

Using TD Learning to Simulate Working Memory Performance in a Model of the Prefrontal Cortex and Basal Ganglia

Ahmed A. Moustafa^{a,*} Anthony S. Maida^{b,**}

^a*Institute of Cognitive Science
University of Louisiana at Lafayette
Lafayette, LA 70504, USA*

^b*Center for Advanced Computer Studies and
Institute of Cognitive Science
University of Louisiana at Lafayette
Lafayette, LA 70504, USA*

Abstract

Delayed-response tasks (DRTs) have been used to assess working memory (WM) processes in human and non-human animals. Experiments have shown that the basal ganglia (BG) and dorsolateral prefrontal cortex (DLPFC) subserve DRT performance. Here, we report the results of simulation studies of a systems-level model of DRT performance. The model was trained using the temporal difference (TD) algorithm and uses an actor-critic architecture. The matrixomes of the BG represent the actor and the striosomes represent the critic. Unlike existing models, we hypothesize that the BG subserve the selection of both motor- and cognitive-related information in these tasks. We also assume that the learning of both processes is based on reward presentation. A novel feature of the model is the incorporation of delay-active neurons in the matrixomes, in addition to DLPFC. Another novel feature of the model is the subdivision of the matrixomal neurons into segregated winner-take-all (WTA) networks consisting of delay- versus transiently-active units.

Our simulation model proposes a new neural mechanism to account for the occurrence of perseverative responses in WM tasks in striatal-, as well as in prefrontal damaged subjects. Simulation results also show that the model both accounts for the phenomenon of time shifting of dopamine phasic signals and the effects of partial reinforcement and reward magnitude on WM performance at both behavioral and neural levels. Our simulation results also found that the TD algorithm can subserve learning in delayed-reversal tasks.

Key words: basal ganglia; prefrontal cortex; dopamine phasic signal; actor-critic architecture; artificial lesioning.

1 Introduction

Computational cognitive neuroscience seeks to understand the neural mechanisms underlying the various cognitive functions. The present paper concerns the cognitive neuroscience of working memory (WM) and reward-based learning as realized in the basal ganglia (BG) and prefrontal cortex (PFC). It is well known that the BG subserve motor processes (Jog et al., 1999; Wickens, 1997). For example, patients with disorders of the BG such as Parkinson’s disease show prominent motor deficits, including impairment in initiating motor actions. Accordingly, many researcher’s have suggested that the BG subserve action selection, that is, selection of which motor response to perform under which conditions (Frank, 2005; Gurney et al., 2001; Prescott et al., 2003). It has also been suggested that the BG serve an analogous function in the cognitive domain (Berns and Sejnowski, 1996; Redgrave et al., 1999), for instance, with respect to the control of WM. Prescott et al. (2003, p. 150), in particular, noted that “the selection and maintenance of specific WM items can be viewed as an extension of action selection by the basal ganglia to the domain of cognition (selecting from the range of potential cognitive representations those which are to be sustained as working memory).”

We label this idea the *uniform selection hypothesis* for cognitive and motor actions and study it in the context of a WM task known as the delayed response task (DRT). In particular, we present a new computational model of that task. The present computational study has two related aims: 1) determining whether the temporal difference (TD) learning algorithm can simulate the acquisition of DRT performance; and, 2) understanding the functional role of the PFC and BG in working memory performance.

A delayed-response task (DRT) requires a human or nonhuman subject to utilize working memory (WM). Typically in a DRT, a subject remembers a presented cue stimulus over a short delay interval and then selects a motor response at the end of the interval based on which cue was presented. Animals are trained to perform such tasks by appropriate reinforcement. We have selected the DRT for our simulation studies because this task has been used to assess the functional role of BG and dorsolateral prefrontal cortex (DLPFC) in WM performance in previous research (Collins et al., 2000; Dunnett et al., 1999; Goldman-Rakic, 1995; Kawagoe et al., 1998; Levitt et al., 2002). Further, there is ample evidence that the BG subserve WM processes in both human and non-human animals (Battig et al., 1960; Collins et al., 2000; Divac

* Current address: Department of Psychology and Program in Neuroscience, University of Arizona, Tuscon, AZ, 85721.

**Corresponding author.

Email address: maida@cacs.louisiana.edu (Anthony S. Maida).

URL: <http://www.cacs.louisiana.edu/~maida> (Anthony S. Maida).

et al., 1967; Gabrieli, 1995; Kawagoe et al., 1998; Levy et al., 1997; Lewis et al., 2003; Owen, 2004; Prescott et al., 2003).

Existing computational models (reviewed in Section 1.5.) simulate either motor or cognitive process of the BG. The new model presented here simulates both the motoric and cognitive processes of BG function using the same mechanism, in a manner consistent with the uniform selection hypothesis. Further, the model assumes that learning to perform selection of both motoric and cognitive processes depends on reward presentation in conjunction with temporal difference (TD) learning (Barto, 1995; Houk et al., 1995). The model accounts for a wide variety of WM data, including neuropsychological and reward-based studies.

Temporal difference (TD) learning is a form of reinforcement learning that was first used by Witten (1977) and Barto et al. (1983). Modern reinforcement learning theory is a framework that describes the interaction between a learning agent and its environment in terms of states, actions, and rewards. It derives from the subfields of the psychology of animal learning beginning with Thorndike (1911) and the theory of optimal control in engineering beginning with dynamic programming (Bellman, 1957). More discussion of the history can be found in Sutton and Barto (1998).

The defining feature of reinforcement learning is that an agent, when in a given state, selects an action with some probability. The mapping from states to actions is called a policy and the agent adjusts its policy to maximize its expected reward (both immediate and future) on the basis of immediate rewards and punishments provided by the environment. Complications arise when an agent must select a sequence of actions in order to receive a reward at the end of a behavioral episode. Since the reward or punishment is received at the end of the episode, how should credit or blame be assigned to the individual actions taken during the episode? This is known as the temporal credit assignment problem, first described by Minsky (1961). TD learning was designed to solve this problem. In part, we use TD learning in this paper because the credit assignment problem arises in the DRT. Specifically, the agent must learn to gate a cue into working memory long before it receives a reward for that action.

The early implementations of TD learning also used an actor-critic architecture. Such an architecture has a memory structure representing the policy for choosing an action and a separate memory structure for evaluating the consequence of an action in an environment. The former is called the actor and the latter is termed the critic. Barto (1995) and Houk et al. (1995) carefully consider the problem of mapping an actor-critic architecture onto the BG. Accordingly, our implementation of the TD algorithm uses an actor-critic framework that follows their approach.

The theory of TD learning is based on the idea of reward prediction failures which are detected by the critic. These failures control modification of the action-selection policy represented in the actor. If there are no prediction failures, the policy is not modified. If an action is taken which immediately leads to an unanticipated reward, then this represents a prediction failure by the critic. In this case, the policy is modified to increase the probability of selecting this action in future occurrences of the same situation because the situation-action association is more predictive of reward than anticipated. Further, the critic modifies its own internals to predict a reward in this situation. In this sense, it is an adaptive critic. Although the modified critic may successfully predict rewards in this particular situation, it will likely give inaccurate predictions in situations leading to this situation. These new failures will cause the critic to modify the actor's policy for earlier actions. In this way, it addresses the credit assignment problem by ratcheting backward in time to adjust the values of actions earlier and earlier within the behavioral episode. The reward characteristics of the critic have a psychological interpretation. The reward predictors earlier in the episode acquire the role of secondary reinforcers and the TD algorithm can be viewed as a generalization of the Rescorla and Wagner (1972) learning rule in which learning is associated with unexpected rewards. The relationship to the Rescorla-Wagner model has been outlined in Sutton and Barto (1987, 1990) and Sutton (1988).

There are also previous lines of research using reinforcement learning for psychological modeling, ranging from the neural level (Foster et al., 2000) to the cognitive level. Some of the models at the cognitive level offer more general approaches. For instance, Sun et al. (2001) and Sun et al. (2005) have used reinforcement learning methods to provide an integrated theory for the acquisition of implicit and explicit knowledge. There are also general cognitive architectures that offer alternative explanations of the functional role of the BG. For instance, Anderson et al. (2004) offers a unified architecture for cognition. In this theory, it is hypothesized that the BG implement a production system and they do not commit themselves to a theory of reinforcement learning. Our model complements this work because it attempts to form a bridge between the neural and cognitive levels. Specifically, it makes reference to systems level neural structures and a dopamine phasic signal, yet it tries to model a WM task with an elementary cognitive component.

Before discussing our own model, we briefly review more specific research upon which the model is based.

1.1 *The Basal Ganglia and Connecting Structures*

This section briefly describes the neuronal connectivity within the BG and its relation to other parts of the brain. This will make it easier to understand our model, which is presented in Section 2. The BG are subcortical structures of the forebrain and midbrain (Prescott et al., 2003; Wilson, 2004). The striatum is the main input structure of the BG. The principal neurons of the striatum, the spiny neurons, receive convergent projections from much of the cortex, including PFC (Eblen and Grabiell, 1995). This is known as the corticostriatal pathway. Spiny neurons receive projections from over 5,000 cortical neurons (Houk, 1995; Wilson, 2004).

Spiny neurons make inhibitory contacts on neurons of the globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr). These in turn make inhibitory contacts on neurons in the thalamus, which sends excitatory projections to PFC. Spiny neurons are normally silent but exhibit brief bursts. Combined with the cascade of two inhibitory contacts, this creates a net excitatory effect that disinhibits the thalamus. The main targets of this pathway, known as the ascending pathway, are motor and cognitive cortical areas (Middleton and Strick, 2000, 2002). Middleton and Strick (2002) found that the cognitive and motor pathways of the GPi/SNr-thalamus-cortex are segregated and that DLPFC is a target of BG output. Further, DLPFC and the caudate nucleus are connected through a closed pathway (Owen, 2004), in contrast to an open pathway projecting to the motor cortex.

The striatum is subdivided into many segregated compartments consisting of spiny neurons. These compartments fall into two categories: striosomes and matrisomes. Approximately 95 percent of the spiny neurons fall within the matrisome compartments (Houk et al., 1995; Wilson, 2004). In addition to distinct neurochemical make-up, both the striosomes and matrisomes have different efferent targets. The striosomes reciprocally connect to the substantia nigra pars compacta (SNc), whereas the matrisomes project to the GPi and SNr — the output system of the BG (Wilson, 2004). It has been hypothesized that the striosomes and matrisomes are involved in different functions (Brown et al., 2002; Houk et al., 1995), with the former subserving reward-based learning and the latter subserving motor learning. This hypothesis is the basis for the mapping of the actor-critic architecture onto the BG (Barto, 1995; Houk et al., 1995). It is also known that microstimulation of different striatal neurons (perhaps matrisomal neurons) elicits different motor responses (Alexander and DeLong, 1985). Mink (1996) argued that inhibitory connectivity between different striatal neurons is the mechanism that enables an animal to select a desired response and inhibit other undesired responses. This supports the idea of using winner-take-all networks to model action selection.

The SNc projects dopamine (DA) to both the striosomes and matrixosomes. Although DA is projected to the striatum via three pathways, the focus here is on the nigrostriatal pathway which regulates learning in the corticostriatal pathway. Synaptic plasticity was found in the corticostriatal pathway (Wickens et al., 1996). In addition to presynaptic and postsynaptic stimulation, DA projected to the striatum is needed for modifying synaptic strength in this pathway (Wickens et al., 1996; Wilson, 2004). The learning rule in the corticostriatal pathway is known as the three-factor rule or DA-based Hebbian learning rule (Reynolds and Wickens, 2002; Wickens et al., 1996).

1.2 Neural Correlates of Reward-based Learning

Schultz and colleagues (Ljungberg et al., 1992; Schultz et al., 1993, 1997; Schultz, 1999; Waelti et al., 2001) have studied the relevance of DA neurons to reward-based learning. By recording from DA neurons of the SNc while monkeys performed an instrumental conditioning task, Schultz et al. (1993) found that on early training trials these neurons increased their firing rate at the time of reward delivery. The DA neurons were phasically activated with a 50 -110 ms latency after the reward was delivered and had a duration of about 200 ms (Schultz, 1999). Further, they found that on late training trials after the animal learned the task, DA phasic responses were time-locked to the appearance of the light (conditioned stimulus; CS), and most importantly, there were no DA phasic signals associated with reward delivery. This phenomenon is known as time shifting of DA phasic responses. Further, in trained animals, nondelivery of a predicted food reward (when the response was incorrect) was associated with depression of DA phasic responses at the time of the expected reward. These findings suggest that DA neurons respond to predictors of reward, rather than to primary reward, and thereby subserve reward prediction learning. Further, the response properties of these neurons are consistent with the TD error signal used in TD learning.

In the above-described instrumental conditioning task, the monkey was always rewarded for making the correct motor response. In the partial (probabilistic) reinforcement paradigm, however, the subject is rewarded on some, but not all, trials when the correct motor response is made (Fiorillo et al., 2003). Fiorillo et al. (2003) recorded from DA neurons while a monkey performed a Pavlovian conditioning task in which different (conditioned) stimuli were associated with receiving a reward at different probabilities: 0.25, 0.5, 0.75, and 1. That is, the reward may or may not be presented to the monkey after stimulus presentation, depending on the probability of receiving reward associated with that stimulus. At the behavioral level, they found that the monkey showed licking behavior (conditioned response) in response to the CS presentation. They also found that the licking behavior increased with reward probability. At the neu-

ral level they found, like some other DA recording studies, that DA neurons transiently responded to reward and reward-predicting stimuli. Further, they found that the DA phasic signal was larger for the conditioned stimulus that was more probably associated with receiving reward. Also, the magnitude of the DA phasic signal at the time of reward was inversely correlated with the probability of receiving reward. Our simulation results show that this is also consistent with the TD error signal used in TD learning.

In another study, Tobler et al. (2005) studied the firing patterns of DA neurons while a monkey performed a Pavlovian conditioning task in which different (conditioned) stimuli were associated with rewards having different magnitudes or amounts of juice. At the behavioral level, Tobler et al. found that the monkey showed licking behavior in response to the presentation of the conditioned stimuli and the amount of licking was positively correlated with the expected reward magnitude. At the neural level, also like some other DA recording studies, they found that DA neurons transiently responded to reward and reward-predicting stimuli. They also found that DA neurons differentiated between reward magnitudes. DA neural responses were larger for CSs that were associated with a larger reward magnitude. Behavioral studies also found that larger rewards were associated with faster learning in the rat (Pearce, 1997). Our simulation results are consistent with all of these findings.

1.3 Maintenance of Information in WM in PFC and BG

A range of methods have indicated that both PFC and the BG are involved in maintenance of information in WM. These methods include imaging, lesion studies, and electrophysiological recordings. We shall survey some of the evidence pertaining to DLPFC first.

Using fMRI imaging, Cohen et al. (1997) found that DLPFC subserved maintenance of information in the n -back task. In this task, a sequence of stimuli is presented in a series of trials. A response on a trial is required only if the current stimulus matches the one presented n trials before. Thus, to perform this task, the subject must maintain the previous n stimuli in WM. Memory load can be increased by changing n from 1 to 2 to 3. Cohen et al. (1997) found that DLPFC showed sustained activity while subjects performed the mnemonic aspect of the task. Increasing the memory load was associated with increased DLPFC activity. Evidence that DLPFC subserves active maintenance of information in WM also comes from reversible lesion studies in monkeys. For instance, Sawaguchi and Iba (2001) found that inactivating DLPFC using muscimol, a GABA agonist, interfered with performance of an oculomotor DRT, while it had only a minor effect on performance of a control S-R task.

Electrophysiology in animals reveals that neurons in dorsolateral prefrontal cortex (DLPFC) remain active during the delay interval of a DRT. In particular, Goldman-Rakic (1995) found a class of neurons, known as delay-active neurons, that had this property. Goldman-Rakic (1995) reasoned that these neurons subserved maintenance of information in WM. It may be the case that such activity is maintained by a DLPFC, basal ganglia, thalamic recurrent loop.

We now turn to evidence supporting the hypothesis that the BG play a role in maintaining information in WM. Using 2-DG imaging, Levy et al. (1997) studied the possible role of the caudate nucleus in performing spatial and object delayed alternation tasks. Levy et al found higher levels of metabolic activity in the caudate nucleus when the animal performed these tasks. Increasing the delay interval in the spatial WM task was associated with an increase in caudate activity, suggesting that it might be involved in active maintenance of information.

Vermersch et al. (1999) reported a case study of a patient with damage to the head of the caudate nucleus. The patient showed impairment when performing a memory-guided saccade task. The patient, most importantly, made more errors when the delay interval increased. The existence of a delay-dependent effect suggests that the caudate subserves maintenance of information in WM.

Electrophysiological recording studies provide evidence that the BG subserve WM processes. Kawagoe et al. (1998) recorded from the striatum of a monkey while it performed an oculomotor DRT. They found that striatal the activations of striatal neurons were time-locked to cue presentation. They also found a class of neurons that showed sustained activity throughout the delay interval, similar to those found in DLPFC.

Apicella et al. (1992) studied the patterns of activity in striatal neurons while a monkey performed a delayed go/nogo task. One type of neuron was activated during the delay interval on both go and nogo trials. Hikosaka et al. (1989) found similar results in a study where the monkey was trained to perform a memory-guided saccade task.

1.4 Perseveration: PFC and BG Damage

Artificial lesion studies on our computational model can cause perseverative responses in our simulated DRT. Consequently, this section discusses perseveration in the more complicated WM tasks used in the experimental literature on humans. It is hoped that the phenomena identified in our model bear some relation to these more complex tasks. Behaviorally, perseveration describes inappropriate repetition of a previously learned response when a new response is

required to correctly perform a task. Specifically in WM tasks, its appearance seems to reflect an inability to learn a new response after a previous one has been learned.

It is well known that damage to DLPFC leads to perseveration in the Wisconsin Card Sorting Task (WCST) (Anderson and Tranel, 2002). In this task, a subject maintains an active sorting rule in WM in order to correctly classify cards that are presented one at a time. After completing some number of correct trials, the correct rule for sorting the cards is changed without the subject's knowledge and the subject must relearn to sort the cards. Although the WCST is more difficult than a DRT, elements of the perseveration phenomenon can appear in a DRT. Further, both tasks are subserved by the BG and PFC working in concert (Amos, 2000; Goldman-Rakic, 1995; Pickett et al., 1998).

Further, damage to DLPFC in animals can lead to perseveration in performing the A-not-B task (Diamond and Goldman-Rakic, 1989). In particular, Diamond and Goldman-Rakic (1989) found that both healthy infants and DLPFC-damaged monkeys perseverate while performing the A-not-B task. For the healthy infants, this could be because of immaturity in their DLPFC (Munakata et al., 2003).

In addition to DLPFC-damaged subjects, some Parkinson's disease (PD) patients show perseveration errors while performing WM tasks, particularly the WCST (Cooper et al., 1991; Lees and Smith, 1983). Amos (2000) assumed that the occurrence of perseverative responses in PD patients is caused by cortical dementia. Lees and Smith (1983) and Cooper et al. (1991) however found that PD patients who showed perseverative errors while performing the WCST were not demented. It is possible that striatal DA reduction is responsible for the occurrence of perseverative responses in PD. This has been modeled by Suri and Schultz (1999) and is discussed in the next section.

It should also be noted that the occurrence of random errors, in contrast to perseverative errors, have also been observed in PD patients performing the WCST (Taylor et al., 1986). The PD patients tested in the Taylor et al. (1986) study were not cortically demented. By the process of elimination, this behavioral deficit may have been associated with striatal DA dysfunction.

Whereas the previous discussion focused on the relation to damage to DLPFC and perseveration, it has also been reported that damage to the BG is also associated with perseverative responses (Fung et al., 1997; Levitt et al., 2002; Pickett et al., 1998). Fung et al. (1997) reported four cases, one of which (Case 3) had damage to parts of the BG and thalamus. The patient showed perseverative movement of the limbs. Pickett et al. (1998) reported a case study of a patient who had damage to the putamen and caudate nucleus. This

patient showed a tendency to perseverate while performing the WCST and the Odd Man Out Task, a task similar to the WCST which tests a subject's ability to form and then to shift abstract categories.

Levitt et al. (2002) studied the performance of schizotypal patients on WM tasks. They found that the size of the caudate nucleus was smaller than that of normal subjects. In this study, schizotypal patients showed perseveration while performing a DRT. They also found an inverse correlation between caudate nucleus size and severity of perseveration. The study suggests that a decrease in caudate nucleus size is associated with severity of perseveration.

Lombardi et al. (1999) reported a study that links perseveration to both BG and DLPFC dysfunction. Using PET scans, Lombardi et al. (1999) studied WCST performance on patients with head injury. They found that both DLPFC and the caudate nucleus showed decreased activity relative to normals while patients performed this task. They also found that these patients showed perseverative responses.

1.5 Systems-Level Models of WM

There are existing models that simulate DRT performance. One of the earliest is that of Dehaene and Changeux (1989). This model simulated both the role of PFC in active maintenance of information in WM and the occurrence of perseverative responses as related to PFC damage. The model had two modules. The first consisted of an input layer directly connected to an output layer. The purpose of this module was to associate stimuli with motor responses. The second module subserved maintenance of information in WM. This module's performance was modified by learning that depended on reward presentation. This model showed that damage to, or not incorporating, the second module, which represented PFC, led to the occurrence of perseverative responses in DRT tasks. This model, however, did not address the role of the BG in WM processes.

Frank et al. (2001) proposed a model to simulate performance in the 1-2-AX task, a WM task that requires the subject to gate two cues into WM in order to correctly select a response to a target sequence. Specifically, the subject is presented with a sequence of stimuli, one at a time, consisting of the stimuli 1, 2, A, B, X, or Y. If the subject last saw a 1, then the target sequence is an A followed by an X. If the subject last saw a 2, then the target sequence is a B followed by a Y. This model assumed that the function of the BG was to gate information into WM, while the function of PFC was active maintenance of information in WM. Gating of information into WM was subserved by the PFC-BG-PFC pathway. The Frank et al. (2001) model

was trained using the Leabra algorithm (O’Reilly, 1998). O’Reilly and Frank (2006) proposed a similar model that simulated action selection performance in the 1-2-AX task. This model incorporated the BG indirect pathway. These models assumed that selecting memory-guided motor responses is subserved by a corticocortico pathway. These models did not incorporate motor processes of the BG.

Braver and Cohen (2000) proposed a model that simulated performance in the AX-CPT task. In this task, a subject is presented with the sequential letter stimuli A, X, B, and Y, and is asked to detect the specific sequence of an A followed by an X. The model incorporated interactions between sensory association cortex, DLPFC, the ventral tegmental area, and cortical motor areas. The model assumed that dopamine neurons of the ventral tegmental area subserved gating of information into WM. One limitation of this model is that it did not incorporate the BG.

The closest existing model to our proposed model is that of Suri and Schultz (1999). This model simulated a spatial DRT. Like our model, this model incorporated an actor-critic architecture and was trained using the TD algorithm. Unlike the Frank et al. (2001) model, it assumed that the striatum subserved memory-guided motor responses. It also assumed that lateral connectivity of matrisomal neurons, simulated by a winner-take-all network, subserved action selection. Further, by training the model using an unconditional reinforcement signal, Suri and Schultz (1999) found that the model generated perseverative responses comparable to those found in PD patients. An unconditional reinforcement signal is one that is always associated with the time of the primary reward, whether predicted or not. It does not time shift to the occurrence of reward predicting stimuli. Their results suggested that inappropriate time shifting of the DA phasic signal can explain the occurrence of perseverative responses in PD. The model did not account for how striatal or prefrontal damage is associated with perseverative errors. Also, the model did not simulate gating of information into WM.

Berns and Sejnowski (1996) proposed an actor-critic model that assumes the BG subserve action selection. Like Suri and Schultz (1999) their model was trained using the TD algorithm. However, the neural substrate of action selection in this model is different from that of Suri and Schultz (1999). In particular, action selection occurs in the GPi which consists of a number of units that control the selection of a different motor response. In this model, the STN inhibits all the GPi units except the winner. Their simulation results successfully predicted that damage to the STN leads to the inability to stop selected actions.

Amos (2000) proposed a model that incorporated interactions between sensory association areas, PFC, and the BG. The model simulated performance in

the WCST. It assumed that PFC maintains the sorting rule (card, color, or shape). The sensory association cortex encoded representations of input stimuli and the striatum integrated cortical information and decided what action to perform. Feedback to PFC from the BG informed PFC whether to maintain or change the sorting rule (not modeled). The model simulated the occurrence of perseverative responses in PFC-damaged subjects and random responses in PD. It did not simulate the occurrence of perseverative responses in striatal-damaged subjects. Dopamine reduction was simulated by decreasing the gain parameters of the sigmoidal activation function and lesioning was simulated by decreasing the output of neurons representing the lesioned area.

1.6 Uniform Selection Hypothesis

Our model embodies the uniform selection hypothesis put forth in the introduction and articulated by Prescott et al. (2003) in the context of WM. Redgrave et al. (1999) argued that since an animal has the ability to select among many different motor actions, such as forage for food or escape from a predator, a brain mechanism is needed to decide which action to select under which conditions. The uniform selection hypothesis extends this idea to the cognitive domain. Berns and Sejnowski (1996, p. 102) articulate a similar position in a more general context stating that “the basal ganglia have classically been considered primarily part of the extrapyramidal motor system, that is, part of the motor system concerned with automatic movement, but a wealth of new data now supports an extended role for the basal ganglia that include an analogous function of cognitive processes.” However, existing models have either incorporated motor or cognitive functions of the striatum but not both (Berns and Sejnowski, 1996; Beiser and Houk, 1998).

In this paper, we integrate features from previous systems level models and combine both motoric and cognitive mechanisms and then study their joint effect in the context of a DRT. Our model hypothesizes that the BG learn to select information for gating that is associated with reward presentation (Houk et al., 1995) in the same way that they learn to make a response which is associated with reward presentation. We hypothesize that by training the model using the TD algorithm and simulating motoric and cognitive gating processes of the BG within the actor-critic framework, our model will be able to account for a wide range of experimental data, including lesion and recent reward-based studies.

2 Methods

As stated at the start of the paper, we have developed and tested a simulation of performance in a DRT. The simulation implements an instantiation of the uniform selection hypothesis. We shall first describe the DRT protocol used in the simulation and then describe the model architecture and algorithms used. The model was implemented using MATLAB and the source code is available for download at: <http://www.cacs.louisiana.edu/~maida> under downloads.

2.1 Task

A DRT is a two-alternative, forced-choice response task in which the subject (in our case the subject is the computer simulation model) is trained to make one of two responses: R1 or R2. Selecting the correct response depends on which cue was presented at the beginning of the delay interval. By default, the model is trained to choose R1 when cue A is presented and R2 when B is presented. In the simulation, time is divided into eleven discrete segments intended to correspond to intervals of a few hundred milliseconds. A cue, either “A” or “B”, is randomly chosen, and presented at the beginning of the second time step. It then disappears at the end of that step. At the beginning of the tenth time step, a trigger stimulus labeled “X”, is presented. This event signals the model to select a response. If the response is correct, a reward, whose value is +1 is given at step 11. If the model responds before the trigger presentation, the trial ends immediately and the model receives a slight negative reward of -0.1 .

Maintaining the presented cue in WM over the delay interval is necessary for selecting the correct motor response with 100% accuracy. After learning to wait for the trigger, the simulation must learn to gate and maintain the presented cue into WM over the delay interval. This is needed so that the model can learn to associate the cue with the correct motor response at the time of the trigger. In this paper, we do not address the question of how the cue is flushed from WM after the trigger is presented and a response is made. Further, the architecture is designed so that once a cue is gated, it is easy for the model to learn to maintain the cue in WM over the delay interval.

2.2 Architecture Overview

The model architecture is shown in Figure 1. It takes the form of an actor-critic architecture, in which the matrisomes and their incoming weights represent

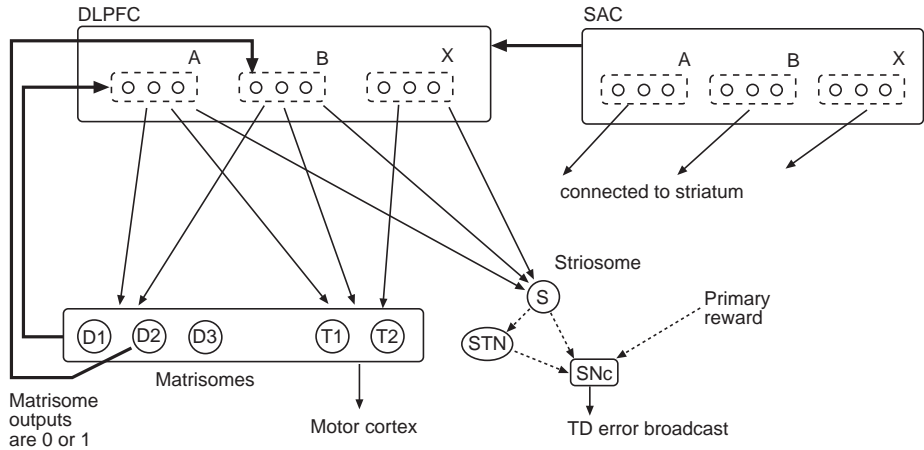


Fig. 1. Model architecture. Thin arrows represent convergent/divergent connections, thick arrows represent topographic connections, and dashed arrows represent connections that are implemented implicitly. SAC and DLPFC are connected to all matrisomes and the single striosome. The striatum is subdivided into matrisomes and striosomes. The matrisomes are subdivided into D-units, T-units. There is one striosome unit (S). WM is maintained by a recurrent loop going through the D-units of the matrisomes.

the actor while the single striosome and its incoming weights represent the critic. The matrisomes and striosomes combined compose the striatum. The architecture models sensory association cortex (SAC), DLPFC, the striatum, and the nigrostriatal pathway. It is assumed that experimental stimuli are initially gated into DLPFC on the basis of earlier perceptual circuits. Except for SAC, these perceptual circuits are not modeled. Representation of a stimulus within DLPFC takes the form of a population code of stimulus-specific features. In the model, the population code consists of three units.

In our model, the striosomes have only one unit, project to the SNc, and compose the critic in the actor-critic architecture. The matrisomes compose the actor which is in turn subdivided into D-units and T-units (our hypothesis). The D-units and T-units form separate winner-take-all (WTA) networks or circuits. Within a circuit, the unit with the highest activation inhibits the other units in the circuit so that their activation becomes zero. Further, they form segregated pathways where the D-units form a closed loop projecting back to DLPFC and the T-units are embedded in an open loop projecting to motor cortex. Appropriate D-units may remain active during the entire delay interval as part of a recurrent, closed loop, circuit to maintain a cue in WM. In contrast, the first T-unit to become transiently active above a threshold initiates a motor response associated with that unit. The activation of a T-unit is not sustained because it is not part of a closed loop. The nine units representing the stimuli (A, B and X) in DLPFC are fully connected to the

Parameter	Value	Parameter	Value
initial weight values in actor	0.3	σ_{noise}	0.025
initial weight values in critic	0.0	gating amplification	2.0
β	3.3	gating threshold	0.75
α	0.003	motor threshold	0.78
correct motor response reward	1.0	premature response reward	-0.1

Table 1

Parameters and their values used in the intact model.

D-units, T-units, and striosome. With two exceptions, the nine units in SAC are also fully connected to the D-units, T-units, and striosome. The exceptions are: 1) One of the connections from SAC A to unit D2 is clamped to zero; 2) One of the connections from SAC B to unit D1 is clamped to zero. These are called dead weights. Their purpose of this is to create sparser connectivity from A to D2 and similarly from B to D1, thereby creating a natural link from A to D1 and B to D2. Since the names of D1 and D2 are arbitrary, this simply means that A projects more strongly to one region of the striatum and B to another.

The striosome units receive linear external input but undergo separate WTA competitions involving the D-units and T-units. The outputs of the WTA competitions are discrete, but for the D-units, they are amplified by a factor of two as they project back to DLPFC. This makes it easier for the DLPFC gating loop to be self sustaining. Learning occurs in all of the connections from the SAC and DLPFC to the striatum. The loop from the D-units to DLPFC represents topographic connections but with the following caveat. There is, for instance, only a single D1 unit, but three DLPFC units representing cue A. Therefore, the D1 unit sends an identical connection to each of the A units in DLPFC. There is no learning in this feedback loop. In order to learn to reliably gate and maintain a cue in WM, learning must take place in the incoming weights to the D-units.

Normally, the activations of the nine SAC and nine DLPFC units are binary. However, if a cue is gated into WM, then the corresponding DLPFC units are active on that time step and have an activation of two instead of one. The other DLPFC units are inactive.

There are two categories of inputs to the model. The first category consists of sensory inputs in the form of cue and trigger presentations. These are initially presented to the sensory association (SAC) cortex. In SAC, each of the three stimuli (cue A, cue B, and trigger X) are represented by a population of three binary units. That is, the representations of A, B, and X are (1, 1, 0), (0, 1, 1), and (1, 0, 1), respectively. There is one exception where the DLPFC

representation is amplified as described above. In this case, the representation for, say, Cue A would be (2, 2, 0). The second category of input to the model is primary reinforcement, presumably coming from the lateral hypothalamus. If the model makes a premature response before trigger presentation, then it receives a small negative primary reward of -0.1 in the following time step. If the model selects a correct response at the time of trigger presentation, it then receives a positive primary reward of 1.0 in the following time step.

SAC sends topographic projections to the units representing A, B, and X in DLPFC. SAC also sends fully connected convergent and divergent projections to the three D-units, the two T-units, and the single striosome unit in the model (see Figure 1). The DLPFC representation of A, B, and X is the same as that in SAC. The DLPFC units project to the D-units, T-units, and striosome unit in exactly the same fashion as does SAC, with the exception that SAC some of the units in SAC have slightly sparser connectivity as explained above. When a stimulus is presented to SAC, the transmission delay across all connections is assumed to be small enough to ignore within a simulation time step. The activations propagate to all of the units in the system within this interval.

2.3 Formal Description of the Model

In regard to the actor-critic architecture, the actor selects an action whereas the critic computes a reward prediction-error value for a chosen action. The learning rule modifies the action selection policy on the basis of reward prediction errors computed by the critic. More specifically, the critic computes a TD error signal on the basis of primary rewards and reward predictions. The signal is based on the difference in predicted reward over two consecutive time steps in combination with any primary reward received on a time step. This particular version of the rule is based on Houk (1995). Although this equation is used, the neural circuit to implement the memory for the previous prediction is not explicit in the model (indicated by dashed arrows in Figure 1). Houk et al. (1995) show how this equation might be implemented using a subthalamic sideloop.

$$\text{TD}(t) = R(t) + P(t) - P(t - 1) \quad (1)$$

In the above, $\text{TD}(t)$ is the TD error signal at time t , $P(t)$ is the predicted reward at time t . This TD signal is not only used to modify the action selection policy. It also modifies the reward prediction policy of the critic. In this particular model, the reward prediction, $P(t)$, is computed as follows. The

quantity, h_{str} , is defined in Formula (3).

$$P(t) = \tanh(\beta \cdot h_{\text{str}}(t)) \quad (2)$$

This is the output of the striosome. $P(t)$ is the hyperbolic tangent of the net input, $h_{\text{str}}(t)$, to the striosome unit. It squashes linear input into a range that varies continuously from -1 to $+1$. This activation function was selected because its output is negative when its input is negative and its output is positive when its input is positive. Thus, it can naturally encode positive and negative reward predictions. β is a gain parameter that makes the function more closely approximate a step function. In all of the simulations, $\beta = 3.3$. This value was chosen so that in the event that all of the inputs were all equal to one, the value of the quantity h_{str} , defined below, would still fall in the linear range of the function. The quantity $h_{\text{str}}(t)$ is the weighted sum over the input cortical units and is defined below.

$$h_{\text{str}}(t) = \sum_{i=1}^{18} w_i(t) \cdot x_i(t) \quad (3)$$

The activations of the 18 cortical units at time step t within a trial are given by $x_i(t)$ and the weight value between unit i and the striosome, S, is given by w_i . Nine of the units are from SAC and nine of the units are from DLPFC. These weights are the incoming weights to the striosome and form the trainable weights for the critic. They are initialized to zero for each simulation run. This has the effect of making the initial reward predictions of the critic equal to zero. The critic refines its predictions over time and the refined policy is implemented by changes in the incoming weights to the striosome. The rule to describe these weight changes is given below. Thus, the change in weight, w_i , at time step t is proportional to the product of the TD error at that time step and the input value to the weight at that time step.

$$w_i(t+1) = w_i(t) + \alpha \cdot \text{TD}(t) \cdot x_i(t) \quad (4)$$

In the above, α is a learning rate parameter. In all of the simulations, $\alpha = 0.003$.

Let us now turn to the actor. Let $M_j(t)$ be the activation of a matrisome unit j — either a D-unit or a T-unit — before the WTA competition. $M_j(t)$ is given by the formula below. The formula is intended to encode the firing rate of the matrisome unit in question.

$$M_j(t) = f(h_j(t)) \quad (5)$$

The activation function, $f(\cdot)$, is the logistic sigmoid given below and the quantity, $h_j(t)$ is defined in Formula (7).

$$f(x) = \frac{1}{1 + e^{-x}} \quad (6)$$

This function has an S-shape similar to the hyperbolic tangent function used in Formula (2), with the exception that its values range from 0 to 1. A gain parameter was not included in Formula (5) because the gain was set to one for all simulations.

Let $h_j(t)$ denote the net input to a matrisomal unit j , either a D-unit or T-unit, at time t . The net input is given by

$$h_j(t) = \sum_{i=1}^{18} u_{ji}(t) \cdot x_i(t). \quad (7)$$

The above formula applies to the actor and is analogous to Formula (3), which applies to the critic.

In order for the model to have some level of nondeterminism in its responses prior to training, the matrisome weights (but not the striosome weights) are perturbed with zero-mean, additive Gaussian noise with $\sigma_{\text{noise}} = 0.025$. The quantity $u_{ji}(t)$ is the perturbed weight from unit i in either SAC or DLPFC to unit j in the matrisome layer. Thus, all of the incoming weights to the matrisomes can be decomposed as follows,

$$u_{ji}(t) = w_{ji}(t) + \delta_{ji}, \quad (8)$$

where $w_{ji}(t)$ is the true, unperturbed weight value at time step t within a trial. Unit i is either one of the nine SAC units or one of the nine DLPFC units. Unit j is either one of the three D-units or one of the two T-units. δ_{ji} is the Gaussian noise value added to the weight before the start of the trial.

When a weight is updated in the actor, the change is made to the true weight, w_{ji} , and not the perturbed weight, u_{ji} . The weight update rule for the actor is

$$w_{ji}(t+1) = w_{ji}(t) + \alpha \cdot \text{TD}(t) \cdot x_i(t) \cdot M'_j(t). \quad (9)$$

This is the three-factor DA-based Hebbian learning rule that was mentioned at the end of Section 1.1. The change in weight, w_{ji} , at time step t is proportional to the product of the presynaptic activity x_i , the postsynaptic activity M'_j , and the TD error at time t . α is the same learning rate parameter that was used

in Formula (4). $M'_j(t)$ is the activation of unit j after the WTA competition. The post-WTA activation will be either 0 or 1. However, the pre-WTA value, $M_j(t)$, is remembered so that it can be tested against a gating threshold. Thus weights are adjusted only for winning units. In both formulas (4) and (9), weights are updated incrementally after each time step within a trial. We note in passing that the dead weights mentioned in Section 2.2 do not undergo learning.

Next we describe the closed-loop feedback from the D-units to DLPFC. The winning D-unit projects its activation to the three DLPFC units representing the associated cue. For example, if the D1 unit wins the competition and its pre-WTA activation is above the gating threshold, then it projects its activation to the three DLPFC units representing Cue A. Further, the activation of these units is amplified by a factor of two and the activation of the other DLPFC units is suppressed (as if they were in a WTA). This has the effect of giving a cue that has been gated into WM, the advantage in terms of being maintained in WM. This positive feedback mechanism assists the TD algorithm in maintaining the cue in WM. Even with this positive feedback, the gated cue does not always stay in WM during a trial. Learning is still necessary for the model to maintain the cue in WM reliably.

Lastly, we describe the open loop behavior of the T-units. They are subjected to a WTA competition, but the pre-WTA activation is also remembered. If a T-unit wins the competition and its pre-WTA activation is greater than a motor threshold, then the response associated with the T-unit is assumed to be selected. In the simulation, the motor response threshold is 0.78 and the gating threshold is 0.75. The threshold values are held fixed for all runs across all simulations, except for one simulation where DLPFC is lesioned. In this case, the motor response threshold was set to 0.65. Other than this, the T-units and D-units are identical with the exception that the T-units are embedded in an open loop and the D-units are embedded in a closed loop. This design decision is a reflection of the uniform selection hypothesis.

2.4 Initial State of the Model

We now make some remarks about the initial state of the system before the simulation begins. In regard to the critic, its weights are initialized to zero. Since the weights are zero, the output of the striosome is initially zero regardless of the input. Thus, the initial reward predictions of the critic are zero, until reward prediction errors, caused by unanticipated primary rewards, trigger weight modification in the critic. Further, if the critic never receives primary reward, there will be no weight modification in the critic, and the weights will remain zero.

In regard to the actor, there must be enough initial random behavior so that the model selects a full range of responses that may be appropriate. In particular, the probability of making either motor response must be nonnegligible. Also, the probability of gating either cue to memory must be nonnegligible. This variability is obtained by setting the initial weights in the actor to a value of 0.3 plus a noise parameter given in Formula (8). Noise is applied to the actor weights at the beginning of each trial. Its mean is zero so that perturbations in the weights occur symmetrically about the weight's value. The standard deviation of 0.025 was selected so that there is sufficient exploratory behavior in the model to explore the entire action space. At the same time, there cannot be so much noise in the weights that the model is unable to form a stable representation of a skill. If the T-unit with highest activation is above some threshold, the unit initiates a motor response. However, the D-units have identical connectivity and identical initial weight values. Further, gating must be able to happen before a motor response is made. This means that the gating threshold must be somewhat lower than the motor response threshold.

3 Results

The simulation results fall into three categories: 1) performance of the intact model under reward conditions using the standard DRT protocol; 2) performance of the lesioned model under reward conditions corresponding to the standard DRT protocol; and, 3) performance of the intact model when the reward conditions are varied, such as by applying partial reinforcement or behavioral extinction.

3.1 *Intact Model*

The intact model was not lesioned and was trained using the basic DRT reward protocol described in Section 2.1. The results in this section serve as the baseline of comparison to the later simulations. Performance of the intact model for three separate runs is shown in the top row of Figure 2. Response choice for each of the above runs is shown in the bottom row of the figure. The first point to note is that the pattern of performance and responses is stable across runs. All simulation runs consisted of 1500 trials. For the purposes of analysis, the data from all simulation runs was divided into 30 blocks of 50 trials each. Blocks of this size were chosen because that was an adequate grain size for summarizing the results. For comparison, Suri and Schultz (1999) used blocks of 40 trials. For the intact model, it took about 10 blocks for motor performance to reach asymptote. It seemed appropriate to run simulations for

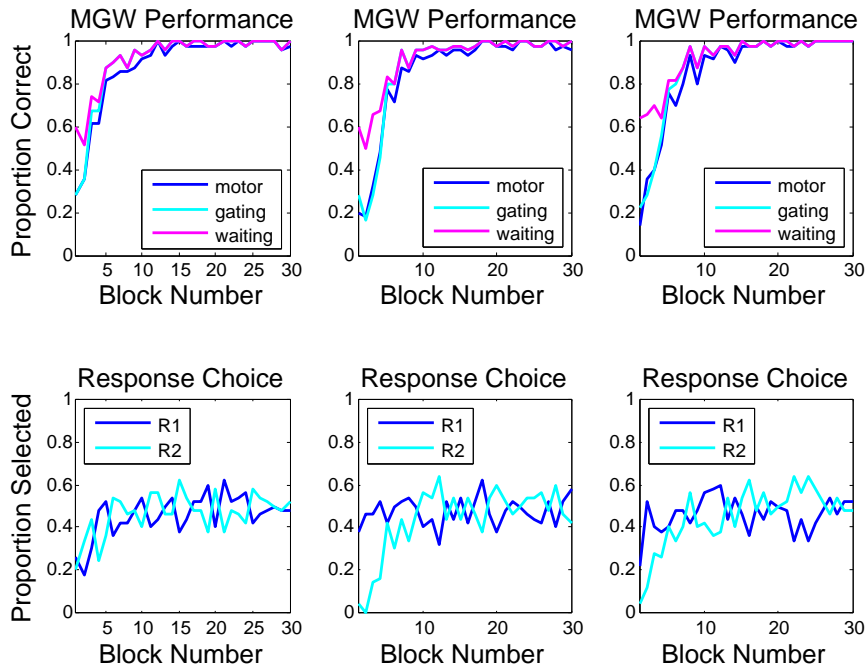


Fig. 2. Intact performance. Top row: Performance as indicated by the measures motor, gating, and waiting versus training for three separate runs. Bottom row: Response choice as a function of training for the corresponding three separate runs shown in the top row.

three times this length to study the lesioned models. The main performance measure for the model was the percent of correct responses for a block of trials. This reflected the model’s ability to learn the task. The model learned to perform the DRT on the basis of primary reinforcement applied according to the protocol given in Section 2.1. Initial performance was below 50% correct because the model had to learn to wait for the presentation of the trigger at the end of the delay interval in order to be able to receive positive reward for a correct response. The plot in this figure also shows waiting performance and gating performance. Waiting performance is the percentage of trials in a block that the model correctly waits for the trigger before responding. This assesses the model’s mastery of the first skill required for learning the task. Since the model must correctly wait in order to respond correctly, waiting performance must always be at least as high as motor performance. Gating performance is the percent of trials in a block that the model gates the correct cue into WM and, once gated, maintains it over the entire delay interval. If the model is to respond correctly on more than 50% of the trials, it must also correctly gate the cue into WM. This is reflected in the result that gating performance exceeds response performance after the first few blocks of training.

Figure 2 (bottom row) shows the response selected by the model over the

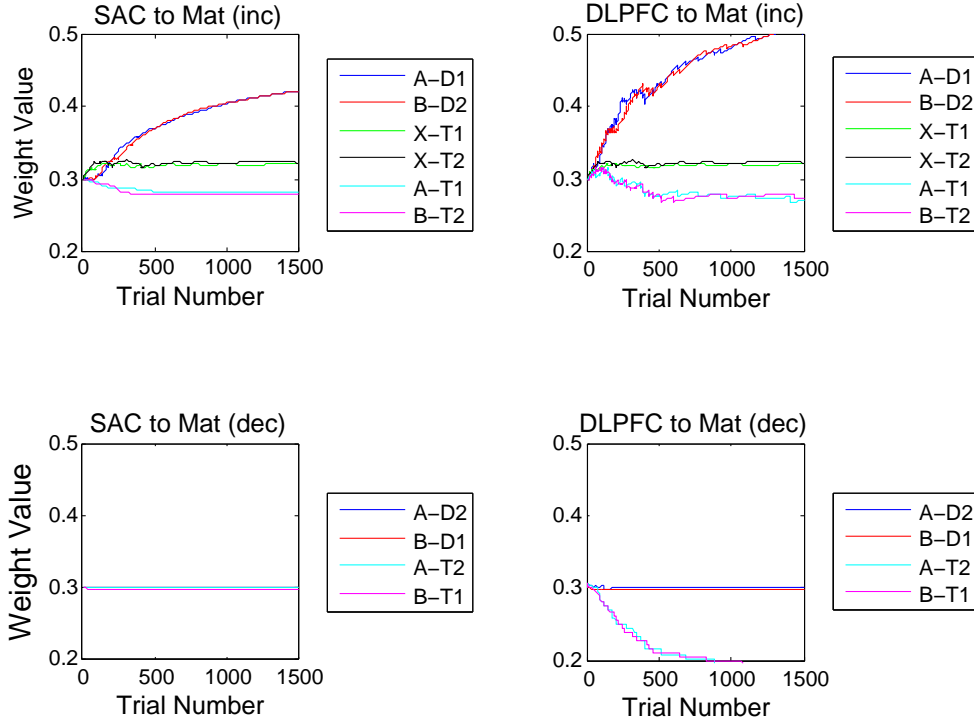


Fig. 3. Weight values versus trial number for the actor in the intact model. These weights are from the simulation shown in the left column of Figure 2. Left: Weights from SAC to the matrisomes. Right: Weights from DLPFC to the matrisomes. Key: “inc” indicates weights expected to increase with training, “dec” weights expected to decrease. The notation “A-D1” means a weight going from one of the three units in A to the D1 unit.

course of training for the corresponding three separate simulation runs. After learning, the model selected each response choice with roughly equal probability and did not exhibit any tendency to perseverate.

Figure 3 shows the weight changes in the actor during the course of learning. These weights are from the simulation shown in the left column of Figure 2. Initial values for all actor weights are 0.3. We classified the weights into four groups: 1) those originating in SAC and expected to increase in strength as the agent learns the task; 2) those originating in SAC and expected to decrease as the agent learns the task; 3) those originating in DLPFC and expected to increase as the agent learns the task; and finally, 4) those originating in DLPFC and expected to decrease as the agent learns the task.

The logic for expecting a weight to increase is as follows. The task requires that the gating, maintenance, and motor operations work correctly. For gating, the connections from the representation for A in SAC must map to D1

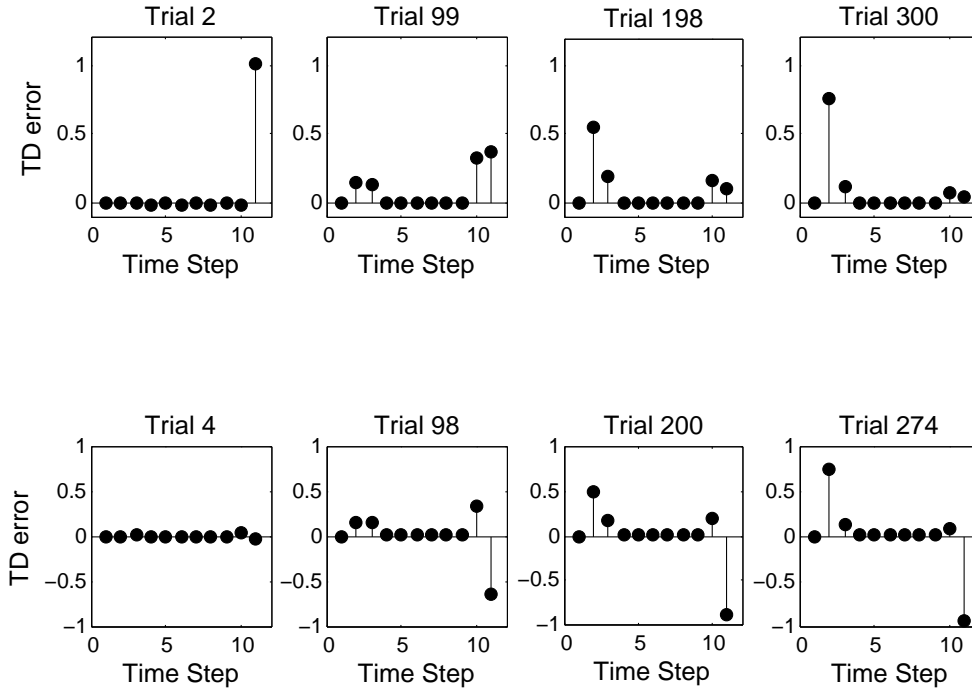


Fig. 4. Top row: Each graph shows the TD error for a single trial on which the model selected a correct response and the primary reward value was 1. For later trials, the TD error signal occurs at the time of cue presentation. Bottom row: Each graph shows the TD error for a single trial on which the model selected the incorrect response, with the constraint that the model did not make a premature response. The primary reward value was zero for all time steps. Thus the error depends on reward predictions only.

in the matrixomes. For gating cue B they must map to D2. For maintenance, the appropriate cue representation must be active in DLPFC. For the proper cues to be maintained in DLPFC, there must be an association from the units representing A in DLPFC to D1 (to create an intact WM loop). Similar logic holds for the association from B to D2. For selecting a motor response, the units representing A in DLPFC must be associated to T1 and those representing B must map to T2. These latter associations should not be so strong that they alone drive a T-unit above threshold. If that were to happen, the model would make a premature response. Thus the presentation of cue X should be necessary to drive a T-unit above threshold to trigger a response. Thus, the strengths of all the above-mentioned connections are expected to increase as learning proceeds. The strengths of connections that have not been mentioned are expected to decrease.

The observed pattern of weight updates is generally consistent with the above

predictions. There are two types of exceptions. The first exception is that many of the weights that were expected to decrease actually remained constant and kept their initial values. The second exception, is that the DLPFC weights from A to T1 and B to T2 actually decreased. However, the weights from A to T2 and B to T1 decreased even more strongly. It seems the model decreased these weights in the processes of learning to wait for cue X before responding. However, the relative strength of the connection from A to T1 as opposed to the connection from A to T2, was the deciding factor in learning to select response R1 when cue A was presented, and so the latter was decreased more.

We now turn to the performance of the critic and describe the pattern of changes in the TD error signal during the course of learning. Each graph in Figure 4 shows the TD error for the successive time steps within some trial of the 1500 trial simulation run. The top row shows the TD error when the model's response was correct. On each of these trials, the model received primary reward on time step 11 whose value was 1. As learning proceeded the TD error signal moved earlier in time, first to time step 10 with the presentation of the trigger X, and then to time step 2 with the presentation of the cue. This is in accordance with the observed phasic response of DA neurons. The TD error tends to jump across the delay interval because, once a cue is gated, there is better than a 50% chance (by virtue of the structure of the maintenance circuit) that it will be maintained for the entire interval without learning. The bottom row of the figure shows the TD error signal for trials that were not rewarded. These graphs were selected with the constraint that the model did not make a premature response. On these trials, the model waited for cue presentation but made the incorrect response, and did not receive a reward on trial 11. In these trials, the time of the depression in the error signal matched the temporal aspect of the phasic response of DA neurons, however the magnitude of the depression was much higher than that observed in such neurons.

3.2 Lesioning the Model

The effects of lesioning different components of the model on DRT performance were also studied. The purpose was to see if the effects of lesioning the model were similar to observations in clinical studies.

3.2.1 Disabling Learning in the Critic

We examined the model's functioning when learning in the critic was disabled. We did this in two ways. First, we fully disabled learning in the critic to implement unconditional reward as previously studied by Suri and Schultz

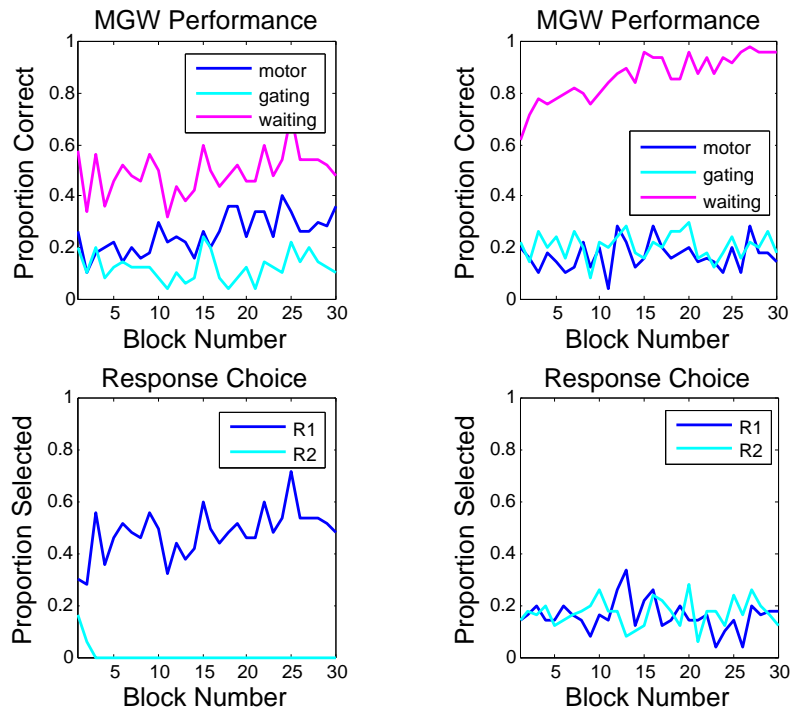


Fig. 5. Typical simulation runs for lesioning the reward system. Left column: Unconditional reward implemented by disabling learning in the critic. Right column: Model does not receive positive reward. Top row: Performance as indicated by the measures motor, gating, and waiting versus training. Bottom row: Response versus training. Key: MGW - motor, gating, and waiting.

(1999). Second, we modified the TD rule in Formula 1, so that it did not use positive primary reward but continued to use negative primary reward under the assumption that, perhaps, positive reward and negative reward use different neural circuits.

When learning is fully disabled in the critic, the effect is to implement unconditional reward which some have hypothesized (Suri and Schultz, 1999) simulates DA reduction. Since the critic’s weights are initialized to zero and learning is disabled in the critic, the critic’s predictions are always zero. From Formula (1), it is apparent that, under these conditions, the TD error signal is exactly equal to the primary reward signal, and is thereby time-locked to the primary reward signal. Thus, the DA phasic signal will not shift in time. Although there is still learning in the actor, it only takes place at the time of the reward signal, so the actor may not learn to gate the correct cue into WM.

Figure 5 (left column) shows performance when learning in the critic is disabled. The model correctly waits for the trigger on about half of the trials.

The model perseverates on all of the trials were it has learned to wait for the trigger. That is, after learning, the model perseverates on all trials in which it is allowed to respond. In the simulation run shown in the figure, it perseverates selecting response R1. In other simulation runs, it may perseverate on R2. In the actor, the weights connecting trigger X with unit T1 increase continuously during training and this is the cause of the perseveration. From examining the individual responses on the simulation runs, the first response rewarded is statistically likely to be the response perseverated on. In regard to gating, the model does not learn to gate at all.

Figure 5 (right column) shows performance when the critic did not use positive reward. Since it still used negative reward, it learned to wait for the trigger on over 90% of the trials, in contrast to the situation when learning in the critic was fully disabled. However, the model was never rewarded for selecting a response, because this would be a positive reward. Consequently the model did not learn to always select a response. The model actually responded on about 40% of the trials and these responses were random. More specifically, the number of times R1 was selected was about equal to the number of times R2 was selected. This was similar to the behavior of the intact model, with the important exception that the motor responses in the negative-reward-only model were uncorrelated with the presentations of the cues. This finding is suggestive of the results of Taylor et al. (1986) and is consistent with the assumption (Amos, 2000) that the striatum serves the function of matching of stimuli with responses.

3.2.2 *Lesioning the Matrisomes and DLPFC*

We examined the effect of lesioning the D-matrisomes and of lesioning DLPFC on the model. The D-matrisomes were lesioned by setting the gating amplification parameter to 1.0 instead of the default value of 2.0. This effectively reduced the output of the D-matrisomes (Amos, 2000) making it more difficult for the model to maintain a WM maintenance loop. Lesioning the D-matrisomes destroyed the model’s ability to learn to gate a cue into WM which, in turn, impaired motor performance. Figure 6 (left column) shows lesioned D-matrisome performance. The model did not learn to gate and the motor performance exhibited perseverative errors starting with the second block of trials. In the simulation run shown, the model perseverated on response R2. In different runs, it perseverated on different responses.

DLPFC was lesioned by removing it from the model (i.e., clamping the activations of the DLPFC units to zero). Figure 6 (right column) shows lesioned DLPFC performance. The motor response threshold was set to 0.65 instead of the default value of 0.78. Since there was no DLPFC input to the T-units, there was not enough input to the T-units for them to reach a threshold of 0.78 and

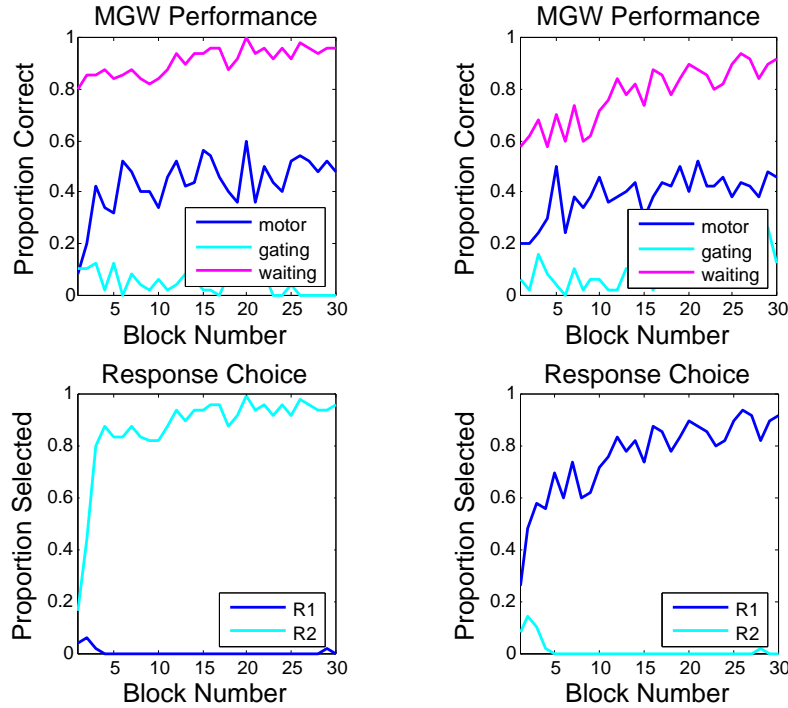


Fig. 6. Lesioning D-matrixes and DLPFC. Left column: Lesioned D-unit matrixes. Right column: Lesioned DLPFC. Top row: Performance as indicated by the measures motor, gating, and waiting versus training. Bottom Row: Response choice versus training. Key: MGW - motor, gating, and waiting.

the model would never select a motor response using the higher threshold. Therefore the model would never receive a reward and, consequently, there would be no learning. With the exception for the need to change the motor response threshold, the model's performance was virtually identical to that exhibited by lesioning the D-matrixes. This is plausible because an intact DLPFC is necessary for gating to be possible. The model did not learn to gate and the model to exhibit perseverative errors, manifested in perseverating on response R1 for the simulation run shown.

In both cases, either lesioning the D-matrixes or lesioning DLPFC, the model exhibited perseverative errors. Also, in both cases, gating performance in the model remained near zero.

3.3 Reward-Based Studies

We studied whether the model accounts for effects found in a selection of reward-based studies, including the partial reinforcement paradigm, the role

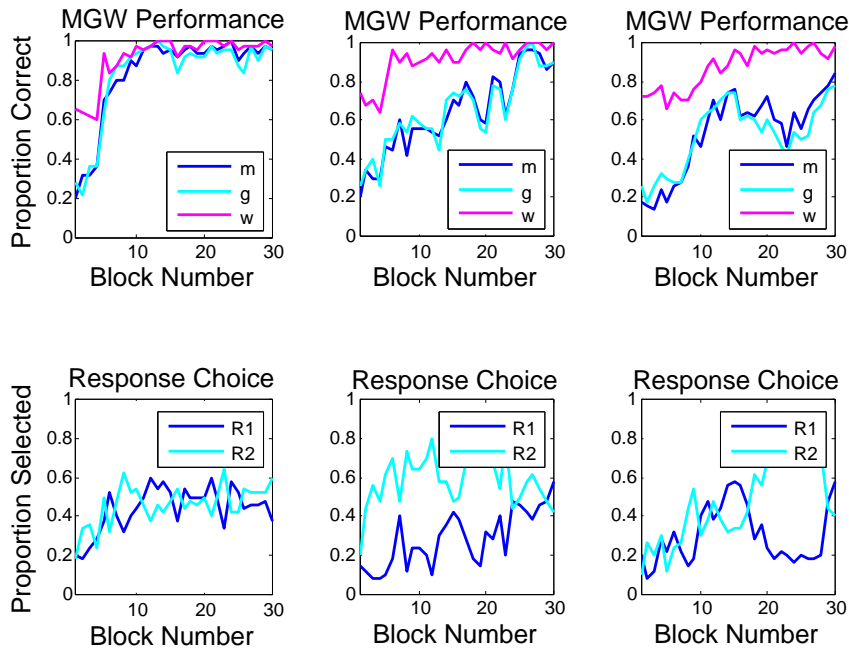


Fig. 7. Partial reinforcement for 1500 trials. Top: Performance as indicated by the measures motor, gating, and waiting versus training. Bottom: Response choice as a function of training. Left: Reward probability equals 0.75. Middle: Reward probability equals 0.5. Right: Reward probability equals 0.25. Key: MGW - motor, gating, and waiting.

of reward magnitude in learning, behavioral extinction, and task reversal. In all of the reward based studies, we used the intact model and the associated learning algorithms were identical. The only difference was the protocol used to apply primary reward or punishment.

3.3.1 Partial Reinforcement

We tested whether the model accounts for the partial reinforcement effect, that is, whether rewarding the model for correct responses on some, but not all, trials affects motor performance and, if so, how (see Section 1.2). The model was trained in three cases where the probability of reward for correct responses was less than one. The reward probabilities for correct responses were 0.75, 0.5, and 0.25. These values were taken from Fiorillo et al. (2003).

In comparison to the case of reward probability of 1.0, the performance in the partial reinforcement conditions was lower on both the motor performance and gating performance measures. This is shown in Figure 7. More important, after training, the TD error signal matched the DA phasic signal reported in

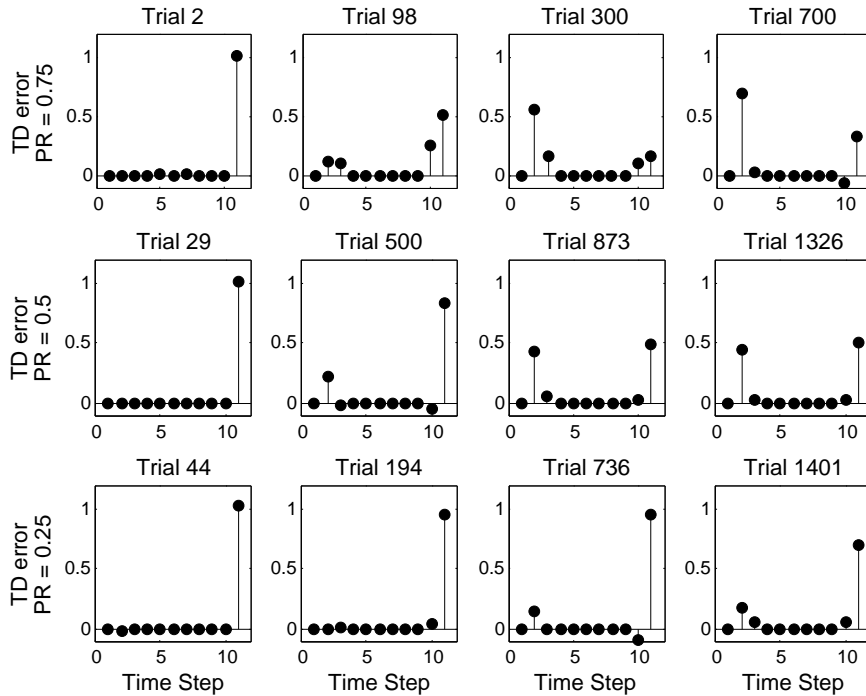


Fig. 8. TD error for partial reinforcement on positive trials. In late trials, TD error magnitude is positively correlated with probability of reward (PR) at the time of cue presentation, but is negatively correlated with PR at the time of reward presentation. Top: TD error when PR for a correct response is 0.75. Middle: TD error when PR is 0.5. Bottom: TD error for when PR is 0.25.

Fiorillo et al. (2003). Figure 8 shows that, after training, the TD error signal was larger at the time of cue presentation when the cue was associated with higher probability for a correct response. Further, the magnitude of the TD error signal at the time of reward was inversely correlated with the probability of reward.

3.3.2 Reward Magnitude

We assessed the effect of manipulating reward magnitude on the model's DRT performance. We used a reward magnitude of 0.5 and 0.25, in addition to a reward magnitude of 1.0, whose performance results are given in Figure 2. The reward was always presented when the model selected a correct response. Motor and gating performance tended to be positively correlated with reward magnitude as can be seen by comparing Figures 2 and 9. Further, motor and gating performance were similar on early training trials, whatever the reward magnitude was. This is in accordance with the fact that motor and gating performance in early training are random, and hence not yet guided by reward presentation. Finally, when the reward magnitude was 0.25, extended training

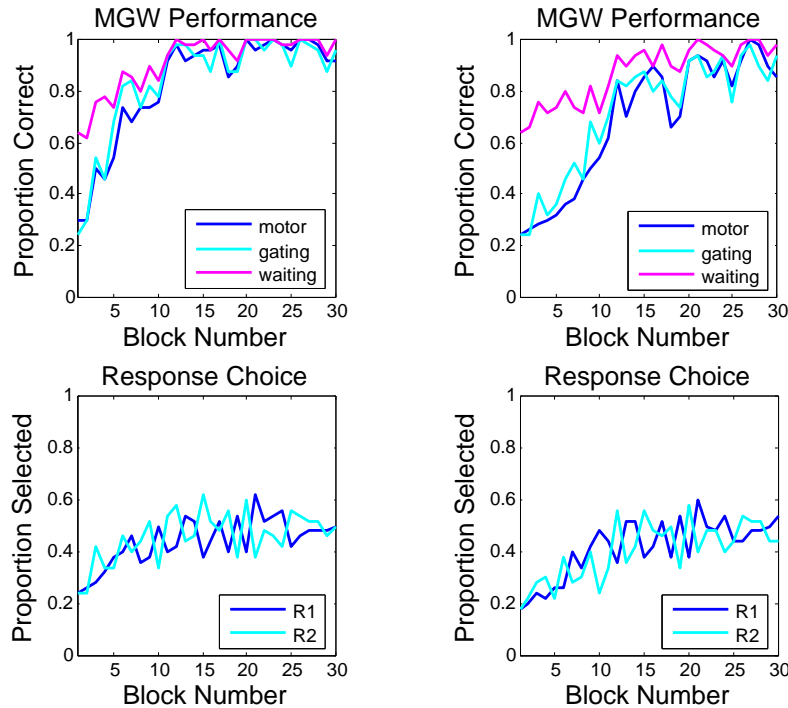


Fig. 9. Manipulating reward magnitude. Left column: Reward value is 0.5. Right column: Reward value is 0.25. Top row: Performance as indicated by the measures motor, gating, and waiting versus training. Bottom row: Response choice versus training. Key: MGW - motor, gating, and waiting.

of the model was required in order for the model to reach a performance level approaching 90%, as shown in the right column of Figure 9.

We also examined the timing and magnitude of the TD error signal. When the reward magnitude is 0.5, the results appear in the top row of Figure 10. When the reward magnitude is 0.25, the results appear in the bottom row of that Figure. The size of the TD error signal was positively correlated with reward magnitude (Tobler et al., 2005). That is, the larger the reward, the larger the effective reinforcement signal at the time of the cue associated with that reward. This can be seen by comparing Figures 4 and 10. These findings are in accordance with the reward prediction hypothesis proposed by Schultz et al. (1997).

3.3.3 Behavioral Extinction and Task Reversal

We assessed the model’s performance under conditions of behavioral extinction and also under conditions of reversal of prior learned responses. Behavioral extinction has to do with losing learned behavior if the predicted reward is

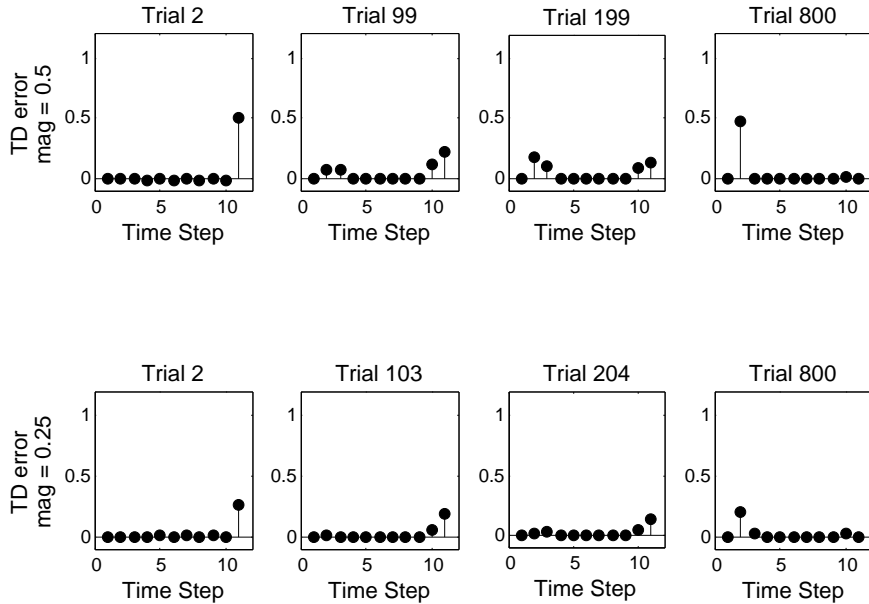


Fig. 10. TD error for positive trials when reward magnitude is manipulated. Top row: Each graph shows the TD error for a single trial on which the model selected a correct response and the primary reward magnitude was 0.5. Bottom row: Same as above, except that reward magnitude was 0.25.

not presented for many trials even when the correct response is made (Pearce, 1997; Suri and Schultz, 1999). We tested whether the model accounts for extinction. We trained the model on the DRT for 1000 trials using the normal reward schedule, then we stopped rewarding the model for selecting a correct response. However, we retained the negative reward that was applied when the model made a premature response. The reason for retaining the negative reward was to ensure the model waited long enough to be allowed to respond. These results are shown in the top row of Figure 11. Like the Suri and Schultz (1999) model, not presenting the reward after the model made the correct motor response led to rapid extinction. Motor performance dropped to zero within 200 trials. In addition, not presenting the reward led to impairment of gating performance as well, and this was not modeled in the Suri and Schultz (1999) model.

We also tested whether the model accounts for delayed-reversal task performance (Pasupathy and Miller, 2005). We trained the model for 1000 trials using the normal reward schedule, then we reversed the values of the correct response. That is, the model was rewarded if response R2 was selected when Cue A was presented and if R1 was selected when Cue B was presented. These

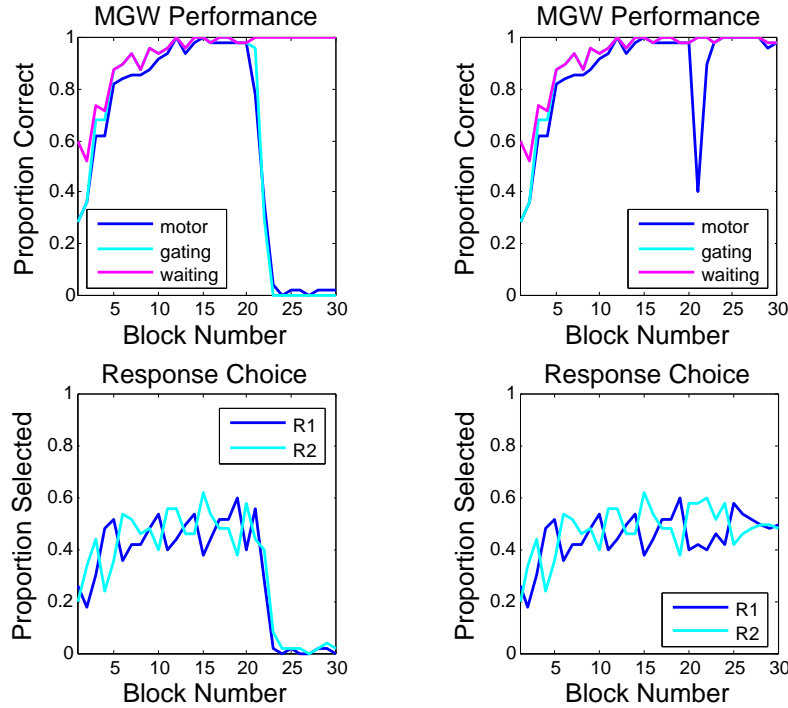


Fig. 11. Left column: Behavioral extinction. Right column: Delayed reversal. Top row: Performance as indicated by the measures motor, gating, and waiting versus training. Bottom row: Response choice versus training. Key: MGW - motor, gating, and waiting.

results appear in the left column of Figure 11. When reversal occurred, motor performance decreased and then increased. The same results applied to gating performance. For reversal, the model was able to learn the reversed response to over 80% accuracy within one 50-trial block after reversal was initiated. Within a 100 trials, the model was able to achieve its near perfect performance on the reversed task.

In regard to TD error shifting, after reversing, the TD error signal went back to the initial state as on early trials in the normal reward schedule (not shown). The TD error signal then shifted in a similar manner to that of normal reward schedule.

4 Discussion

Our model accounts for the essential features of DRT performance. In particular, it accounts for a wide range of experimental data, including lesion and reward-based data. Because the architecture uses the same mechanism

for gating and action selection, the simulation results provide support for the uniform selection hypothesis. In addition, previous models have not explored the computational significance of delay-active neurons in the striatum.

4.1 Perseveration

The occurrence of perseverative responses in our simulation model is qualitatively similar to those observed behaviorally in PD, striatal-, and prefrontal-damaged subjects. Like the Suri and Schultz (1999) model, our model shows that inappropriate shifting of DA phasic signals (unconditional reward) may be the reason that PD patients show perseverative responses in WM tasks. This hypothesis can possibly be tested by recording from DA neurons of a monkey model of PD (induced by administering either MPTP or 6-OHDA) while the animal engages in a DRT. The prediction is that striatal DA reduction in PD will lead to nonshifting of DA responses, which in turn will lead to making perseverative errors.

Further, training the model using an unconditional reinforcement signal reveals the importance of the TD algorithm — the reward prediction hypothesis — for learning to perform DRTs. That is, nonshifting of the TD signal leads to impairment in learning to perform reward-based behavioral tasks. Suri and Schultz (1998) found a similar result for modeling a sequential movement task using the TD algorithm.

The model also accounts for the findings that striatal damage is associated with the occurrence of perseverative errors (Fung et al., 1997; Levitt et al., 2002; Pickett et al., 1998). Damage to the D-matrisomes led to perseveration as follows. Due to inability to gate the cue into WM, it is only the cortical representation of trigger X that determines what motor response to choose. Accordingly, the motor response that will be associated with receiving a reward on early training trials is more likely to be made than the other motor response. This is because of TD learning. Weights connecting trigger X in DLPFC and one T-unit will be larger than that of the other T-unit. Accordingly, the motor response that happens to be associated with the former T-unit will be made on most of the trials in the remainder of the experiment. In short, it could be that damage to the striatum disrupts cortical information that is used to guide the subject in learning to select the correct motor response, and therefore the subject perseverates.

The simulation results also show that damage to DLPFC leads to the occurrence of perseverative responses (Amos, 2000; Anderson and Tranel, 2002). Although both the DLPFC-lesioned model and the D-matrisome-lesioned model break the DLPFC-striatum-DLPFC gating loop, the neural mechanism under-

lying the occurrence of perseverative responses in the two models is not identical. In the DLPFC-lesioned model, the SAC-striatum pathways are solely responsible for selecting motor responses. However, in the D-matrisome-lesioned model, both SAC and DLPFC are responsible for selecting motor responses.

In summary, the model predicts that damage to an element within the DLPFC-striatum-DLPFC gating loop will lead to perseveration of motor responses. This is consistent with the observation of Lombardi et al. (1999) who note, "...it is more accurate to consider the entire dorsolateral frontal-subcortical circuit, rather than the DLPFC alone, as contributing to perseveration in the WCST." A possible common factor among the occurrence of perseverative responses in PD, striatal-damaged, and PFC-damaged subjects may be the impairment of gating performance.

Studies reviewed in the introduction have shown that PD patients make either perseverative (Cooper et al., 1991; Lees and Smith, 1983) or random (Taylor et al., 1986) errors in WM tasks. In those studies, the PD patients were not cortically demented. Assuming that striatal DA degeneration is the only deficiency they had, it is unclear how that can lead to both symptoms in different patient groups. The simulation results suggest that these different behavioral deficits could be due to different characteristics of the DA phasic signal. Striatal DA reduction may have two different effects, perhaps depending on the severity of the disease. Striatal DA reduction may lead to *nonshifting* of the DA phasic signal to the time of the reward predictors, or conditioned stimuli, which in turn leads to the occurrence of perseverative responses (Suri and Schultz, 1999). Our simulation that disabled learning in the critic as a means to implement unconditional reinforcement supports this hypothesis. Alternatively, striatal DA reduction may lead to *nonoccurrence* of DA phasic signals by modifying the TD formula to treat positive rewards as zero. This leads to no synaptic modification in the actor, which in turn leads to random responses.

Amos (2000) modeled the occurrence of random errors in PD patients differently. In his model (not an actor-critic model), striatal DA reduction was simulated by reducing the gain of the sigmoidal units that simulated striatal neurons. Amos did not use a learning algorithm and thus the occurrence of random errors in his model was not related to learning. In our model, striatal DA reduction plays a role in reward prediction learning. Our model suggests that making random errors could be related to learning to perform the task.

4.2 *Reward-Based Studies*

Regarding the reward-based studies, our model also accounts for the partial reinforcement and reward magnitude effects at the behavioral level and the DA

phasic signal at the biological levels. Regarding the partial-reinforcement DRT that we modeled, our simulations predict that DA phasic responses at the time of cue presentation will show similar characteristics to DA phasic responses at the time of the CS in the partial-reinforcement Pavlovian conditioning task run by Fiorillo et al. (2003). In other words, the simulation results predict that the Fiorillo et al. (2003) results can be extended to the cognitive domain involving conditioned stimuli associated with WM. It should be noted, however, that we have not simulated the partial reinforcement extinction effect, and our model in its present form, does not account for this result. Our simulations also suggest that the results of Tobler et al. (2005) for a Pavlovian task may also apply to a DRT that controls for reward magnitudes. These predictions can, in principle, be tested by recording from DA neurons while a monkey performs a DRT in which the partial reinforcement paradigm is used, or different reward magnitudes are delivered. Our model also suggests that the strength of long-term potentiation in the corticostriatal pathway is positively correlated with the size of the DA phasic signal.

4.3 Limitations and Future Work

One limitation of the model concerns the level of detail included in the actor-critic architecture. Although it is biologically inspired, it is not biologically detailed. Our model, for example, does not incorporate the BG indirect pathway (Frank, 2005; O'Reilly and Frank, 2006). Future work will attempt to incorporate the role of the BG indirect pathway in both motor inhibition and WM processes. Also, there is evidence that the DA mesocortical pathway subserves WM processes (Goldman-Rakic et al., 2000). Our model did not incorporate this pathway.

Further, by simultaneously recording from both the caudate nucleus and DLPFC of the monkey while it learned to perform a DRT, Pasupathy and Miller (2005) found that these structures showed different temporal activations. Both the striatum and DLPFC showed increased activation while learning to perform the task, but the increase in activation in the striatum preceded that in DLPFC. The authors argued that learning in the cortex is dependent on and follows learning in the BG. Our model does not account for these findings for no learning takes place in the cortex in our model.

Another limitation of our model is that, in its current form, it simulates gating of only one cue into WM. However, we believe that an extended architecture can be used to simulate gating of more than one cue into WM, such as is needed to perform the 1-2-AX task (Frank et al., 2001; O'Reilly and Frank, 2006). Future research will address the relationship of time shifting of the DA phasic signal and performance in the 1-2-AX task.

In addition to DLPFC and the striatum, delay-active neurons were also reported in the frontal eye fields, while a rhesus monkey performed an oculomotor DRT (Funahashi et al., 1989). Whether they have a different function than that of delay-active neurons in DLPFC is to our knowledge unknown. Our model did not provide an explanation for the existence of delay-active neurons in frontal eye fields.

Finally, DLPFC directly projects to motor cortex. Frank et al. (2001) assumed that this pathway subserves the selection of memory-guided responses. In our model, like the Suri and Schultz (1999) model, selection of memory guided responses is subserved by the BG. The difficulty here concerns determining the separate functional contributions of each pathway. Our model does not incorporate the DLPFC-motor cortex pathway.

References

- Alexander, G. E., DeLong, M. R., 1985. Microstimulation of the primate neostriatum. II. Somatotopic organization of striatal microexcitable zones and their relation to neuronal response properties. *Journal of Neurophysiology* 53, 1417–1430.
- Amos, A., 2000. A computational model of information processing in the frontal cortex and the basal ganglia. *Journal of Cognitive Neuroscience* 12 (3), 505–519.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of mind. *Psychological Review* 111 (4), 1036–1060.
- Anderson, S. W., Tranel, D., 2002. Neuropsychological consequences of dysfunction in human dorsolateral prefrontal cortex. In: Grafman, J. (Ed.), *Handbook of Neuropsychology*. Elsevier, New York, pp. 145–156.
- Apicella, P., Scarnati, E., Ljungberg, T., Schultz, W., 1992. Neuronal activity in monkey striatum related to the expectation of predictable environmental events. *Journal of Neurophysiology* 68, 945–960.
- Barto, A. G., 1995. Adaptive critics and the basal ganglia. In: Houk, J. C., Davis, J. L., Beiser, D. G. (Eds.), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, pp. 215–232.
- Barto, A. G., Sutton, R. S., Anderson, C. W., 1983. Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on System, Man, and Cybernetics* 13, 835–846.
- Battig, K., Rosvold, H. E., Mishkin, M., 1960. Comparison of the effects of frontal and caudate lesions on delayed response and alternation in monkeys. *Journal of Comparative and Physiological Psychology* 53, 400–404.
- Beiser, D. G., Houk, J. C., 1998. Model of cortical-basal ganglionic processing:

- Encoding the serial order of sensory events. *Journal of Neurophysiology* 79, 3168–3188.
- Bellman, R. E., 1957. *Dynamic Programming*. Princeton University Press, Princeton.
- Berns, G. S., Sejnowski, T. J., 1996. How the basal ganglia make decisions. In: Damasio, A., Damasio, H., Christen, Y. (Eds.), *The Neurobiology of Decision-making*. Springer Verlag, pp. 101–113.
- Braver, T. S., Cohen, J. D., 2000. On the control of control: The role of dopamine in regulating prefrontal function and working memory. In: Monsell, S., Driver, J. (Eds.), *Attention and Performance*. Vol. XVII. MIT Press, pp. 713–737.
- Brown, L. L., Feldman, S. M., Smith, D. M., Cavanaugh, J. R., Ackerman, R. F., Grabel, A. M., 2002. Differential metabolic activity in the striosome and matrix compartments of the rat striatum during natural behaviors. *Journal of Neuroscience* 22 (1), 305–314.
- Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, C. D., Jonides, J., et al, 1997. Temporal dynamics of brain activation during a working memory task. *Nature* 386, 604–608.
- Collins, P., Wilkinson, L. S., Everitt, B. J., Robbins, T. W., Roberts, A. C., 2000. The effect of dopamine depletion from the caudate nucleus of the common marmoset (*Callithrix jacchus*) on tests of prefrontal cognitive function. *Behavioral Neuroscience* 114, 3–17.
- Cooper, J. A., Sagar, H. J., Jordan, N., Harvey, N. S., Sullivan, E., 1991. Cognitive impairment in early untreated Parkinson’s disease and its relationship to motor disability. *Brain* 114, 2095–2122.
- Dehaene, S., Changeux, J. P., 1989. A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience* 1, 244–261.
- Diamond, A., Goldman-Rakic, P. S., 1989. Comparison of human infants and rhesus monkeys on Piaget’s A \rightarrow B task: Evidence for dependence on dorso-lateral prefrontal cortex. *Experimental Brain Research* 74, 24–40.
- Divac, J., Rosvold, H. E., Szwachart, M. K., 1967. Behavioral effects of selective ablation of the caudate nucleus. *Journal of Comparative and Physiological Psychology* 63, 184–190.
- Dunnett, S. B., Nathwani, F., Brasted, P. J., 1999. Medial prefrontal and neostriatal lesions disrupt performance in an operant delayed alternation task in rats. *Behavioural Brain Research* 106, 13–28.
- Eblen, F., Grabel, A. M., 1995. Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. *Journal of Neuroscience* 15 (9), 5999–6013.
- Fiorillo, C. D., Tobler, P. N., Schultz, W., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898–1902.
- Foster, D. J., Morris, R. G. M., Dayan, P., 2000. Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus* 10, 1–16.

- Frank, M. J., 2005. Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience* 17 (1), 51–72.
- Frank, M. J., Loughry, B., O’Reilly, R. C., 2001. Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience* 1 (2), 137–160.
- Funahashi, S., Bruce, C. J., Goldman-Rakic, P. S., 1989. Mnemonic coding of visual space in the monkey’s prefrontal cortex. *Journal of Neurophysiology* 61, 331–349.
- Fung, V. S. C., Morris, J. G. L., Leicester, J., Soo, Y. S., Davies, L., 1997. Clonic perseveration following thalamofrontal disconnection: A distinctive movement disorder. *Movement Disorder* 12, 378–385.
- Gabrieli, J., 1995. Contribution of the basal ganglia to skill learning and working memory in humans. In: Houk, J. C., Davis, J. L., Beiser, D. G. (Eds.), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, pp. 277–294.
- Goldman-Rakic, P. S., 1995. Cellular basis of working memory. *Neuron* 14, 477–485.
- Goldman-Rakic, P. S., Muly, M. C., Williams, G. V., 2000. D1 receptors in prefrontal cells and circuits. *Brain Research Reviews* 31, 295–301.
- Gurney, K., Prescott, T. J., Redgrave, P., 2001. A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics* 84 (6), 401–410.
- Hikosaka, O., Sakamoto, M., Usui, S., 1989. Functional properties of caudate neurons. III. activities related to expectation of target and reward. *Journal of Neurophysiology* 61, 814–832.
- Houk, J. C., 1995. Information processing in modular circuits linking basal ganglia and cerebral cortex. In: Houk, J. C., Davis, J. L., Beiser, D. G. (Eds.), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, pp. 3–9.
- Houk, J. C., Adams, J. L., Barto, A. G., 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, J. C., Davis, J. L., Beiser, D. G. (Eds.), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, pp. 249–270.
- Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., Grabiell, A. M., 1999. Building neural representations of habits. *Science* 286 (5445), 1745–1749.
- Kawagoe, R., Takikawa, Y., Hikosaka, O., 1998. Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience* 1, 411–416.
- Lees, A. J., Smith, E., 1983. Cognitive deficits in the early stages of Parkinson’s disease. *Brain* 106, 257–270.
- Levitt, J. J., McCarley, R. W., Dickey, C. C., Voglmaier, M. M., Niznikiewicz, M. A., Seidman, L. J., et al, 2002. MRI study of caudate nucleus volume and its cognitive correlates in neuroleptic-naive patients with schizotypal personality disorder. *American Journal of Psychiatry* 159 (7), 1190–1197.
- Levy, R., Friedman, H. R., Davachi, L., Goldman-Rakic, P. S., 1997. Differ-

- ential activation of the caudate nucleus in primates performing spatial and nonspatial working memory tasks. *Journal of Neuroscience* 17 (10), 3870–3882.
- Lewis, S. J. G., Dove, A., Robbins, T. W., Barker, R. A., Owen, M. A., 2003. Cognitive impairments in early Parkinson’s disease are accompanied by reductions in activity in frontostriatal neural circuitry. *Journal of Neuroscience* 23 (15), 6351–6356.
- Ljungberg, T., Apicella, P., Schultz, W., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology* 67, 145–163.
- Lombardi, W. J., Anderson, P. J., Sirocco, K. Y., Rio, D. E., Gross, R. E., Umhau, J. C., Hommer, D. W., 1999. Wisconsin Card Sorting Test performance following head injury: Dorsolateral fronto-striatal circuit activity predicts perseveration. *Journal of Clinical and Experimental Neuropsychology* 21 (1), 2–16.
- Middleton, F. A., Strick, P. L., 2000. Basal ganglia output and cognition: Evidence from anatomical, behavioral, and clinical studies. *Brain and Cognition* 42 (2), 183–200.
- Middleton, F. A., Strick, P. L., 2002. Basal-ganglia ‘projections’ to the prefrontal cortex of the primate. *Cerebral Cortex* 12 (9), 926–935.
- Mink, J. W., 1996. The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology* 50 (4), 381–425.
- Minsky, M. L., 1961. Steps toward artificial intelligence. In: *Proceedings of the Institute of Radio Engineers*. Vol. 49. pp. 8–30.
- Munakata, Y., Morton, J. B., Stedron, J. M., 2003. The role of prefrontal cortex in perseveration: Developmental and computational explorations. In: Quinlan, P. (Ed.), *Connectionist Models of Development*. Psychology Press, East Sussex, UK, pp. 83–114.
- O’Reilly, R. C., 1998. Six principles of biologically based computational models of cortical cognition. *Trends in Cognitive Science* 2, 455–462.
- O’Reilly, R. C., Frank, M. J., 2006. Making working memory work: A computational model of learning in the frontal cortex and basal ganglia. *Neural Computation* 18, 283–328.
- Owen, A. M., 2004. Cognitive dysfunction in Parkinson’s disease: The role of frontostriatal circuitry. *The Neuroscientist* 10, 525–537.
- Pasupathy, A., Miller, E. K., 2005. Different time courses of learning-related activity in the prefrontal cortex and basal ganglia. *Nature* 433, 873–876.
- Pearce, J. M., 1997. *Animal Learning and Cognition*. Psychology Press, East Sussex, UK.
- Pickett, E. R., Kuniholm, E., Protopapas, A., Friedman, J., Lieberman, P., 1998. Selective speech motor, syntax, and cognitive deficits associated with bilateral damage to the putamen and head of the caudate nucleus: A case study. *Neuropsychologia* 2, 173–188.
- Prescott, T. J., Gurney, K., Redgrave, P., 2003. Basal ganglia. In: Arbib, M. A. (Ed.), *The Handbook of Brain Theory and Neural Networks*, 2nd Edition.

- MIT Press, Cambridge, MA, pp. 147–151.
- Redgrave, P., Prescott, T. J., Gurney, K., 1999. The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience* 89, 1009–1023.
- Rescorla, R. A., Wagner, A. R., 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A. H., Prokasy, W. F. (Eds.), *Classical Conditioning: II. Current Research and Theory*. Appleton-Century-Crofts, New York, pp. 64–99.
- Reynolds, J. N. J., Wickens, J. R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks* 15 (4-6), 507–521.
- Sawaguchi, T., Iba, M., 2001. Prefrontal cortical representation of visuospatial working memory in monkeys examined by local inactivation with muscimol. *Journal of Neurophysiology* 86, 2041–2053.
- Schultz, W., 1999. The reward signal of midbrain dopamine neurons. *News in Physiological Science* 14 (6), 249–255.
- Schultz, W., Apicella, P., Ljungberg, T., 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience* 13, 900–913.
- Schultz, W., Dayan, P., Montague, P. R., 1997. A neural substrate for prediction and reward. *Science* 275, 1593–1599.
- Sun, R., Merrill, E., Peterson, T., 2001. From implicit skills to explicit knowledge: A bottup-up model of skill learning. *Cognitive Science* 25, 203–244.
- Sun, R., Slusarz, P., Terry, C., 2005. The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review* 112 (1), 159–192.
- Suri, R. E., Schultz, W., 1998. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research* 121, 350–354.
- Suri, R. E., Schultz, W., 1999. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* 91 (3), 871–890.
- Sutton, R. S., 1988. Learning to predict by the methods of temporal difference. *Machine Learning* 3, 9–44.
- Sutton, R. S., Barto, A. G., 1987. A temporal-difference model of classical conditioning. In: *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*. pp. 355–378.
- Sutton, R. S., Barto, A. G., 1990. Time-derivative models of Pavlovian reinforcement. In: Gabriel, M., Moore, J. (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. MIT Press, Cambridge, MA, pp. 497–537.
- Sutton, R. S., Barto, A. G., 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Taylor, A. E., Saint-Cyr, J. A., Lang, A. E., 1986. Frontal lobe dysfunction in Parkinson’s disease. The cortical focus of neostriatal outflow. *Brain* 109, 845–883.
- Thorndike, E. L., 1911. *Animal Intelligence*. Hafner, Darien, CT.

- Tobler, P. N., Fiorillo, C. D., Schultz, W., 2005. Adaptive coding of reward value by dopamine neurons. *Science* 307, 1642–1645.
- Vermersch, A. I., Graymard, B. M., Rivaud-Pechoux, S., Ploner, C. J., Agid, Y., Pierrot-Deseilligny, C., 1999. Memory guided saccade deficit after caudate nucleus lesion. *Journal of Neurology, Neurosurgery and Psychiatry* 66 (4), 524–527.
- Waelti, P., Dickinson, A., Schultz, W., 2001. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412, 43–48.
- Wickens, J., 1997. Basal ganglia: Structure and computations. *Network: Computation in Neural Systems* 8 (4), R77–R109.
- Wickens, J. R., Begg, A. J., Arbuthnott, G. W., 1996. Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neuroscience* 70, 1–5.
- Wilson, C. J., 2004. Basal ganglia. In: Sheperd, G. M. (Ed.), *The Synaptic Organization of the Brain*. Oxford University Press, New York, pp. 361–413.
- Witten, I. H., 1977. An adaptive optimal controller for discrete-time Markov environments. *Information and Control* 34, 286–295.