experimental data correspond with Fig. 10.).

DISCUSSION

This simulation study demonstrates that characteristics of dopamine neuron activity and striatal dopamine modulation are advantageous for exploration, sensorimotor learning, planning, and behaviorally silent improvement of associative strengths. Three types of membrane potential-dependent influences of dopamine on striatal medium spiny neurons are simulated: long-term adaptation of corticostriatal transmission, transient membrane effects, and influences on the durations of up and down-states. In our

simulations, all three types of dopamine effects significantly improve planning capabilities. Sensorimotor learning requires long-term adaptations of corticostriatal transmission and the transfer of the dopamine-like signal to the first reward-predictive stimulus. Dopamine-like novelty responses lead to increases in the number of exploratory acts, which significantly improves planning capabilities.

Dopamine Neuron Activity

Dopamine neuron activity was simulated with the Extended TD model (Suri and Schultz, submitted). Learning of reward predictions may involve plastic changes in projections from cortex and to striatal striosomes (Houk *et al.* 1995; Montague *et al.* 1996; Schultz *et al.* 1997; Suri & Schultz 1999; Brown *et al.*, in press), whereas stimulus predictions and novel associative chains may be evaluated in reciprocal projections between cortical areas (Suri and Schultz, submitted). As the Extended TD model was primarily motivated by studies of animal learning and neural activities, it does not closely correspond to adaptive processes in these structures. Therefore, substantial differences between the Extended TD model (Critic) and the proposed model for pathways from cortex via matrisomes to dopamine neurons (Actor) are a result of our modeling technique and may not correspond to substantial differences in their biological substrate.

After learning stimulus-reward associations, the reward prediction and the dopamine-like reward prediction error of the Extended TD model are equal to these signals of the TD model (compare Fig. 9B with Sutton and Barto, 1990). We model biphasic dopamine responses to novel stimuli by initializing certain weights with positive values (Suri and Schultz, 1999). With this assumption, the model reproduces biphasic dopamine novelty responses (Fig. 7A). Furthermore, these novelty responses decrease when the same neutral stimulus is presented repeatedly (Fig. 7B), which resembles habituation of dopamine novelty responses for repeated presentation of a neutral stimulus (see Introduction).

The reward prediction error of an internal model approach reproduces dopamine neuron activity (Suri and Schultz, submitted). This approach is simplified for the current study and termed the "Extended TD model." In contrast to the TD model, the dopamine-like reward prediction error signal of both models is influenced by the formation of novel associative chains. Whereas the previously proposed internal model approach computed many signals that were almost identical, the Extended TD model computes only one prediction error signal and one prediction signal for each sensory event. The prediction and prediction error signals of both models, including the dopamine-like signal, do not differ after learning of stimulus-reward associations (compare Fig. 9B with Suri and Schultz, submitted). Prediction signals of both models resemble anticipatory activity in cortex and striatum (Suri and Schultz, submitted). However, the time courses of the dopamine-like reward prediction errors differ when both models are tested for the formation of novel associative chains (unpublished).

Recently, Brown *et al.* (in press) modelled dopamine neuron activity with an alternative to TD models. Their simulations demonstrate that learning the time of reward occurrence may depend on metabotropic glutamate receptor-mediated calcium currents of striatal medium spiny neurons. Unfortunately, their model may fail to reproduce dopamine neuron activity for stimulus and reward offsets. In their simulations, stimulus and reward offsets were both at the end of the tasks (see appendix in Brown *et al.*), which is inconsistent with the experimental paradigm. In addition, the simulated dopamine-like signal can not be completely compared with dopamine neuron activity, since these offsets were outside the time intervals shown in their figures.

Sensorimotor Learning

Previous studies related sensorimotor learning to dopamine-dependent plasticity in projections from cortex to striatal matrisomes (Houk *et al.* 1995; Suri and Schultz, 1998, 1999). These models assume that corticostriatal transmission increases when "eligibility traces" of pre- and postsynaptic firing are active together with the dopamine-like signal. In the current model, corticostriatal transmission adapts without eligibility traces according to the product of dopamine activity, presynaptic activity, and postsynaptic membrane potential (eq. 5). Consistent with a result of a previous simulation study (Suri and Schultz

1998), successful sensorimotor learning requires an adaptive dopamine-like signal that can be computed with the Extended TD model (standard parameter values) or with the original TD model (parameter $k = 0$; Sutton and Barto, 1990).

Planning

The simulated planning task requires the formation of novel associative chains and the selection of the act that predicts the optimal outcome. During the reaction time, act preparation activity in striatal matrisomes alternates between both possible acts, driven by the fluctuating membrane potentials of these neurons. Indeed, activity of a subset of striatal neurons is related to act preparations (Schultz *et al.*, 1995). The reward-predictive values of both acts are evaluated by the Extended TD model and are reflected in the dopamine-like signal. Increases in the dopamine-like signal reinforce preparatory firing of the neuron that corresponds to the outcome with the best reward prediction. This reinforcement prolongs and increases reward-promising preparatory striatal activities. Thus, the dopamine-like signal guides reward-promising act preparations and elicits acts.

Our model simulations showed that successful planning requires that striatal activities serve as act preparation signals and influence dopamine neuron activity before the act is elicited. Successful planning requires the Extended TD model, as novel associative chains cannot be formed with the TD model. Therefore, previous models using a dopamine-like reinforcement signal for learning of sensorimotor associations are not able to solve the planning task (Houk *et al*. 1995; Montague *et al*., 1996; Schultz *et al*., 1997; Suri and Schultz, 1998, 1999). Since the model evaluates sequentially the reward predictions for both alternative acts, reaction times for planning are longer than reaction times for sensorimotor responding. As reaction times for planning are about 600 *m*sec (standard model, Fig. 11), the transient influence of dopamine on membrane potentials of striatal medium spiny neurons contributes substantially to the selection of the correct act in the planning phase. Also the persistent influence of dopamine on synaptic transmission contributes to planning capabilities. When the paradigm is adapted to prolong the reaction time in the planning task, the simulated long-term changes of cortico-striatal transmission lead to an improvement in planning capabilities (see last paragraph of Results). Thus, the model improves corticostriatal weights by simulating experience with the Extended TD model (Sutton and Barto, 1981; Sutton and Barto, 1998). Consistent with this simulation result, motor performance for human and animals can be improved by stimulating planning processes (Dickinson and Balleine, 1994; Decety, 1996).

Striatal Membrane Potential Fluctuations

Animal learning theorists suggest that animals vary their behavior in repetitive trials in order to find the optimal behavior (Skinner 1938; Hull 1952). According to these theories, animals repeat rewarded variations because a reward that follows a certain behavioral variation strengthens the processes that led to this variation. In the proposed model, each striatal neuron uses this strategy to optimize the dopamine-like reward prediction error. Spontaneous striatal membrane potential fluctuations lead to variations in the act selections and variations in the act preparation signals. Therefore, acts are varied in the exploration phase, and striatal act preparation signals are varied before they are executed in the planning phase. Dopamine membrane effects and dopamine long-term effects on corticostriatal transmission are adapted proportionally to both the dopamine-like signal and this variation in the firing rate. Therefore, processes that caused a behavioral variation are strengthened if the outcome is better than expected and weakened if the outcome is worse. A reinforcement learning algorithm with such random variations was proposed for tasks that require continuous output signals (Gullapalli, 1990).

Dopamine-Like Novelty Responses Increase Exploration and Improve Planning Capabilities

The presented simulations demonstrate that dopamine neuron activity could serve as an effective reinforcement signal for sensorimotor learning. However, it has been claimed that dopamine responses to

novel and physically salient stimuli are not consistent with this hypothesis (Pennartz 1995; Salamone *et al*. 1997). Nevertheless, our simulations demonstrate that dopamine-like biphasic novelty responses do not substantially impair sensorimotor learning, since the effects of the two phases cancel out and these responses extinguish during the exploration phase. Novelty responses increase the number of acts in the exploration phase (Table 4). This simulation result suggests that dopamine neuron responses stimulate exploration, which is consistent with experimental evidence (Stahle, 1992). Such stimulation of exploratory acts increases the number of task experiences in the exploration phase, which improves planning capabilities in the test phase (Fig. 10).

In reinforcement learning studies, an agent often moves between different places in a maze and learns to find the maximal amount of rewards. Performance of such algorithms depends on a trade-off between exploration and exploitation as well as on the exploration strategy (Fe'ldbaum 1965). In reinforcement learning studies, one popular technique is to be optimistic for places that have never been explored (Thrun 1992; Dayan and Sejnowski 1996; Sutton and Barto, 1998). Similar to the proposal to reproduce dopamine novelty responses with positive initial weights (Suri and Schultz, 1999), this has been achieved for TD models by attributing optimistic initial values to novel places (Sutton 1990; Dayan and Sejnowski 1996; Sutton and Barto 1998). When a novel place is visited that is preceded and followed by familiar places that are not associated with the reward, the reward prediction error increases above zero when the novel place is visited and decreases below the baseline level when the next place is visited. This biphasic reward prediction error in studies of TD learning resembles biphasic dopamine responses to novel stimuli. Furthermore, biphasic responses of dopamine neurons and of the reward prediction error signal progressively diminish when the same situation is presented repeatedly. The current study suggests that dopamine-like novelty responses may stimulate exploratory behaviors of animals as does the novelty bonus of reinforcement learning agents.

Striatal Dopamine Concentration

Tonic dopamine concentration in the striatum and nucleus accumbens is increased or decreased in some aversive situations that do not affect or decrease firing of dopamine neurons (Pennartz, 1995; Schultz, 1998). Tonic increases of dopamine concentration in these aversive situations could originate from release of dopamine via contacts of cortical terminals on dopaminergic axons, from dopamine release by striatal dopaminergic neurons (Betarbet *et al.*, 1996), from slow changes in the firing rates of midbrain dopamine neurons, or from dopamine novelty responses. Aversive stimuli can affect dopamine levels in different regions of the striatum or the nucleus accumbens to a different extent and even in the opposite direction (Besson and Louilot, 1995). Therefore, local changes in striatal dopamine concentration could indicate that dopamine is locally regulated to serve as a local reinforcement signal that is specific for the subtasks, such as orienting responses, processed in these local striatal areas. This hypothesis would explain why dopamine antagonists worsen performance in some aversive tasks such as active avoidance behaviors (Salamone, 1992). In reinforcement learning studies, such hierarchical architectures using subtask-specific reinforcement signals have been shown to be advantageous if the agent’s performance can be separated in subtasks that are sufficiently independent (Dayan and Hinton, 1993).

According to the proposed dopamine-dependent adaptation rules for long-term synaptic changes and short-term membrane effects, increases in dopamine levels in the striatum increase the ratio between high firing rates and low firing rates, which may be called an increase in the signal-to-noise ratio (see section “Striatal Dopamine Modulation *In Vivo*”). Indeed, dopamine application increases signal-to-noise ratio of striatal firing rates *in vivo* (Rolls, 1984; reviewed by Servan-Schreiber *et al.*, 1998a). In addition, pharmacological studies with dopamine agonists and antagonists indicate that dopamine decreases reaction times for behavioral responses and is involved in behavioral activation (Salamone *et al.*, 1997; Robbins *et al.*, 1998). In the proposed model, manipulations of dopamine concentration in the matrisomes have a complex influence on the number of acts in the exploration phase, since the simulated dopamine effects on elevated and on hyperpolarized membrane potentials are antagonistic and therefore sensitive to

model parameters values (unpublished results). Moreover, sustained dopamine concentration increases do not serve as a reinforcement signal, since the direction of dopamine effects depends on the fluctuating membrane potential. The most reliable influence of sustained dopamine increases on the model performance is probably a focussing effect in some tasks due to an increase in the signal-to-noise ratio of striatal firing rates (Servan-Schreiber *et al.*, 1998a,b).

Cortex

In hippocampal slices, pulsative application of dopamine D1 agonists increases the number of spontaneous bursts of CA1 pyramidal cells if the agonist application is contingent on spontaneous bursts but not if it is applied noncontingently (Stein *et al.*, 1993; Stein, 1997). This L-type calcium current mechanism was seen as a cellular substrate of operant conditioning and closely resembles the simulated effects of dopamine on striatal medium spiny neurons. Dopamine membrane effects seem to stabilize the activity patterns of cortical pyramidal neurons by suppressing weak and enhancing strong presynaptic activation (Durstewitz *et al.*, 1999a,b). Thus, dopamine enhances cortical working memory activity (Sawaguchi and Goldman-Rakic, 1991; Durstewitz *et al.*, 1999a,b). These mechanisms could contribute to dopamine-induced signal-to-noise ratio enhancement (reviewed by Servan-Schreiber *et al.*, 1998a). Since dopamine effects in the cortex resemble those in the striatum, cortical dopamine modulation may have similar functions as those simulated in the current study for the striatum.

Saccadic Eye Movements

Planning is a basic concept of modern control algorithms used in reinforcement learning and in engineering sciences (Garcia et al. 1989; Sutton and Barto, 1998). Planning may not only be involved in "cognitive" tasks but also in "motor" tasks that cannot be solved with sensorimotor learning (Wolpert et a., 1995). Pure sensorimotor learning fails to explain behavior in tasks that require novel movements to reach a goal. For example, sensorimotor learning cannot explain learning of saccadic eye movements to fixate a behaviorally important object if the muscle commands vary due to varying start positions of the gaze. The proposed model suggests how such goal-directed movements could be generated. The proposed Extended TD model (Critic) learns prediction signals that precede sensory consequences of planned acts (Fig. 7B). Neural recording studies during saccadic eye movement tasks report direct evidence for this postulated phenomenon. Sensory and sensorimotor neural activity in frontal eye fields (Goldberg & Bruce 1990; Umeno & Goldberg 1997), superior colliculus (Walker *et al.* 1995), and lateral intraparietal area (Duhamel *et al.* 1992), striate and extrastriate cortex (Nakamura and Colby, 1999) anticipates the retinal consequences of saccades about 100 *m*sec before these saccades are elicited. The saccade preparation-dependent transformation of these receptive fields has been proposed to occur in frontal eye fields (Goldberg and Bruce, 1990) or in lateral intraparietal area (Dominey and Arbib, 1992). Instead of assuming a nonadaptive task-specific mechanism for this transformation (Dominey and Arbib, 1992; Dominey, Arbib, and Joseph, 1995), the proposed model suggests that neurons in cortical areas learn to anticipate sensory consequences of intended saccades. Moreover, we suggest that this anticipatory activity can select a reward-promising target via increases in dopamine neuron activity. This hypothesis would explain the surprising fact that these largely retinotopically organized areas can select context-dependent saccade targets for arbitrary start and target positions. This hypothesis is also consistent with the view that dopamine attributes salience to novel and reward-related stimuli and thereby triggers the animal's visual and internal attention to such targets (Pennartz 1995; Salamone *et al*. 1997; Redgrave *et al.* 1999).

Acknowledgement

This study was supported by a fellowship of the Swiss National Science Foundation (R.S.) and by a Program Project grant from the Human Brain Project (M.A.).

REFERENCES

Alexander GE, DeLong MR, Strick PL. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 1986;9:357-81

Alexander GE, Crutcher MD. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 1990 Jul;13(7):266-71

Albin RL, Young AB, Penney JB. The functional anatomy of basal ganglia disorders. Trends Neurosci 1989 Oct;12(10):366-75

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

Arbib, M.A. (1972). *The metaphorical brain*. New York: Wiley- Interscience.

Balleine BW, Dickinson A Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 1998 Apr-May;37(4-5):407-19

Besson C, Louilot A Asymmetrical involvement of mesolimbic dopaminergic neurons in affective perception. Neuroscience 1995 Oct;68(4):963-8

Betarbet R, Turner R, Chockkan V, DeLong MR, Allers KA, Walters J, Levey AI, Greenamyre JT. Dopaminergic neurons intrinsic to the primate striatum. J Neurosci 1997 Sep 1;17(17):6761-8

Berns G. S. and Sejnowski *T*.J. (1996). How the basal ganglia make decisions. In: Neurobiology of decision-making (Damasio A. R. *et al.*, eds.), Springer Verlag Berlin Heidelberg.

Blank T, Nijholt I, Teichert U, Kugler H, Behrsing H, Fienberg A, Greengard P, Spiess J The phosphoprotein DARPP-32 mediates cAMP-dependent potentiation of striatal N-methyl-D-aspartate responses. Proc Natl Acad Sci U S A 1997 Dec 23;94(26):14859-64

Brown, J., Bullock, D, and Grossberg, S. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. J Neuroscience, in press.

Calabresi P, Mercuri N, Stanzione P, Stefani A, Bernardi G. Intracellular studies on the dopamine-induced firing inhibition of neostriatal neurons *in vitro*: evidence for D1 receptor involvement. Neuroscience 1987 ; 20 (3): 757-71.

Calabresi P., Pisani A., Mercuri N. B. and 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, Pisani A, Centonze D, Bernardi G Synaptic plasticity and physiological interactions between dopamine and glutamate in the striatum. Neurosci Biobehav Rev 1997a Jul;21(4):519-23

Calabresi P, Saiardi A, Pisani A, Baik JH, Centonze D, Mercuri NB, Bernardi G, Borrelli E. Abnormal synaptic plasticity in the striatum of mice lacking dopamine D2 receptors. J Neurosci 1997b Jun 15;17(12):4536-44

Cepeda C, Buchwald NA, Levine MS. Neuromodulatory actions of dopamine in the neostriatum are dependent upon the excitatory amino acid receptor subtypes activated.Proc Natl Acad Sci U S A 1993; 90 (20): 9576-80

Cepeda C, Chandler SH, Shumate LW, Levine MS Persistent Na+ conductance in medium-sized neostriatal neurons: characterization using infrared videomicroscopy and whole cell patch-clamp recordings. J Neurophysiol 1995; 74 (3): 1343-8.

Cepeda C, Colwell CS, Itri JN, Chandler SH, Levine MS. Dopaminergic modulation of NMDA-induced whole cell currents in neostriatal neurons in slices: contribution of calcium conductances. J Neurophysiol 1998;.79.(1):.82-94.

Cepeda C, Levine MS. Dopamine and N-methyl-D-aspartate receptorinteractions in the neostriatum. Dev Neurosci 1998; 20 (1): 1-18.

Cherubini E, Herrling PL, Lanfumey L, Stanzione P Excitatory amino acids in synaptic excitation of rat striatal neurones *in vitro*. J Physiol (Lond) 1988; 400: 677-90.

Cohen JD, Servan-Schreiber D Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. Psychol Rev 1992 Jan;99(1):45-77

Cooper DC, White FJ. Cocaine alters the bistable membrane potential in the nucleus accumbens: and *in vivo* intracellular study in the mice. Soc Neurosci. Abstr. 1998; vol. 24.

Craik, K. (1943). *The nature of explanation*. Great Britain. Cambridge University Press.

Dayan P. and Hinton G.E. 1993. Feudal reinforcement learning. In: Advances in Neural Information Processing Systems 5 (S.J. Hanson and J.D. Cowan and C.L. Giles eds.), San Mateo, CA, Morgan Kaufmann: 271-278.

Dayan P and Sejnowski TJ (1996). Exploration bonuses and dual control. Machine Learning, 25, 5-22.

Decety J. The neurophysiological basis of motor imagery. Behav Brain Res. 1996 May;77(1-2):45-52.

Dickinson, A. (1980). Contemporary animal learning theory. Cambridge University press.

Dickinson, A. and Balleine B (1994). Motivational control of goal-directed action. Animal Learning and Behavior 22 (1): 1-18.

Dominey PF, Arbib MA. A cortico-subcortical model for generation of spatially accurate sequential saccades. Cereb Cortex 1992 Mar-Apr;2(2):153-75

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.

Durstewitz D, Kelc M, Gunturkun O. A neurocomputational theory of the dopaminergic modulation of working memory functions. J Neurosci 1999a Apr 1;19(7):2807-22

Durstewitz, D., Seamans, J. K., and Sejnowski, *T*. J. (1999b) Dopaminergic modulation of activity states in the prefrontal cortex. Soc. Neurosci. Abstr., 25: 1216.

Fel'dbaum, A. A. (1965). Optimal Control Systems. New York, NY: Academic Press.

Fellous JM, Linster C Computational models of neuromodulation. Neural Comput 1998 May 15;10(4):771-805

Fienberg AA, Hiroi N, Mermelstein PG, Song W, Snyder GL, Nishi A, Cheramy A, O'Callaghan JP, Miller DB, Cole DG, Corbett R, Haile CN, Cooper DC, Onn SP, Grace AA, Ouimet CC, White FJ, Hyman SE, Surmeier DJ, Girault J, Nestler EJ, Greengard P. DARPP-32: regulator of the efficacy of dopaminergic neurotransmission. Science 1998 Aug 7;281(5378):838-42 .

Gallistel CR (1990) The organization of learning. A Bradford Book, MIT press, Massachusetts.

Garcia, C.E., Prett, D.M., & Morari, M. (1989). Model Predictive Control: Theory and Practice-a survey. *Automatica, 25*, 335-348.

Garcia-Munoz M, Young SJ, Groves PM. Presynaptic long-term changes in excitability of the corticostriatal pathway. Neuroreport 1992 Apr;3(4):357-60

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.

Gonon F. Prolonged and extrasynaptic excitatory action of dopamine mediated by D1 receptors in the rat striatum *in vivo*.J Neurosci 1997; 17 (15): 5972-8.

Graybiel AM. Neurotransmitters and neuromodulators in the basal ganglia. Trends Neurosci 1990 Jul;13(7):244-54

Gullapalli V. A stochastic reinforcement algorithm for learning real valued functions. Neural Networks, 3: 671-692 (1990)

Hernandez-Lopez S, Bargas J, Reyes A, Galarraga E. Dopamine modulates the afterhyperpolarization in neostriatal neurones. Neuroreport 1996; 7 (2): 454-6.

Hernandez-Lopez S, Bargas J, Surmeier DJ, Reyes A, Galarraga E. D1 receptor activation enhances evoked discharge in neostriatal medium spiny neurons by modulating an L-type Ca2+ conductance. J Neurosci 1997; 17 (9): 3334-42

Houk JC, Adams JL, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of Information Processing in the Basal Ganglia (Houk JC, Davis JL, Beiser DG eds), Massachusetts Institute of Technology: 215-232.

Hull, C. L. (1952). A behavioral system: An introduction to behavior theory concerning the individual organism. New Haven: Yale University Press.

Jaeger D, Kita H, Wilson CJ Surround inhibition among projection neurons is weak or nonexistent in the rat neostriatum. J Neurophysiol 1994 Nov;72(5):2555-8.

Katayama Y, Tsubokawa T, Moriyasu N Slow rhythmic activity of caudate neurons in the cat: statistical analysis of caudate neuronal spike trains. Exp Neurol 1980; 68(2):310-21

Kearns M and Singh S. "Bias-variance" error bounds for temporal difference updates. Submitted.

Kita H. Glutamatergic and GABAergic postsynaptic responses of striatal spiny neurons to intrastriatal and cortical stimulation recorded in slice preparations. Neuroscience 1996; 70 (4): 925-40.

Kitai ST, Surmeier DJ. Cholinergic and dopaminergic modulation of potassiumconductances in neostriatal neurons. Adv Neurol 1993; 60: 40-52.

Lange KW, Robbins TW, Marsden CD, James M, Owen AM, Paul GM L-dopa withdrawal in Parkinson's disease selectively impairs cognitive performance in tests sensitive to frontal lobe dysfunction. Psychopharmacology (Berl) 1992;107(2-3):394-404

Levine MS, Li Z, Cepeda C, Cromwell HC, Altemus KL. Neuromodulatory actions of dopamine on synaptically-evoked neostriatal responses in slices. Synapse 1996; 24 (1): 65-78.

MacCorquodale K. and Meehl P.E. "Section 2: Edward C. Tolman". pp. 177-266. In: Modern Lerning Theory (Estes W.K. ed.), Appleton-Century-Crofts, New York, 1954.

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

Montague PR, Dayan P, Sejnowski TJ A framework for mesencephalic dopamine systems based on predictive Hebbian learning J Neurosci 1996 Mar 1;16(5):1936-47.

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

Nicola SM, Malenka RC Modulation of synaptic transmission by dopamine and norepinephrine in ventral but not dorsal striatum. J Neurophysiol 1998 Apr;79(4):1768-76

Nisenbaum ES, Berger TW, Grace AA. Depression of glutamatergic and GABAergic synaptic responses in striatal spiny neurons by stimulation of presynaptic GABAB receptors. Synapse 1993;14 (3): 221-42.

Nisenbaum ES, Xu ZC, Wilson CJ Contribution of a slowly inactivating potassium current to the transition to firing of neostriatal spiny projection neurons. J Neurophysiol 1994 Mar;71(3):1174-89

Pacheco-Cano MT, Bargas J, Hernandez-Lopez S, Tapia D, Galarraga E. Inhibitory action of dopamine involves a subthreshold Cs(+)-sensitive conductance in neostriatal neurons. Exp Brain Res 1996; 110 (2): 205-11

Pennartz CM The ascending neuromodulatory systems in learning by reinforcement: comparing computational conjectures with experimental findings. Brain Res Rev 1995 Nov;21(3):219-45

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic books.

Pineda JC, Galarraga E, Bargas J, Cristancho M, Aceves J Charybdotoxin and apamin sensitivity of the calcium-dependent repolarization and the afterhyperpolarization in neostriatal neurons. J Neurophysiol 1992 Jul;68(1):287-94

Redgrave P, Prescott TJ, Gurney K. Is the short-latency dopamine response too short to signal reward error? Trends Neurosci 1999 Apr;22(4):146-51

Robbins TW, Granon S, Muir JL, Durantou F, Harrison A, Everitt BJ. Neural systems underlying arousal and attention. Implications for drug abuse. Ann N Y Acad Sci 1998 Jun 21; 846:222-37.

Rolls ET, Thorpe SJ, Boytim M, Szabo I, Perrett DI. Responses of striatal neurons in the behaving monkey. 3. Effects of iontophoretically applied dopamine on normal responsiveness. Neuroscience 1984 Aug;12(4):1201-12

Rutherford A, Garcia-Munoz M, Arbuthnott GW An afterhyperpolarization recorded in striatal cells 'in vitro': effect of dopamine administration. Exp Brain Res 1988;71(2):399-405

Salamone JD Complex motor and sensorimotor functions of striatal and accumbens dopamine: involvement in instrumental behavior processes. Psychopharmacology (Berl) 1992;107(2-3):160-74

Salamone JD, Cousins MS, Snyder BJ Behavioral functions of nucleus accumbens dopamine: empirical and conceptual problems with the anhedonia hypothesis. Neurosci Biobehav Rev 1997 May;21(3):341-59

Sawaguchi *T*, Goldman-Rakic PS. D1 dopamine receptors in prefrontal cortex: involvement in working memory. Science 1991 Feb 22;251(4996):947-50

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

Schultz, W., Apicella, P., Romo, R. and Scarnati, E.: Context-dependent activity in primate striatum reflecting past and future behavioral events. In: Models of Information processing in the basal ganglia (Eds. J.C. Houk, J.L. Davis and D.G. Beiser. MIT Press, Cambridge, MA, USA, pp. 11-28, 1995.

Schultz W, Dayan P, Montague PR A neural substrate of prediction and reward. Science 1997 275(5306):1593-9

Servan-Schreiber D, Bruno RM, Carter CS, Cohen JD. Dopamine and the mechanisms of cognition: Part I. A neural network model predicting dopamine effects on selective attention. Biol Psychiatry 1998 May 15; 43(10): 713-22

Servan-Schreiber D, Carter CS, Bruno RM, Cohen JD Dopamine and the mechanisms of cognition: Part II. D-amphetamine effects in human subjects performing a selective attention task. Biol Psychiatry 1998 May 15; 43(10): 723-9

Skinner, B. F. (1938). The behavior of organisms: An experimental analysis. New York: D. Appleton Century.

Smith DA and Bolam JP (1990) The neural network of the basal ganglia as revealed by the study of synaptic connections of identified neurons. TINS 13 (7):259-265.

Stahle L Do autoreceptors mediate dopamine agonist--induced yawning and suppression of exploration? A critical review. Psychopharmacology (Berl) 1992;106(1):1-13

Stern EA, Kincaid AE, Wilson CJ. Spontaneous subthreshold membrane potential fluctuations and action potential variability of rat corticostriatal and striatal neurons *in vivo.* J Neurophysiol 1997; 77(4) :1697-715.

Suri, R.E., and Arbib, M.A. Modeling sensorimotor learning in striatal projection neurons. Soc. Neurosci. Abstr. vol 24: p. 174. 1998.

Suri RE, Marmol JS, and Arbib MA. A documented online model of striatal dopamine modulation. 1999. In preparation at http://latte.usc.edu/~bmw.

Suri RE, Schultz W Learning of sequential movements by neural network model with dopamine-like reinforcement signal Exp Brain Res 1998 Aug;121(3):350-4

Suri RE, Schultz W A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience 1999;91(3):871-90

Suri and Schultz, submitted. Internal model reproduces anticipatory neural activity. Available at http://www.cnl.salk.edu/~suri

Surmeier DJ, Bargas J, Hemmings HC Jr, Nairn AC, Greengard P Modulation of calcium currents by a D1 dopaminergic protein kinase/phosphatase cascade in rat neostriatal neurons. Neuron 1995 Feb;14(2):385-97

Surmeier DJ, Kitai ST. D1 and D2 dopamine receptor modulation of sodium and potassium currents in rat neostriatal neurons. Prog Brain Res 1993; 99: 309-24.

Surmeier DJ, Eberwine J, Wilson CJ, Cao Y, Stefani A, Kitai ST. Dopamine receptor subtypes colocalize in rat striatonigral neurons.Proc Natl Acad Sci U S A 1992; 89 (21): 10178-82.

Sutton RS, Barto AG (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. Time derivative models of Pavlovian reinforcement. In: Learning and computational neuroscience: Foundations of adaptive networks (eds Gabriel M. and Moore. J.) MIT Press, Cambridge: 539-602, 1990.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Bradford Books, Cambridge, MA. (Available: http://envy.cs.umass.edu /~rich/book/the-book.html)

Sutton RS Pinette B The learning of world models by connectionist networks. Proceedings of the seventh annual conference of the cognitive science society, Lawrence Erlbaum, Irvine, California, August 1985: 54-64

Taylor AE, Saint-Cyr JA The neuropsychology of Parkinson's disease. Brain Cogn 1995 Aug;28(3):281-96

Thistlethwaite D., A critical review of latent learning and related experiments. Psychological Bulletin 48 (2): 97-129, 1951.

Thrun, S. B. (1992). The role of exploration in learning control. In D.A. White and D.A. Sofge, eds, Handbook of Intelligent Control: Neural, fuzzy and adaptive approaches. New York, NY: Van Nostrand Reinhold.

Umemiya M, Raymond LA Dopaminergic modulation of excitatory postsynaptic currents in rat neostriatal neurons. J Neurophysiol 1997 Sep;78(3):1248-55

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.

Wallesch CW, Karnath HO, Papagno C, Zimmermann P, Deuschl G, Lucking CH. Parkinson's disease patient's behaviour in a covered maze learning task. Neuropsychologia 1990;28(8):839-49

Wickens JR, Begg AJ, Arbuthnott GW Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex *in vitro*. Neuroscience 1996 Jan;70(1):1-5

Wickens JR, Wilson CJ. Regulation of action-potential firing in spiny neurons of the rat neostriatum *in vivo*.J Neurophysiol 1998; 79 (5): 2358-64

Williams GV, Millar J. Concentration-dependent actions of stimulated dopamine release on neuronal activity in rat striatum. Neuroscience 1990; 39 (1):1-16.

Wilson CJ The generation of natural firing patterns in neostriatal neurons. Prog Brain Res 1993;99:277-97.

Wilson CJ. Dendritic morphology, inward rectification, and the functional properties of neostriatal neurons. In: Single Neuron Computation (McKenna *T*, Davis J, and Zornetzer SF eds.) Academic Press, San Diego. 1992

Wilson CJ, Groves PM Spontaneous firing patterns of identified spiny neurons in the rat neostriatum. Brain Res 1981 Sep 7;220(1):67-80

Wilson CJ, Kawaguchi Y The origins of two-state spontaneous membrane potentialfluctuations of neostriatal spiny neurons. J Neurosci 1996; 16 (7): 2397-410.

Wolpert DM, Ghahramani Z, Jordan MI An internal model for sensorimotor integration. Science 1995 Sep 29;269(5232):1880-2.

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.