ARTICLE **Communicated by Joshua Gold**

Integration of Reinforcement Learning and Optimal Decision-Making Theories of the Basal Ganglia

Rafal Bogacz
R.Bogacz@bristol.ac.uk
Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K.

Tobias Larsen
larsent@tcd.ie
Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K., and Trinity College Institute of Neuroscience, Dublin 2, Ireland

This article seeks to integrate two sets of theories describing action selection in the basal ganglia: reinforcement learning theories describing learning which actions to select to maximize reward and decision-making theories proposing that the basal ganglia selects actions on the basis of sensory evidence accumulated in the cortex. In particular, we present a model that integrates the actor-critic model of reinforcement learning and a model assuming that the cortico-basal-ganglia circuit implements a statistically optimal decision-making procedure. The values of corico-striatal weights required for optimal decision making in our model differ from those provided by standard reinforcement learning models. Nevertheless, we show that an actor-critic model converges to the weights required for optimal decision making when biologically realistic limits on synaptic weights are introduced. We also describe the model's predictions concerning reaction times and neural responses during learning, and we discuss directions required for further integration of reinforcement learning and optimal decision-making theories.

1 Introduction

The basal ganglia are a set of subcortical nuclei critically involved in action selection. This article seeks to integrate two sets of theories concerning action selection in the basal ganglia: reinforcement learning (RL) and optimal decision-making (DM) theories. The RL theories describe the process of learning which action to select for a given stimulus to maximize the reward (Frank, Seeberger, & O'Reilly, 2004; Montague, Dayan, & Sejnowski, 1996; O'Doherty et al., 2004; Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1998). By contrast the DM theories assume that the animal has acquired a stimulus-response mapping, and they describe the process of selecting an action corresponding to the stimulus most supported by incoming sensory

evidence (Brown, Bullock, & Grossberg, 2004; Frank, 2006; Gurney, Prescott, & Redgrave, 2001; Humphries, Stewart, & Gurney, 2006; Lo & Wang, 2006; Redgrave, Prescott, & Gurney, 1999). The DM theories further assume that the stimulus and its neural representation are noisy, and they describe the process of integration of sensory evidence over time, until a criterion of confidence in stimulus identity is met. It has recently been proposed that the cortico-basal-ganglia circuit selects actions on the basis on noisy inputs in a statistically optimal way, thereby minimizing decision time (Bogacz & Gurney, 2007), and we refer to this theory as the optimal DM theory. Thus, the RL and DM theories concern two types of uncertainty that need to be dealt with during choice: the RL theories assume that the animal is uncertain of the expected rewards for selecting different actions, while the DM theories assume that initially after the stimulus onset, the animal is uncertain of the stimulus identity.

The RL and optimal DM theories have been used to describe both neurophysiologic and behavioral data. The RL theories have been used to explain firing properties of dopaminergic neurons (Kakade & Dayan, 2002; Schultz et al., 1997; Tobler, Fiorillo, & Schultz, 2005; Ungless, Magill, & Bolam, 2004) and counterintuitive patterns of choices that subjects make in learning tasks (Bogacz, McClure, Li, Cohen, & Montague, 2007; Frank et al., 2004). The optimal DM theory describes firing properties of neurons in the basal ganglia (Bogacz & Gurney, 2007) and distributions of reaction times due to its equivalence with the diffusion model (Ratcliff, 2006; Ratcliff, Gomez, & McKoon, 2004; Ratcliff & Smith, 2004).

So far, the two sets of theories have mostly been used to address the data from different phases of task acquisition. The RL theories focus on the learning phase when the animal has to discover the stimulus-response mapping, and they assume that this mapping is initially learned in the basal ganglia (Atallah, Lopez-Paniagua, Rudy, & O'Reilly, 2007; Frank et al., 2004; O'Doherty et al., 2004; Samejima, Ueda, Doya, & Kimura, 2005). By contrast, the DM theories often focus on the proficient phase when the stimulus-response mapping has been kept constant for long periods, and they assume that the mapping is stored in the cortex (Bogacz & Gurney, 2007; Gurney et al., 2001). (A similar assumption is made by models describing DM processes in the cortex: Mazurek, Roitman, Ditterich, & Shadlen, 2003; Shadlen & Newsome, 2001; Usher & McClelland, 2001; Wang, 2002.)

RL and DM theories describe different aspects of information processing in the cortico-basal-ganglia circuit, but it has not been investigated before if this circuit can simultaneously implement RL and optimal DM. This question is addressed in this article which is organized as follows. Section 2 reviews relevant elements of RL and optimal DM theories. Then section 3 presents a new model integrating RL and optimal DM, and section 4 shows results of its simulations. Finally, section 5 discusses predictions of the model and further research required to integrate RL and optimal DM theories.

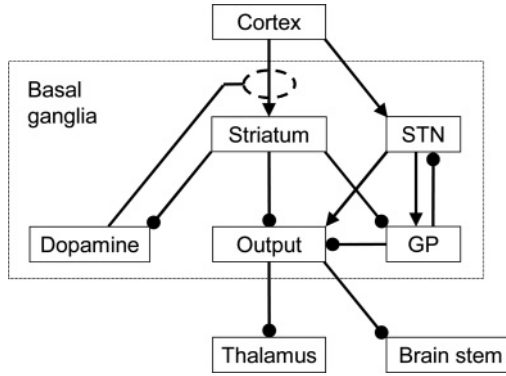


Figure 1: Main connections of the basal ganglia, based on Gurney, Prescott, and Redgrave (2001). Boxes denote brain areas: Output—output nuclei: substantia nigra pars reticulate and entopeduncular nucleus (or its homologue, GPi in primates); STN—subthalamic nucleus, GP—globus pallidus (or its homologue GPe in primates); and dopamine–areas releasing dopamine (substantia nigra pars compacta and ventral tegmental area). Nuclei included in the basal ganglia are within the dotted rectangle. Arrows denote excitatory connections, lines ending with circles denote inhibitory connections, and lines ending with a dashed circle denote modulatory dopaminergic projection.

2 Review of RL and Optimal DM theories

This section briefly reviews relevant aspects of basal ganglia anatomy, the optimal DM model[1] and the actor-critic model of RL.

2.1 Functional Anatomy of the Basal Ganglia. The main connections of the basal ganglia are shown in Figure 1. Different neurons within the basal nuclei are selective for different movements; hence, it has been proposed that the basal ganglia are divided into channels corresponding to individual actions that traverse all nuclei (Alexander, DeLong, & Strick, 1986). The connections between the nuclei are predominantly within a channel (e.g., striatal neurons selective for right-hand movement project to output neurons selective for right-hand movement), with an exception of connections from the subthalamic nucleus (STN), which are more diffused (Parent & Smith, 1987).

[1]The model reviewed here differs from that described by Bogacz and Gurney (2007) in that the basal ganglia sends feedback to the cortex as described by Bogacz (2009). Also, the model is introduced in a novel, more intuitive way (Bayes' theorem is explicitly mapped on the cortico-basal-ganglia circuit).

In a default state, the output nuclei send tonic inhibition to the thalamus and brain stem, blocking cortical control over muscles. An action is selected when the inhibition from the corresponding channel in the output nuclei is lowered (Chevalier, Vacher, Deniau, & Desban, 1985; Deniau & Chevalier, 1985). This may happen when the striatal neurons in the corresponding channel become sufficiently active to block the output neurons.

All basal nuclei receive input from dopaminergic neurons, which modulate the activity and synaptic plasticity in the basal ganglia. The strongest dopaminergic input is provided to the striatum.

2.2 Optimal DM in the Cortico-Basal-Ganglia Circuit. In this section, we describe relevant experimental studies of neural bases of DM, a statistically optimal DM procedure, and its implementation in the model of the cortico-basal-ganglia-thalamic circuit (Bogacz, 2009; Bogacz & Gurney, 2007).

2.2.1 Decision Mechanisms in the Cortex. Neural bases of DM are typically studied in a task in which a monkey is presented with a stimulus consisting of moving dots. The majority of dots are moving randomly, while a certain proportion is moving coherently left or right. The subject is required to identify the direction of the coherent motion and make an eye movement in this direction in order to receive a reward (Britten, Shadlen, Newsome, & Movshon, 1993). Animals are typically performing this task for several months before neural responses are studied.

Single-cell recordings suggest that in this task, the neurons in the medial temporal (MT) area, which are involved in motion processing, provide sensory input that depends on the stimulus presented. The MT neurons have a preferred direction of motion, and their firing rate is proportional to the magnitude of motion in their preferred direction (Britten et al., 1993). If the coherent fraction of dots is moving left, then on average, the activity of MT neurons selective for leftward motion is higher than of the neurons selective for rightward motion. Thus, the decision problem faced by brain regions "listening to MT" may be formulated as identifying which population of sensory neurons has the highest mean (Gold & Shadlen, 2001, 2002). However, this identification is not trivial because the responses of sensory neurons are noisy (due to noise present in the stimulus and neural representation); hence, they need to be "observed" for a period of time before an accurate decision may be made.

Indeed, the neural correlates of information accumulation in this paradigm have been observed in neurons in the lateral intraparietal (LIP) area and the frontal eye field (FEF). These neurons respond selectively before and during saccades in their preferred direction. During the motion discrimination task, as the monkey's confidence in one of the responses grows, the neurons selective for this response gradually increase their firing rate (Kiani & Shadlen, 2009; Roitman & Shadlen, 2002; Schall, 2001;

Shadlen & Newsome, 2001). In this article, we refer to these neurons as integrators.

2.2.2 DM Procedure. An optimal procedure for making a choice between two alternatives on the basis of noisy sequentially incoming data is provided by the sequential probability ratio Test (SPRT) (Wald, 1947). The SPRT minimizes decision time for any required accuracy (Wald & Wolfowitz, 1948). Several studies suggested how SPRT could be implemented in simple networks in the cortex (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Gold & Shadlen, 2002, 2007). A generalization of SPRT to multiple alternatives has been developed. Called the multihypothesis SPRT (MSPRT) (Baum & Veeravalli, 1994), it has been shown analytically to be asymptotically optimal, that is, to minimize the decision time for required accuracy when the required accuracy converges to 100% (Dragalin, Tertakovsky, & Veeravalli, 1999). In simulations, it appears to make as fast or faster choices than any other well-known algorithm when the required accuracy is lower (McMillen & Holmes, 2006). For example, in simulations with the level of noise in sensory input estimated from experimental data, the MSPRT achieved approximately 20% faster decision times with accuracy at 99% in choosing among 10 alternatives than simpler models did (race and leaky competing accumulator models) (Bogacz & Gurney, 2007). In this article, we conjecture that MSPRT has the same optimality property as SPRT and also refer to it as optimal.

The MSPRT is a statistical test between N hypotheses, which we now define for the context of choice based on sensory input. Assume that time is divided into discrete intervals, and let $x_i(t)$ denote the total number of spikes produced by the population of sensory neurons selective for alternative i during interval t. Let hypothesis H_i correspond to alternative i being correct, so let H_i state that sensory input $x_i(t)$ has the highest mean. To define the hypotheses precisely, one needs to assume the probability distribution of $x_i(t)$. For example, Bogacz and Gurney (2007) assumed that the input $x_i(t)$ produced by sensory population i can be approximated by a gaussian distribution[2] with mean I_i and variance σ^2, and defined hypotheses H_i: $I_i = I_+$, $I_{j\neq i} = I_-$, where I_+ and I_- are the mean numbers of spikes per interval produced by sensory neurons for preferred and nonpreferred stimuli, respectively.

Let $x(t)$ denote the total sensory input during interval t (thus, $x(t)$ is a vector $[x_1(t), \ldots, x_N(t)]$). According to the MSPRT, at each moment of time

[2]Zhang and Bogacz (2010) showed that the model of basal ganglia also performs MSPRT for a more realistic assumption that spikes produced by the sensory neurons can be described by Poisson processes. However, for simplicity of explanation, in this article, we assume that the inputs can be approximated by gaussian distributions.

and for each alternative, one computes the probability of this alternative being correct given the sensory input so far, which we denote by $P_i(t)$:

$$P_i(t) = P(H_i|x(1, \ldots, t)). \tag{2.1}$$

According to the MSPRT the choice should be made when any of $P_i(t)$ exceeds a fixed threshold; otherwise, the decision process should continue.

Let us consider how $P_i(t)$ can be calculated. Let us assume that at the start of each trial, we do not have any prior expectations about the direction of movement, so our initial estimates of the probabilities of left and right are equal to $P_i(0) = 1/N$. Let us further assume that $x(t)$ are statistically independent across different time intervals t. After each time interval during which we observe the sensory input $x(t)$, the probabilities can be updated according to Bayes' theorem:

$$P_i(t) = \frac{P_i(t-1)P(x(t)|H_i)}{P(x(t))}. \tag{2.2}$$

According to equation 2.2, the prior estimates $P_i(t–1)$ are updated by multiplying them by the probability densities of observing the sensory input given the corresponding hypotheses. Additionally, to ensure that the updated probabilities $P_i(t)$ add up to 1, the product in equation 2.2 is divided by a normalization term $P(x(t))$ that is equal to the following probability density:

$$P(x(t)) = \sum_{i=1}^{N} P_i(t-1)P(x(t)|H_i). \tag{2.3}$$

To illustrate how the computation of the posterior probabilities could be performed in a neural circuit, let us consider an example of the moving dots task with two alternatives: left (L) and right (R). Figure 2a represents equation 2.2 as a network of computational elements. In order to calculate the posterior probabilities $P_L(t)$ and $P_R(t)$, the prior probabilities and the likelihoods of the observed sensory input given the two hypotheses need to be multiplied (top circles in Figure 2a), and these products need to be divided by the normalization term. The normalization term is simply the sum of these products (computed in the right circle in Figure 2a). The posterior probabilities become the priors for the next time interval; hence, they need to be fed back with a time delay (arrows connecting bottom circles to top circles in Figure 2a).

The network in Figure 2a involves multiplication and division, which are difficult to compute by neurons. This problem can be solved by applying a logarithm to all terms computed in the network as shown in Figure 2b. The logarithm changes multiplication into addition and division into

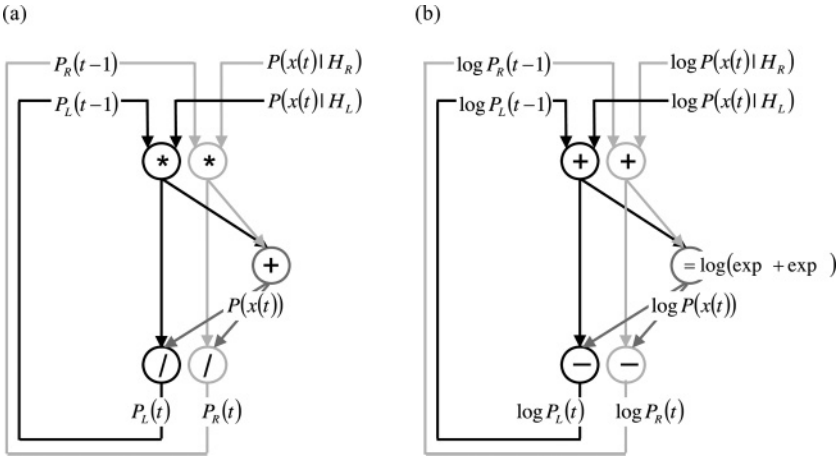


Figure 2: Representation of the computation of the posterior probabilities of dots moving left and right (a) and their logarithms (b). Circles denote mathematical operations, and arrows denote the flow of information. Black and gray pathways show computations of the posterior probabilities of dots moving left and right, respectively.

subtraction. The computation of the normalization term becomes slightly more complex, as the inputs to the node computing it (the right circle in Figure 2b) need to be exponentiated before addition, and a logarithm of the resulting sum has to be taken.

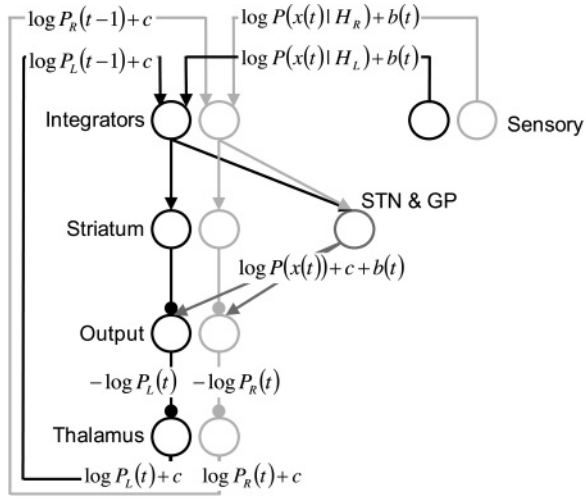
The application of logarithm brings another benefit: it has been shown that $\log P(x(t)\,|\,H_i)$ can be decomposed (for commonly used assumptions about distribution of $x_i(t)$) (Bogacz & Gurney, 2007; Gold & Shadlen, 2001, 2002; Zhang & Bogacz, 2010):

$$\log P(x(t)|H_i) = g x_i(t) - b(t), \tag{2.4}$$

where g is a constant (for the hypotheses defined above, $g = (I_+ – I_-)/\sigma^2$; Bogacz & Gurney, 2007) and $b(t)$ has the same value for all i. Thanks to equation 2.4, the information about $\log P(x(t)\,|\,H_i)$ can be encoded in the firing rate of sensory neurons (as we show below).

2.2.3 Neural Implementation. The network of Figure 2b can be mapped on the known anatomy of the cortico-basal ganglia-thalamic circuit, as shown in Figure 3a. Before describing the details of the model, we provide an overview of probabilistic quantities computed in the model (indicated in labels in Figure 3a; activations $b(t)$ and c in the labels will be discussed later).

(a)



(b)

Region	Equation
Cortical integrators	$INT_i(0) = \log 1/N + c$ $INT_i(t) = TH_i(t-1) + g\, x_i(t)$, for $t>0$
Striatum	$STR_i(t) = INT_i(t)$
Input from STN and GP	$SG(t) = \log \sum_{i=1}^{N} \exp INT_i(t)$
Output nuclei	$OUT_i(t) = -STR_i(t) + SG(t)$
Thalamus	$TH_i(t) = c - OUT_i(t)$

Figure 3: A subset of cortico-basal-ganglia-thalamic circuit required to implement MSPRT. (a) Architecture of the model. Pairs of circles correspond to brain areas: sensory—sensory cortex encoding relevant aspects of stimuli (e.g., MT in motion discrimination task); integrators—cortical region integrating sensory evidence. The circle labeled "STN & GP" denotes a circuit composed of subthalamic nucleus and globus pallidus. Arrows denote excitatory connections, and lines ending with circles denote inhibitory connections. Black and gray pathways correspond to two sample channels. (b) Equations describing the model.

The cortical integrators[3] sum the inputs encoding the logarithm of the prior probability provided by a feedback from the thalamus and $\log P(x(t)\,|\,H_i)$ provided by the input from the sensory neurons. The logarithm of the normalization term is computed in the model by a network of two nuclei: STN

[3]The integrators in the model correspond to neurons integrating sensory inputs (see section 2.2.1). They may correspond to neurons located in the frontal eye field (in tasks with saccadic response), or premotor cortex (in tasks with motor response), as it is known that these areas project to basal ganglia. However, the exact roles of each of these cortical areas in decision formation are currently unknown.

and globus pallidus (GP) (as described below). The output nuclei compute the difference between the logarithm of the normalization term and the input from the integrators, as they receive an inhibition from the integrators via the striatum. Hence, the activity of the neurons in the output nuclei is proportional to –log $P_i(t)$. But the output nuclei send inhibition to the thalamus, thus, the activities of the thalamic neurons are proportional to log $P_i(t)$. These posterior probabilities are feedback to the cortical integrators.

Figure 3b lists equations describing the model. In the model, the circuit of STN and GP computes $SG(t)$ defined in Figure 3b as a nonlinear summation of its inputs from cortical integrators. Bogacz and Gurney (2007) showed that $SG(t)$ can be computed in a model of the STN-GP circuit if the STN and GP neurons have particular input-output transfer functions. These input-output relationships required for computation of $SG(t)$ agree with those observed in in vivo studies of STN and GP neurons (Hallworth, Wilson, & Bevan, 2003; Nambu & Llinas, 1994; Wilson, Weyrick, Terman, Hallworth, & Bevan, 2004).

The computations in the model of Figure 3 are analogous to those in the network of Figure 2b, but additional excitatory inputs are included that ensure that neural activities in the model are not negative. Note that some of the values in the labels in Figure 2b are negative, because a logarithm of a probability is negative (as a probability is lower than 1), but the firing rates cannot be negative. Thus, in the model, the activities of cortical integrators are initialized to the logarithms of prior probabilities of alternatives increased by a constant c (see Figure 3b). Similarly, the thalamus also receives an excitatory input equal to c. The integrators in the model receive input from sensory neurons equal to $g\, x_i(t)$, which according to equation 2.4 is equal to $\log P(x(t) \mid H_i) + b(t)$.

We now show that adding the above excitatory inputs ($b(t)$ and c) does not affect the activity of the output nuclei in the model. In particular, we show that for the model defined in Figure 3b, the activity of the output nuclei is

$$OUT_i(t) = -\log P_i(t). \tag{2.5}$$

We show that equation 2.5 holds using mathematical induction. We consider the values of $OUT_i(t)$ at different time steps (i.e., we perform induction on t). In the first step, we show that equation 2.5 holds for $t = 0$. According to Figure 3b the input provided by STN and GP at $t = 0$ is

$$SG(0) = \log \sum_{i=1}^{N} \exp\left(\log \frac{1}{N} + c\right) = \log\left(\left(\sum_{i=1}^{N} \exp \log \frac{1}{N}\right) \exp c\right) = c. \tag{2.6}$$

Hence, according to Figure 3b, the activity of output nuclei is

$$OUT_i(0) = -INT_i(0) + SG(0) = -\log 1/N - c + c. \tag{2.7}$$

Note that c cancels in equation 2.7, which illustrates the property of the model that any value added to the activity of all integrators does not affect the activity of output nuclei. Equation 2.7 implies that equation 2.5 holds for $t = 0$ (as we assumed that $P_i(0) = 1/N$).

Now we will show that if the inductive hypothesis of equation 2.5 is satisfied at time $t - 1$, it is also satisfied at time t (which, according to mathematical induction, will imply that equation 2.5 is satisfied for all $t \geq 1$). The activity of integrators is (from Figure 3b and equations 2.4 and 2.5; also see the labels in Figure 3a)

$$INT_i(t) = \log P_i(t-1) + c + \log P(x(t)|H_i) + b(t). \tag{2.8}$$

The input provided by STN and GP becomes (using manipulations as in equation 2.6)

$$SG(t) = \log \sum_{i=1}^{N} P_i(t) P(x(t)|H_i) + c + b(t). \tag{2.9}$$

Using equation 2.3, we get

$$SG(t) = \log P(x(t)) + c + b(t). \tag{2.10}$$

The activity of output nuclei becomes

$$\begin{aligned} OUT_i(t) = &- \log P_i(t-1) - c - \log P(x(t)|H_i) - b(t) \\ &+ \log P(x(t)) + c + b(t). \end{aligned} \tag{2.11}$$

Note that $b(t)$ and c cancel, and equation 2.11 together with Bayes' theorem (see equation 2.2), give equation 2.5, which completes the proof.

In the model, a choice is made when the activity of any channel in the output nuclei, $-\log P_i(t)$, decreases below a threshold (consistent with selection by disinhibition; see section 2.1). This is equivalent to making a choice as soon as any $P_i(t)$ exceeds a threshold; thus, the model implements MSPRT.

The integrators receive input $g\ x_i(t)$, which seems to suggest that the correct value of constant g (satisfying equation 2.4) needs to be known. However, if g is set to a value higher than that for which equation 2.4 is satisfied, the performance of the model does not decrease (shown numerically in Figure 3B in Bogacz & Gurney 2007, and justified analytically in appendix C in Bogacz & Gurney, 2007). Hence the precise value of g does not need to be known.

2.3 RL in Basal Ganglia Circuit. RL theories are typically used to describe performance in tasks in which rewarded responses for particular

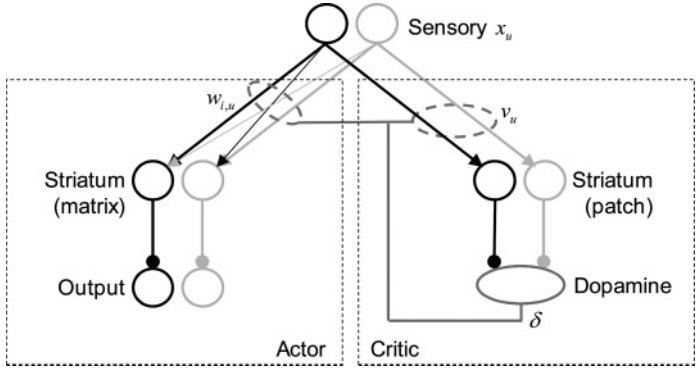


Figure 4: An actor-critic model mapped on a subset of the basal ganglia circuit. Notation as in Figure 3.

stimuli need to be learned by trial and error, but the stimulus is clearly presented and thus easy to identify. Hence, the temporal integration is not considered, and the model variables do not change within the duration of an individual selection process. Let x_u be a binary variable denoting the presence or absence of stimulus u. RL theories typically assume that x_u is computed in the sensory cortex (Frank & Claus, 2006; Frank et al., 2004). In RL models, the probabilities of selecting actions i depend on action values, which we denote by y_i. RL theories assume that y_i are computed by neurons in striatal channel i on the basis of their cortical inputs:

$$y_i = \sum_u w_{i,u} x_u, \tag{2.12}$$

where $w_{i,u}$ are the weights of connections from cortical neurons representing stimulus u to striatal neurons in channel i. The network that could support this calculation is shown in the part of Figure 4 labeled "Actor." The stratial neurons project to the output nuclei; hence, the higher the action value computed by striatal neurons in a given channel, the higher the chance is that the corresponding action will be executed. Thus in this network, the stimulus-response mapping is encoded in the weights of connections between sensory and striatal neurons. Note that we now use two different indices for stimuli and actions (u and i), because at the start of RL experiments, it is usually not known which action is correct for a given stimulus or whether this mapping can change in the course of an experiment.

After each choice, the weights $w_{i,u}$ are modified. Several learning rules have been proposed, and here we focus on an actor-critic model (Sutton & Barto, 1998), which is supported by experimental data (O'Doherty et al., 2004). This model is illustrated in Figure 4 and has two parts. The first part,

an actor, learns which action i to select for stimulus u, and it corresponds to the network discussed so far. A second additional part, a critic, learns an average expected reward associated with the stimulus alone, that is, it learns the average reward obtained on all trials when stimulus u was presented; we denote it by v_u.

Doya (2000) proposed that v_u are encoded in the synaptic weights of striatal neurons in striosomes or "patches" that project to dopaminergic areas, while $w_{i,u}$ are encoded in the synaptic weights of striatal neurons in the surrounding "matrix" that project to the output nuclei (Gerfen, 1992). The patches are more common in the ventral striatum (Gerfen, 1992) that has been proposed to be involved in the computations of the critic (O'Doherty et al., 2004), while the matrix is more prevalent in the dorsal striatum associated with the actor.

Much evidence suggests that the firing rate of dopaminergic neurons represents the reward prediction error δ, defined as the difference between obtained reward (r) and expected reward (Montague et al., 1996; Schultz et al., 1997; Tobler et al., 2005). Doya (2000) proposed that the dopaminergic neurons can compute δ as they receive excitation from areas encoding obtained reward and inhibition from striatal patch neurons. Thus, the firing rate of the dopaminergic neurons encodes

$$\delta = r - v_{presented}. \tag{2.13}$$

The dopaminergic neurons send projections to striatum that modulate synaptic plasticity of cortico-striatal connections (Reynolds, Hyland, & Wickens, 2001). Hence, the expected reward for the stimulus presented, encoded in the critic, is modified proportionally to δ:

$$v_{presented} \leftarrow v_{presented} + \eta\delta. \tag{2.14}$$

In equation 2.14, η denotes the learning rate. The weights in the actor $w_{i,u}$ are also modified proportionally to δ, and a few versions of the actor learning rule have been proposed (Sutton & Barto, 1998). For example, in one of the versions, the weights between sensory neurons encoding presented stimulus and striatal neurons in a channel corresponding to the chosen action are updated according to Sutton and Barto (1998):

$$w_{chosen,presented} \leftarrow w_{chosen,presented} + \eta\delta. \tag{2.15}$$

2.4 What Do the RL and DM Models Optimize? Both RL and DM models optimize criteria connected with reward. RL models learn which actions to select to maximize the expected reward per choice in a learning task. The optimal DM model maximizes the reward rate, defined as the

average reward per unit of time. The particular expression for the reward rate depends on the task. Gold and Shadlen (2002) consider a task in which a subject receives a unit of reward for correct choices, and there is a delay D between the response and the onset of the next trial. In this task, the expected reward rate is equal to the ratio of a probability of making a correct choice and the average duration of a trial:

$$RR = \frac{Accuracy}{RT + D}. \tag{2.16}$$

In this task, the average duration of a trial depends on reaction time RT; hence, to maximize reward rate, the choices also need to be fast. The optimal DM model maximizes the reward rate in a wide range of sequential choice tasks (including the above) when the threshold parameter, controlling the speed-accuracy trade-off, is chosen optimally (Bogacz, 2009; Bogacz et al., 2006). This analysis suggests that to maximize the reward rate in learning tasks, both RL and optimal DM need to be employed.

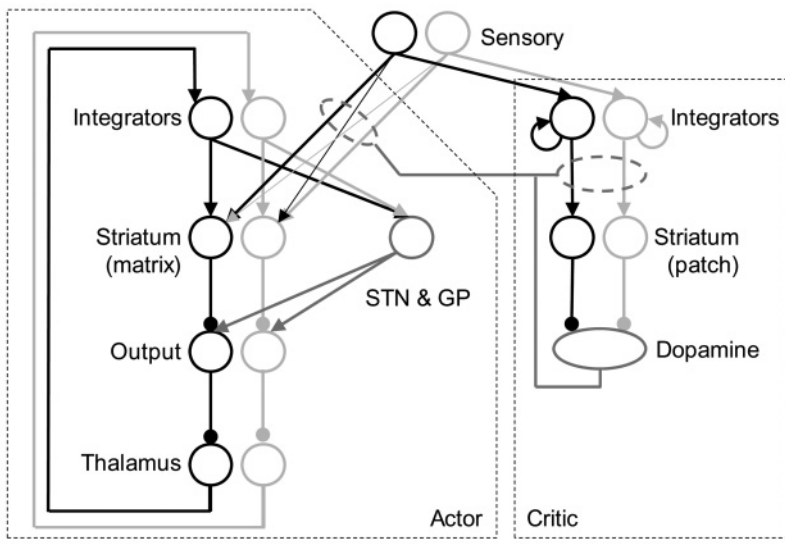
3 Integrated Model

3.1 Overview of the Model. We now present a model of the cortico-basal-ganglia circuit that performs DM approximating MSPRT on the basis of stimulus-response mapping learned with RL. Figure 5a shows the architecture of the model, which combines the models of Figures 3 and 4. The model consists of two parts: an actor and a critic. The actor selects an action on the basis of the noisy sensory input, which it integrates, and the weights of striatal matrix neurons, which encode the stimulus-response mapping. The critic computes the expected reward associated with the stimulus on the basis of the noisy sensory input that is also integrated by the critic, and the weights of striatal patch neurons. The dopaminergic neurons compute the reward prediction error on the basis of the expected reward found by the critic and control the modifications of all striatal weights.

Figure 5b shows equations describing the model; they use a notation combining notations from sections 2.2 and 2.3 and denote the input from the sensory neurons selective for stimulus u at time t by $x_u(t)$. The actor in this integrated model is formed by a modification of the optimal DM model (see Figure 3) in which the sensory neurons project to the striatum rather than cortical integrators. According to Figure 5b, the activity of striatal matrix neurons associated with particular action depends on weights $w_{i,u}$.

The critic also includes cortical integrators that accumulate sensory information on stimulus identity. For simplicity, we consider a very basic model of this integration (see Figure 5b), because in our simulations, it is only

(a)



(b)

Part	Region	Equation
Actor	Cortical integrators	$INT_i(0)$ are based on rewards (see Section 3.2) $INT_i(t) = TH_i(t-1)$, for $t>0$
	Striatum (matrix)	$STR_i(t) = INT_i(t) + g\sum_u w_{i,u} x_u(t)$
	Input from STN and GP	$SG(t) = \log \sum_{i=1}^{N} \exp INT_i(t)$
	Output nuclei	$OUT_i(t) = -STR_i(t) + SG(t)$
	Thalamus	$TH_i(t) = c - OUT_i(t)$
Critic	Cortical integrators	$INT_u^{critic}(0) = 0$ $INT_u^{critic}(t) = INT_u^{critic}(t-1) + x_u(t)$, for $t>0$
	Striatum (patch)	$STR_u^{critic}(t) = v_u INT_u^{critic}(t)$

Figure 5: The integrated model. (a) Architecture of the model. Notation as in Figure 3. (b) Equations describing the model.

important which $INT_u{}^{critic}$ is the highest (as explained later), and extending the model of integration (e.g., adding inhibitory connections or integrating via cortico-basal-ganglia-thalamic loops) would not change which $INT_u{}^{critic}$ is the highest.

The integration in the actor is performed until the OUT_i decreases below some threshold indicating that the decision has been made with the required precision. At this time, the integration in the critic is stopped as well, and the stimulus with the highest INT_u^{critic} is selected as the best guess

for the stimulus presented. In the model, this is represented by setting INT_u^{critic} to[4]

$$INT_u^{critic} = \begin{cases} 1 \text{ if } INT_u^{critic} = \max_u(INT_u^{critic}) \\ 0 \text{ otherwise.} \end{cases} \tag{3.1}$$

Because the connections between critic integrators and the striatal patch neurons are weighted according to the expected reward for each stimulus (see Figure 5b), the activity of the striatal patch neurons selective for the selected stimulus u is equal to

$$STR_u^{critic} = v_u INT_u^{critic} = v_u. \tag{3.2}$$

The striatum then projects the information about the expected reward v_u via inhibitory connections to the dopaminergic neurons, which will also receive excitatory input indicating the reward given by the action taken, and through the combination of the two inputs calculate prediction error δ as in equation 2.13. The prediction error, together with the learning rate, determines how the weights are modified according to equations 2.14 and 2.15.

For this model to approximate optimal DM, $w_{i,u}$ need to have different values from those provided by standard RL models, as we describe in the section 3.2. Nevertheless, we show in section 3.3 that desired $w_{i,u}$ can be learned when biologically realistic limits on synaptic weights are introduced, and in section 3.4 that with such weights, the proposed model approximates MSPRT.

3.2 Striatal Weights Required for Optimal DM. The striatal weights required for optimal DM differ from those provided by standard RL models in tasks in which some actions may be more rewarded. Let $r_{i,u}$ denote the reward delivered after choosing action i for stimulus u. An example of a task with unequally rewarded actions is a modification of a motion discrimination task in which a subject receives two drops of juice for leftward saccade when dots are moving left ($r_{1,1} = 2$), one drop of juice for rightward saccade when dots are moving right ($r_{2,2} = 1$), and no juice for saccades opposite to the direction of motion ($r_{1,2} = r_{2,1} = 0$). We refer to this as the biased reward task.

[4]We do not model equation 3.2 explicitly, but we note that setting the winning INT_u^{critic} to the ceiled value of 1 could result from a transient increase in gain of integrators, as it has been proposed that such gain increase may occur at the moment of decision (Shea-Brown, Gilzenrat, & Cohen, 2008). Conversely, setting the other INT_u^{critic} to 0 can be a result of inhibition from the winning integrator.

For such tasks, in standard RL models, $r_{i,u}$ determine striatal weights such that the higher $r_{i,u}$ is, the higher $w_{i,u}$ is. However, to maximize the probability of reward in a given trial of the biased reward task, one needs to choose an action that is most rewarded for the presented stimulus, and thus to implement optimal DM, the striatal weights should be equal to

$$w_{i,u} = \begin{cases} 1 \text{ if action } i \text{ is the most rewarded action for stimulus } u \\ 0 \text{ otherwise.} \end{cases} \tag{3.3}$$

Such weights allow choosing the most rewarded action for the noisy stimulus, while the weights provided by standard RL models may not allow the most rewarded choice, as we show with an example. Let us denote the mean input from sensory neurons selective for stimulus u by I_u. If the dots moved right with low coherence, so that the mean input from right sensory neurons (e.g., $I_2 = 4$) was just slightly higher than the left sensory neurons (e.g., $I_1 = 3$), then with weights of equation 3.3, the average input received by the right channel ($w_{2,2}I_2 = I_2 = 4$) will be higher than the average input received by the left channel ($w_{1,1}I_1 = I_1 = 3$). Thus, after long enough integration, INT_2 will eventually become higher than INT_1, and the more rewarded action of the right saccade will be chosen. Note that this may not happen if the striatal weights were set according to classical RL models (e.g., if $w_{i,u} = r_{i,u}$), because then the mean input to the left channel ($w_{1,1}I_1 = 6$) may be higher than the mean input to the right channel ($w_{2,2}I_2 = 4$) and the unrewarded left saccade is most likely to be chosen.

The above argument shows that with weights of equation 3.3, the most rewarded action can be chosen if enough time is allowed. However, to maximize the reward rate defined in section 2.4, it is often necessary to trade accuracy for speed. If the decision time is limited, the presented stimulus may not be identified with 100% accuracy, and reward magnitudes ($r_{i,u}$) may provide useful information in guiding the choice. The reward magnitudes can be estimated from previous trials; thus, they provide information prior to the stimulus in the current trial, so they should influence a prior probability of selecting the alternatives. As explained in section 2.2.3, the prior probabilities determine the initial values of the integrators. Bogacz et al. (2006) have shown that to maximize the reward rate in the biased reward task, the initial values of the integrators $INT_i(0)$ should be modified proportionally to the logarithms of the reward for corresponding actions:

$$INT_i(0) = \alpha \log r_{i,i} + c, \tag{3.4}$$

where α is a proportionality constant that depends on task parameters (see Bogacz et al., 2006, for details). According to equation 3.4, the initial value of the integrator corresponding to a more rewarded response needs to be higher than the one corresponding to a less rewarded response (thus, less

sensory input is required to trigger the more rewarded response). Such shifts in $INT_i(0)$ are consistent with observed increases in the activity of neurons in LIP (Platt & Glimcher, 1999) and premotor cortex (Roesch & Olson, 2004) selective for more rewarded action before stimulus onset.

With weights of equation 3.3, and initial values of integrators depending on reward magnitudes, a subject can correctly estimate which action is more likely to be rewarded throughout the whole decision process (and hence can choose the most rewarded actions at different speed-accuracy trade-offs). In particular, at the beginning of the decision process, when little sensory evidence has been provided, the activities of integrators depend mostly on the prior reward expectations. And later, during the decision process, the relative magnitudes of integrators depend to a greater and greater extent on the accumulated sensory input and to a lesser and lesser extent on the prior reward expectations.

To implement optimal DM in tasks with biased rewards, the magnitudes of rewards need to be learned and maintained somewhere to influence $INT_i(0)$. It has been recently shown that $r_{i,u}$ can be learned even when stimulus uncertainty is present (Larsen, Leslie, Collins, & Bogacz, 2010). With stimulus uncertainty, it is not obvious to which stimulus the obtained reward should be attributed. But Larsen et al. (2010) showed that a model in which estimates of $r_{i,u}$ are updated proportionally to the confidence that stimulus u was present (computed on the basis of sensory input and reward magnitude) converges to the correct estimates.[5] Similarly like Frank and Claus (2006), we propose that $r_{i,u}$ are learned in the orbitofrontal cortex, because a large amount of experimental data suggests that the magnitudes of rewards associated with stimuli are learned and maintained during choice in this area (Roesch & Olson, 2004; Wallis, 2007; Wallis & Miller, 2003). However, since this article focuses on the basal ganglia, we do not explicitly model learning of the optimal values of $INT_i(0)$.

3.3 Learning Striatal Weights Optimizing DM. Here we show that the actor-critic model (see equations 2.13 to 2.15) learns weights described by equation 3.3 when the model additionally incorporates the biologically realistic assumption that the weights $w_{i,u}$ are bounded, that is, they cannot be negative, and cannot exceed a maximum value, which we set to 1. Thus, if a weight exceeds the [0, 1] range due to modification according to equation 2.15, it is set to the boundary value: 0 or 1.

We additionally make the following two assumptions: the stimulus is correctly identified by the critic, and on a proportion of trials e, suboptimal or exploratory actions are selected. These assumptions may be violated with

[5] Although in both this article and in Larsen et al. (2010) we address the problem of RL for noisy stimuli, we develop algorithms that learn different quantities. Here we develop an algorithm that learns weights given by equation 3.3, whereas Larsen et al. develop an algorithm that estimates $r_{i,u}$.

noisy stimuli, and in section 4, we investigate in simulations how violating these two assumptions affects the weight convergence.

Under the above assumptions, the values estimated by the critic converge to

$$v_u = r_{best,u} - \varepsilon, \tag{3.5}$$

where $r_{best,u} = \max_i r_{i,u}$ and ε is a small, positive constant. This happens because when stimulus u is presented, on the majority of trials the best action is chosen, and hence the expected reward for this stimulus is close to the reward to this action ($r_{best,u}$). But on a fraction of trials, actions are chosen with lower rewards, and thus the expected reward for this stimulus is lower by ε. The value of ε depends on the proportion of exploratory choices e and the differences between r_{best} and rewards for choosing other actions. We now consider two cases.

First, when the best action for stimulus u is chosen, then δ (see equation 2.13) is equal to

$$\delta = r_{best,u} - v_u = r_{best,u} - r_{best,u} + \varepsilon = \varepsilon. \tag{3.6}$$

Hence, according to equation 2.15, the weight $w_{best,u}$ increases, and, as shown in the appendix, the step size of this increase ($\eta\delta$) does not converge to 0 over iterations; thus, $w_{best,u}$ will eventually exceed 1. Once $w_{best,u}$ reaches 1, it will stay at 1 due to the boundaries on $w_{i,u}$ in our model.

Second, when a suboptimal action for stimulus u is chosen, then δ is equal to

$$\delta = r_{subopt,u} - v_u = r_{subopt,u} - r_{best,u} + \varepsilon. \tag{3.7}$$

If there are only two possible actions, then v_u will take a value between $r_{subopt,u}$ and $r_{best,u}$; hence, δ in equation 3.7 is negative. In a more general case of multiple alternative actions, $r_{subopt,u} < r_{best,u}$, hence δ in equation 3.7 is negative as long as ε is sufficiently small, which can be guaranteed by choosing e sufficiently low. Thus, the weight $w_{subopt,u}$ decreases, and over iterations it will eventually decay below 0 and stay at 0 due to the boundaries on $w_{i,u}$ in our model. This completes the argument.

3.4 DM in the Integrated Model. In this section, we show that the integrated model can approximate MSPRT on the trials following learning when the striatal weights are equal to values given by equation 3.3.

To simplify notation and without the loss of generality, let us assume that action i is correct for stimulus $u = i$ (for other stimulus-response mappings, this could be satisfied by renumbering stimuli). Under this assumption, the weight set according to equation 3.3 is equal to $w_{i,i} = 1$ and $w_{i,u\neq i} = 0$.

Hence, the striatal activity is equal to (the third line in the equation below comes from equation 2.4):

$$\begin{aligned} STR_i(t) &= INT_i(t) + g \sum_u w_{i,u} x_u(t) = \\ &= INT_i(t) + g x_i(t) = \\ &= INT_i(t) + \log P(x(t)|H_i) + b(t). \end{aligned} \tag{3.8}$$

We now show, using mathematical induction (analogously as in section 2.2.3), that for $t \geq 1$, the activity of channel i in the output nuclei in the model is

$$OUT_i(t) = -\log P_i(t-1) - (\log P(x(t)|H_i) + b(t)). \tag{3.9}$$

Let us consider the network activity at the first time step. The activity of integrators is equal to $INT_i(1) = \log P_i(0) + c$. The input provided by the STN and GP is (using manipulations as in equation 2.6)

$$SG(1) = \log \sum_{i=1}^{N} P_i(0) + c. \tag{3.10}$$

Since the prior probabilities $P_i(0)$ add up to 1, $SG(1) = c$. Hence, the activity of the output nuclei is

$$\begin{aligned} OUT_i(1) &= -STR_i(1) + SG(1) \\ &= -\log P_i(0) - c - (\log P(x(1)|H_i) + b(1)) + c. \end{aligned} \tag{3.11}$$

Thus, at the first time step, the activity of the output nuclei follows equation 3.9. Now we will show that if equation 3.9 is satisfied at time $t-1$, it is also satisfied at time t. If equation 3.9 is satisfied at time $t-1$, then the activity of integrators at time t is (see the labels in Figure 6)

$$\begin{aligned} INT_i(t) &= c - OUT_i(t-1) \\ &= \log(P_i(t-2)P(x(t-1)|H_i)) + c + b(t-1). \end{aligned} \tag{3.12}$$

Hence, when equation 2.3 is used, the activity of the STN in the model becomes

$$SG(t) = \log P(x(t-1)) + c + b(t-1). \tag{3.13}$$

The activity of the output nuclei becomes

$$\begin{aligned} OUT_i(t) = &-\log(P_i(t-2)P(x(t-1)|H_i)) - (\log P(x(t)|H_i) \\ &+ b(t)) + \log P(x(t-1)). \end{aligned} \tag{3.14}$$

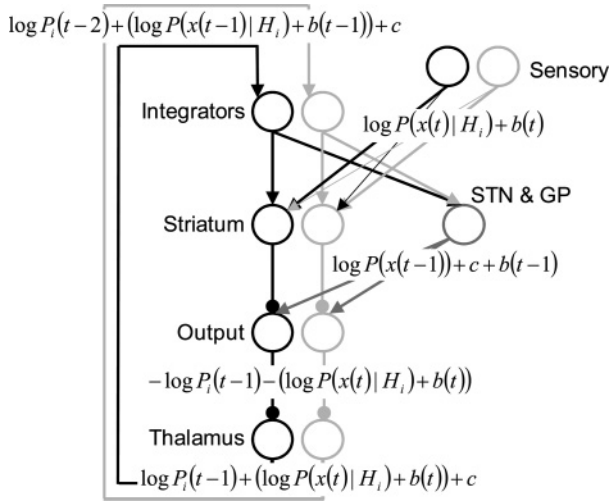


Figure 6: Probabilities encoded by neural activity in the actor. Notation as in Figure 3.

According to Bayes' theorem (see equation 2.2), the activity of the output nuclei in equation 3.14 is equal to that in equation 3.9, which completes the proof.

Note that the normalization term (equation 3.13), is computed on the basis of the input from the previous time step. This happens because the STN receives sensory input from the basal ganglia, thalamus, and cortex, which introduce delays. Consequently, the activity of the output nuclei is proportional to $-\log P_i(t-1)$ decreased by the sensory input gathered during the last time step, where the length of the time step in the model is equal to the time necessary to traverse the cortico-basal-ganglia-thalamic loop. Thus, this network only approximates MSPRT, and how close this approximation is depends on how short the time necessary for information to traverse the cortico-basal-ganglia-thalamic loop is relative to the decision time.

3.5 Consolidation of Stimulus-Response Mapping. It is interesting to add that the above model supports a memory consolidation mechanism proposed by Ashby, Ennis, and Spiering (2007) for learning stimulus-response mapping in the connections between sensory and integrator neurons. In particular, note that once the stimulus-response mapping is formed in cortico-striatal connections, a given integrator will be most active on trials when the corresponding sensory population has the highest mean. For example, in Figure 6, the black sensory and integrator units will be coactive on some trials, and gray sensory and integrator units will be coactive on other trials. Thus simple Hebbian learning (forming connections between

coactive neuronal populations) will form direct connections between sensory neurons and corresponding integrators (i.e., the connections present in Figure 3a).

In summary, once the stimulus-response mapping is initially learned in cortico-striatal connections, the cortico-basal-ganglia circuit can approximate MSPRT. Additionally, at this point, sensory neurons representing stimuli are coactive with integrator neurons representing corresponding actions, and the stimulus-response mapping can be learned in cortico-cortical connections using simple Hebbian rules. This will then allow the network to implement MSPRT more precisely.

4 Simulation of Striatal Weights Learning

This section describes simulations of learning the striatal weights in the integrated model. Recall that while proving convergence of striatal weights to values required for optimal DM in section 3.3, it was assumed that (1) stimuli are correctly identified and that (2) exploratory choices are made. First, two simulations (see sections 4.2 and 4.3) investigate how the value of the decision threshold parameter controlling speed-accuracy trade-off in the model influences the convergence of the weights. It is shown that when it is set to the values emphasizing too much speed or accuracy, assumptions 1 or 2 respectively become violated, which leads to different departures of striatal weights from the optimal values. Then section 4.4 suggests modifications to the model, allowing convergence of the weights to the optimal values.

4.1 Methods of Simulations. A single simulation consists of 150 trials, each consisting of one randomly chosen stimulus presented, one action taken, and one reward received. The simulation of a trial is divided into time steps of 0.001 s, and at each time step, the sensory input $x_u(t)$ is sampled from gaussian distribution with mean I_u and variance σ^2. The mean input of sensory neurons selective for the stimulus u^* presented on a given simulated trial is set to $I_{u^*} = 0.0045$, the means of activity of the other neurons are set to $I_{u \neq u^*} = 0.003$, and the variance is $\sigma^2 = 0.00011$ (these values result in a model performance very close to that of the sample participant performing the moving dots task in the study of Bogacz et al. (2006). The activity of integrators is initialized to $INT_i(0) = 0$.

Depending on the chosen action, a reward is received according to the scheme described in section 3.2 where the reward for correct leftward saccade is $r_{1,1} = 2$, the reward for correct rightward saccade is $r_{2,2} = 1$, and the reward for incorrect saccades is $r_{1,2} = r_{2,1} = 0$. The weights v_u and $w_{i,u}$ are then updated with learning rate $\eta = 0.1$, and another trial commences.

4.2 Simulations with Speed Emphasis. In the first simulation, the decision threshold is set to 0.6. As can be seen in Figures 7a and 7b, the striatal

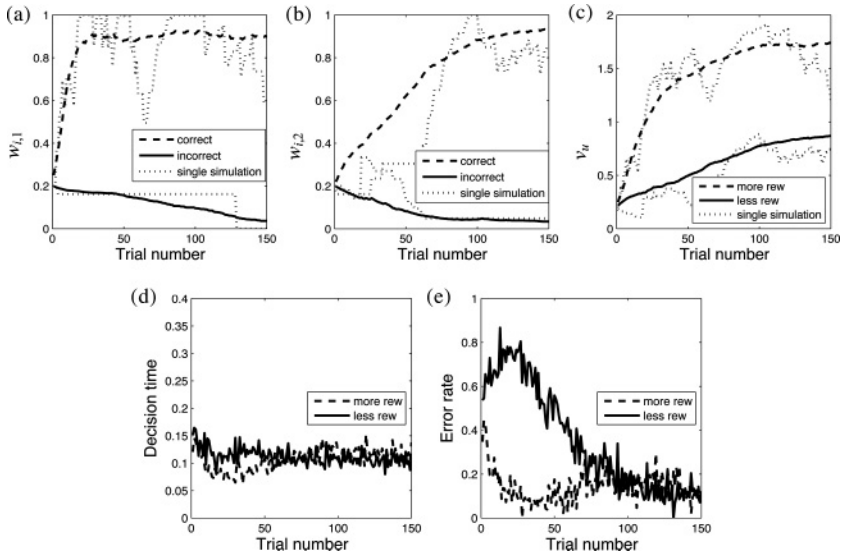


Figure 7: Striatal weights learning in the integrated model during the biased reward task ($r_{1,1} = 2, r_{2,2} = 1, r_{1,2} = r_{2,1} = 0$). The convergence of the striatal weights for the actions chosen in the context of the stimulus with the higher reward (a) and the lower reward (b). The dashed lines represent the correct responses, and the solid lines represent the incorrect responses. (c) The estimated rewards values. (d) The average decision time for more and less rewarded stimuli during learning. (e) The average error rate for more and less rewarded stimuli during learning.

weights converge to the vicinity of the desired values but do not always reach them. An example is shown in Figure 7a where the weight $w_{1,1}$ never converges to 1 as required for optimal behavior. This happens because for the parameters used in the simulations, assumption 1 is not satisfied; that is, the stimulus is incorrectly identified on some trials. In particular, when stimulus_2 is presented but the simulated subject mistakenly believes that stimulus_1 is presented and chooses action_1, then a reward of 0 is received, the δ term becomes negative, and the weight is pushed away from 1. For the same reason, weight $w_{2,2}$ in Figure 7b does not converge to 1. The expected reward value as estimated by the critic is plotted in Figure 7c, which shows that the values learned during single simulations are close to the values averaged over multiple trials, which suggests that the estimation is consistent between trials.

In Figures 7d and 7e, we can see that as the striatal weights are learned, the error rates and, to a lesser extent the decision times decrease. There is, however, a difference in the error rates for the more and the less rewarded stimuli. Early on, the error rate on trials on which the less rewarded stimulus

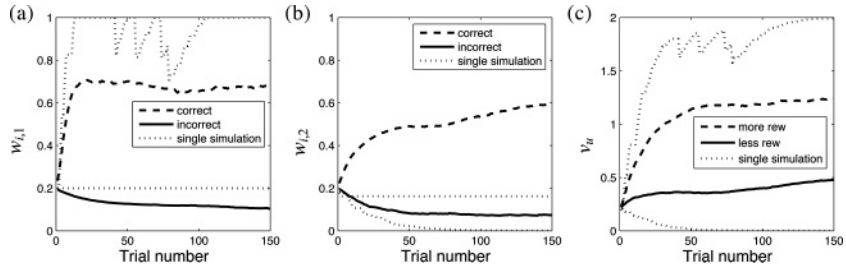


Figure 8: Striatal weights learning in the integrated model during the biased reward task ($r_{1,1} = 2, r_{2,2} = 1, r_{1,2} = r_{2,1} = 0$). The convergence of the striatal weights for the actions chosen with the decision threshold lowered to 0.4, in the context of the stimulus with the higher reward (a) and the lower reward (b). (c) The estimated reward values. In all panels, the dashed lines represent the correct responses, the solid lines represent the incorrect responses, and the dotted lines represent single simulations.

is presented is above the initial 50%, whereas the error rate for the more rewarded stimulus decreases rapidly. This is due to the difference in convergence rates for the weights associated with the two stimuli resulting in $w_{1,1} > w_{2,2}$, and the correct action for the more rewarded (and more converged) stimulus therefore being selected even though the stimulus predicts otherwise. As an example, imagine the stimulus_2 being shown at trial 50. The mean inputs from sensory neurons would be $I_1 = 0.003$, $I_2 = 0.0045$, and the weights at trial 50 are $w_{1,1} \approx 1$, $w_{1,2} \approx 0.2$, $w_{2,2} \approx 0.3$, $w_{2,1} \approx 0.2$ (see Figures 7a and 7b). The mean input to the striatal unit representing action_1 ($w_{1,1}I_1 + w_{1,2}I_2 \approx 1 \times 0.003 + 0.2 \times 0.0045 = 0.0039$) is higher than to the striatal unit representing action_2 ($w_{2,1}I_1 + w_{2,2}I_2 \approx 0.2 \times 0.003 + 0.3 \times 0.0045 = 0.00195$) even though I_1 is lower than I_2. However, once both weights have converged, both stimuli get similar error rates.

Figure 7d also shows that due to the different convergence rates for the more and the less rewarded stimuli, there is a (smaller) difference in the decision times for the two stimuli during learning, but not before or after.

4.3 Simulations with Accuracy Emphasis. In the simulations described in the previous section, the weights did not converge to values of equation 3.3 because stimuli were misinterpreted on some trials. Therefore, one could anticipate that better convergence can be expected when fewer errors are made, and hence we repeated the simulation with the decision threshold lowered to 0.4 (lowering the threshold below which the activity of the output nuclei needs to decrease results in more accurate decisions). Surprisingly, this change leads to much poorer weight convergence, as can be seen in Figures 8a and 8b. When looking at the weight changes for the single simulation, we can see that it is not because of too many errors by the

critic, but rather because of too few errors by the actor leading to lack of exploration—that is, assumption (2) is not satisfied. In the single simulation in Figures 8a and 8b, it is clear that only two of the weights are modified. These are the weights $w_{1,1}$ and $w_{1,2}$, which implies that in this simulation, only action$_1$ is chosen, regardless of the stimulus. Notice that within a few trials, one of the weights becomes much higher than the others, which leads to the corresponding integrator receiving larger inputs than the alternative, independent of the stimulus presented. For example, consider a situation when on trial 20, stimulus$_2$ (i.e., right) is presented; thus, $I_1 = 0.003$, $I_2 = 0.0045$. Although action$_2$ is the correct response, the striatal unit representing action$_1$ gets much higher average input ($w_{1,1}I_1 + w_{1,2}I_2 \approx 1 \times 0.003 + 0.07 \times 0.0045 = 0.00332$) than the striatal unit representing action$_2$ ($w_{2,1}I_1 + w_{2,2}I_2 \approx 0.2 \times 0.003 + 0.2 \times 0.0045 = 0.0015$). Since in different simulations different actions become the preferred action, the averaged values of the weights converge to some value in between the extremes. In Figure 8c, the value estimated by the critic shows the same behavior with the single simulations converging to either 0 or $r_{best,u}$. The averages converge to values in between (1 for more rewarded, 0.5 for less rewarded alternative), showing that each action becomes the preferred action on 50% ofthe simulations.

4.4 Improving Weight Convergence. In this section, we present three approaches to improving weight convergence. The first two approaches aim at satisfying assumptions 1 and 2, that is, ensuring that the critic should make as few errors as possible, while the actor should explore all the options. First, this property would arise naturally in tasks in which, after a choice, the subject can continue to observe the stimulus and hence refine the stimulus information in the critic. In such tasks, the decision threshold in the actor should be set such that occasionally exploratory choices are made, while the integration in the critic should continue beyond the actor's choice until the stimulus is identified with high accuracy.

In tasks in which the stimulus cannot be observed after choice, a second approach can be used in which additional exploration is introduced at the time of the action decision. This approach is inspired by the findings of Daw, O'Doherty, Dayan, Seymour, and Dolan (2006) that certain cortical areas (frontopolar cortex and intraparietal sulcus) are active only on explorative trials. We introduced the exploration in the following way. After the decision threshold is reached, there is a fixed probability p that the actor chooses the winning action and a probability $1 - p$ for choosing the alternative. In the simulations, this p is set to 0.85, with the other parameters kept the same as in the simulation of Figures 8a to 8c. Figures 9a and 9b show that all the weights converge to the desired values as a result of identifying the stimulus on most trials, while still occasionally choosing actions other than the optimal. Figure 9c shows that the value estimation for each stimulus converges to the correct region for both the single simulations and the

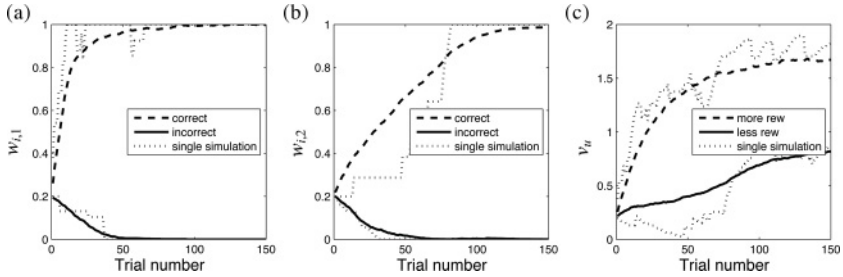


Figure 9: Striatal weights learning in the integrated model during the biased reward task ($r_{1,1} = 2, r_{2,2} = 1, r_{1,2} = r_{2,1} = 0$). The convergence of the striatal weights for the actions chosen with the decision threshold at 0.4 but with added exploration, in the context of the stimulus with the higher reward (a) and the lower reward (b). (c) The estimated reward values. In all panels, the dashed lines represent the correct responses, the solid lines represent the incorrect responses, and the dotted lines represent single simulations.

averages, which suggests that the critic makes a reasonable estimation of the expected reward, also in individual simulations.

The strategy of modifying the decision threshold has some intuitive justification in that it seems reasonable to be more careful with decisions if the reward scheme is unknown. This strategy works by avoiding mistakes, but if the difficulty of the perceptual choice is increased to the level for which mistakes cannot be avoided, the weights will fluctuate around the boundaries at 0 and 1, and the average weights will never converge.

A third approach, offering a possible solution to the above problem, is to assume that the weight parameters $w_{i,u}$ are not restricted to the range between 0 and 1, but that the information transmitted through them is. In particular, this model assumes that the information transmitted through the synapse is equal to $w_{i,u} \times x_u$ if $w_{i,u} \in [0, 1]$, is equal to 0 if $w_{i,u} < 0$ and equal to x_u if $w_{i,u} > 1$.

When simulating this, we put an additional boundary on the weights so that they cannot exceed -1 and 2 (recall that the numbers are arbitrary) and run the simulation with the same parameters as in Figure 9 except that the decision threshold is 0.6 as in Figure 7.

As Figure 10 shows, the weights converge to the required values. The less rewarded correct action, $w_{2,2}$, is slow in its convergence (and hence the simulation is extended to 250 trials), though it still converges.

5 Discussion

In this article, we argue that the time is ripe for an integration of RL and optimal DM theories. As first steps toward it, we have shown that in tasks

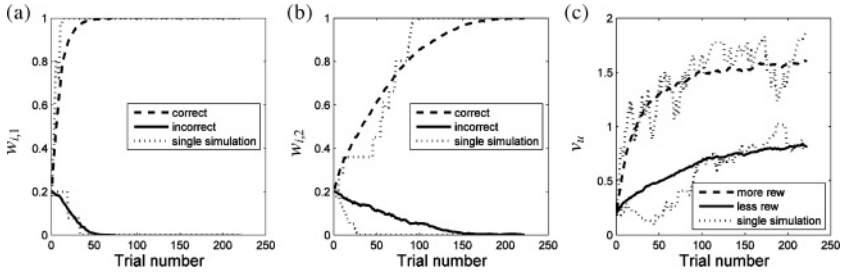


Figure 10: Striatal weights learning in the integrated model with double boundaries during the biased reward task ($r_{1,1} = 2, r_{2,2} = 1, r_{1,2} = r_{2,1} = 0$). The convergence of the striatal weights for the actions chosen in the context of the stimulus with the higher reward (a) and the lower reward (b). The dashed lines represent the correct responses, the solid lines represent the incorrect responses, and dotted lines represent single simulations. (c) The estimated rewards values.

in which a stimulus response mapping has been learned using RL, the cortico-basal-ganglia network can approximate optimal DM. Furthermore, we have described the conditions that cortico-striatal weights encoding stimulus-response mappings need to satisfy in order to allow optimal DM, and we developed a modified actor-critic model for learning these weights. In this section, we discuss predictions of the model and the directions in which the integration of RL and optimal DM theories may be developed.

5.1 Predictions of the Model. The proposed integrated model makes several predictions that could be tested in behavioral and neurophysiological data. Figures 7d and 7e illustrate a predicted pattern of behavioral data in the biased reward task. Namely, on initial trials, reaction times and error rates should on average be lower for the more rewarded stimulus (due to larger changes in its corresponding weights), but these differences across stimuli should be reduced as the training progresses (as the striatal weights approach the values given in equation 3.3).

In our simulations, the decision times were identical for both stimuli after learning (see Figure 7d), but it is important to note that we did not simulate learning of optimal initial values of the integrators. Contrary to this simulation, it has been observed that human participants make faster responses to more rewarded stimuli in highly practiced tasks, but these faster responses may arise due to the increases in initial values of integrators (Simen, Buck, Hu, Holmes, & Cohen, 2009). Consequently, our model predicts that the decision times should be equal for more and less rewarded stimuli once the weights have converged (as in Figure 7d) in patients who do not modify the initial values of integrators (e.g., in patients with orbitofrontal damage who are impaired in estimating reward magnitudes). In such patients, the

relative response times for different stimuli are most likely to depend only on striatal weights, as in our simulations.

The model's predictions on striatal weights and initial values of the integrators can also be tested by employing mathematical models of decision making such as LATER (Carpenter & Williams, 1995) or a diffusion model (Ratcliff, 1978; Vandekerckhove & Tuerlinckx, 2007). These models allow estimating the rate of evidence accumulation, which according to our model should depend on striatal weights, and initial values of integrators. Our model predicts that in the biased reward task, the estimated rates of evidence accumulation for the two stimuli should differ early in the training, but should become equal after learning. Our model also predicts that after learning, the estimated initial starting point of the integrator corresponding to the more rewarded action should be higher than for the starting point estimated for the other integrator.

During learning in the model, the cortical integrators receive input from sensory neurons via striatum, output nuclei, and thalamus, while in a highly practiced task, the integrators receive the input directly from sensory neurons (see section 3.5). Thus, the model predicts that the neural integrators should start to increase their activity earlier after stimulus onset in highly practiced tasks than during learning. Law and Gold (2008) have investigated the activity of LIP neurons while the monkey learned the moving dots tasks. Although their published data do not allow proper statistical verification of this prediction, they seem to be consistent with it (e.g., the black curve in right panel of their Figure 4a, corresponding to a high level of practice, starts to depart from baseline 25 ms to 50 ms earlier than in the middle panel of Figure 4a, corresponding to an intermediate level of practice). The model also predicts that microstimulations of MT neurons should affect LIP activity with a longer latency during learning than they do in highly practiced tasks (Hanks, Ditterich, & Shadlen, 2006).

5.2 Combining Uncertainties in RL and DM. When considering how to integrate RL and DM theories, it is pivotal to consider how the uncertainties related to each theory must combine if the resulting behavior is to be optimal. As mentioned at the beginning of the article, the main uncertainties concerned by the RL and DM models are those in stimulus-response mapping and stimulus-identity, respectively.[6]

[6]There are also other uncertainties in RL. In many RL tasks, the rewards are assigned probabilistically, for example, selecting the same action for the same stimulus may result in a reward on 75% randomly chosen trials. This introduces an uncertainty, to which Yu and Dayan (2005) refer as the expected uncertainty, because a subject can learn over trials to expect the probabilistic nature of rewards. Daw, Niv, and Dayan (2005) propose another type of uncertainty, which is produced by a high-level model of the task in the prefrontal cortex. For simplicity, we do not consider these uncertainties further.

The stimulus-response uncertainty, traditionally associated with RL models, influences decision times, traditionally associated with DM models. Pasupathy and Miller (2005) reported longer reaction times on trials with higher stimulus-response uncertainty (trials following a change in stimulus response mapping). This suggests that learning and decision processes interact while dealing with the uncertainties.

When attempting to improve performance, a subject would have to decrease the uncertainties; however, there is a fundamental difference between the ways these uncertainties can be decreased. The stimulus-response uncertainty can be reduced by reinforcement learning over trials, while the stimulus-identity uncertainty can be lessened by longer observation times within a trial.[7]

To optimally reduce these two uncertainties, one needs to optimally choose the parameter controlling the speed-accuracy trade-off (see section 2.4). The above observations of Pasupathy and Miller (2005) might indicate that in more uncertain circumstances, subjects want to be more certain of the stimulus before making an action. This strategy may be beneficial. In the presence of stimulus-response uncertainty, correct identifications of the stimuli increase the probability that the stimulus-response mapping will be learned correctly. Thus, one of the important directions in integrating RL and DM models will be the identification of an optimal value of threshold parameter controlling the speed-accuracy trade-offs that will allow maximization of reward rate and how this value can be learned (Simen, Cohen, & Holmes, 2006).

5.3 Related Work. Work on combining the uncertainty of RL with stimulus uncertainty has been also done in the domain of robot control. Kaelbling, Littman, and Cassandra (1998) formalized the stimulus uncertainty in a framework of partially observable Markov decision processes (POMDPs), which assumes that an agent is not fully certain of its state (e.g., position in the environment), but instead perceives observations that depend on the agent's state and action. In this framework, if an agent is very uncertain of its state, in order to maximize the long-term reward, it may be beneficial to choose actions that decrease its state uncertainty (e.g., to observe). In fact, their framework is so general that it can describe the task we consider in section 3.2: in each time point during the stimulus presentation, the agent would be choosing from left saccade, right saccade, and continuation of observation (which would be described as a third action).

The aim of Kaelbling et al. (1998) was to develop an efficient algorithm that could run on a robot's computer and find the optimal policy for any

[7]It is worth noting that even in experiments in which subjects are told the underlying reward probabilities (and hence, in theory, eliminating stimulus-response uncertainties), feedback over multiple trials is needed for the behavior to fully reflect the reward probabilities (Jessup, Bishara, & Busemeyer, 2008).

POMDP, while we investigate how simple learning problems can be solved in the known anatomy of cortico-basal-ganglia circuits. These different perspectives led to differences in our solutions, and we discuss three of them. First, although their algorithm is very elegant, it is so complex that it is difficult to imagine how it could be implemented in neural circuits. Second, the algorithm assumes that the agent fully knows the model of the world (e.g., set of states, transition between them due to actions) and computes the optimal policy "offline" on the basis of the model, while we assume that the agent learns by interacting with the environment (as it is the case for biological organisms). Third, in their framework, observation is treated as any other action, while we treat it in a special way because motor actions produce very different neural responses in the cortico-basal-ganglia circuit than continuation of observation.

More recently, Dayan and Daw (2008) developed an elegant probabilistic framework in which both RL and DM can be described. In this framework, a subject estimates the state of the environment on the basis of external cues and can either execute an action or continue observation. Dayan and Daw (2008) discussed neural implementation of the computations in their framework, but did not explicitly map them on the cortico-basal-ganglia circuit.

Law and Gold (2009) recently published a very interesting computational model describing decision making and learning in the moving dots task. In their model, LIP integrates sensory input from MT during the decision process, and after the feedback, the weights between MT and LIP neurons are modified proportionally to the reward prediction error. Although Law and Gold (2009) modify in their model the weights between MT and LIP, they write "Our model is not informative about where in the brain the actual changes occur. Rather, the model establishes principles governing how functional (that is, direct or indirect) connectivity . . . is modified by experience" (p. 661). We feel that our model describes how first indirect and then direct connections between sensory neurons and integrators can be formed in accordance with the principles in their model.

5.4 Further Directions. There are several directions in which a theory integrating RL and optimal DM can be developed:

- So far we have focused on learning from rewards, but it has been proposed that learning from punishments involves a separate population of striatal neurons that project to GP (Frank et al., 2004). It would be interesting to investigate if the cortico-basal-ganglia circuit can simultaneously support learning from punishments and optimal DM.
- In recent work, Dimperio, Jessup, and Busemeyer (2010) showed that in nonperceptual DM, the different behavior seen with and without feedback can be captured by a model combining decision field theory

and RL. For a full understanding of choices in humans, the common aspects of this and our model should be investigated.

- The convergence proof in section 3.3 assumed that rewards are deterministic, and it would be interesting to also consider tasks in which rewards are delivered stochastically (with means defined by a payoff matrix).
- It would also be interesting to investigate how the optimal value of the learning rate η in RL models depends on stimulus uncertainty.
- So far in this article, dopamine was discussed only in the context of RL, but it also plays an important role during DM, and two interesting theories on this role have been proposed (Gurney, Humphries, Wood, Prescott, & Redgrave, 2004; McClure, Daw, & Montague, 2003). It would be interesting to analyze how they relate to optimal DM.

Appendix

Here we show that when the best action is chosen in the modified actor-critic model of section 3.3, term δ does not converge to zero but can be bounded from below by a positive constant that does not decrease over iterations. This condition is necessary to guarantee that $w_{best,u}$ will reach 1, because if $\delta > 0$ but $\delta \to 0$, it would be still possible that $w_{best,u}$ would increase and converge to a value lower than 1 (David Leslie, personal communication, August 2007).

Let us note that the expected rewards for stimuli learned by the critic converge to

$$v_u \to \sum_i P_i r_{i,u}. \tag{A.1}$$

We now assume that v_u is equal to the value given in equation A.1 and compute the lower bound on δ after choosing the best action for stimulus u:

$$\begin{aligned} \delta &= r_{best,u} - v_u = r_{best,u} - \sum_i P_i r_{i,u} = \\ &= r_{best,u}(1 - P_{best}) - \sum_{i \neq best} P_i r_{i,u}. \end{aligned} \tag{A.2}$$

Let $r_{2nd,u}$ denote the reward for selecting the second-best action for stimulus u, so by definition, $r_{2nd,u} \geq r_{i,u}$ for $i \neq best$, hence:

$$\delta \geq r_{best,u}(1 - P_{best}) - \sum_{i \neq best} P_i r_{2nd,u}. \tag{A.3}$$

Now we use the fact that P_i sum to 1, so condition A.3 is equivalent to

$$\delta \geq (r_{best,u} - r_{2nd,u})(1 - P_{best}). \tag{A.4}$$

The first term in brackets is strictly positive if $r_{best,u} > r_{2nd,u}$. The second term in brackets is equal to e, which was assumed to be positive and constant in section 3.3.

Acknowledgments

This work was supported by EPSRC grants EP/C516303/1 and EP/C514416/1. We thank David Leslie, Peter Trimmer, and Patrick Simen for reading earlier versions of the manuscript and making very useful comments, and Sean Collins for discussion.

References

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci., 9*, 357–381.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114*, 632–656.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., & O'Reilly, R. C. (2007). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat. Neurosci., 10*(1), 126–131.

Baum, C. W., & Veeravalli, V. V. (1994). A sequential procedure for multihypothesis testing. *IEEE Transactions on Information Theory, 40*, 1996–2007.

Bogacz, R. (2009). Optimal decision-making theories. In J.-C. Dreher & L. Tremblay (Eds.), *Handbook of reward and decision making*. Orlando, FL: Academic Press.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review, 113*, 700–765.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation, 19*, 442–477.

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Res., 1153*, 111–121.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1993). Responses of neurons in macaque MT to stochastic motion signals. *Vis. Neurosci., 10*(6), 1157–1169.

Brown, J. W., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Netw., 17*(4), 471–510.

Carpenter, R. H., & Williams, M. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature, 377*(6544), 59–62.

Chevalier, G., Vacher, S., Deniau, J. M., & Desban, M. (1985). Disinhibition as a basic process in the expression of striatal functions. I. The striato-nigral influence on tecto-spinal/tecto-diencephalic neurons. *Brain Res., 334*(2), 215–226.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci., 8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*, 876–879.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci., 8*(4), 429–453.

Deniau, J. M., & Chevalier, G. (1985). Disinhibition as a basic process in the expression of striatal functions. II. The striato-nigral influence on thalamocortical cells of the ventromedial thalamic nucleus. *Brain Res., 334*(2), 227–233.

Dimperio, E., Jessup, R. K., & Busemeyer, J. R. (2010). *Integrating sophisticated choice models with basic learning processes to more fully account for complex choice behavior*. Manuscript submitted for publication.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr. Opin. Neurobiol., 10*(6), 732–739.

Dragalin, V. P., Tertakovsky, A. G., & Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests—part I: Asymptotic optimality. *IEEE Transactions on Information Theory, 45*, 2448–2461.

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Netw., 19*(8), 1120–1136.

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-Orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol. Rev., 113*(2), 300–326.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science, 306*(5703), 1940–1943.

Gerfen, C. R. (1992). The neostriatal mosaic: Multiple levels of compartmental organization in the basal ganglia. *Annu. Rev. Neurosci., 15*, 285–320.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci., 5*(1), 10–16.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36*(2), 299–308.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci., 30*, 535–574.

Gurney, K., Humphries, M., Wood, R., Prescott, T. J., & Redgrave, P. (2004). Testing computational hypotheses of brain systems function: A case study with the basal ganglia. *Network, 15*(4), 263–290.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol. Cybern., 84*(6), 401–410.

Hallworth, N. E., Wilson, C. J., & Bevan, M. D. (2003). Apamin-sensitive small conductance calcium-activated potassium channels, through their selective coupling to voltage-gated calcium channels, are critical determinants of the precision, pace,

and pattern of action potential generation in rat subthalamic nucleus neurons in vitro. *J. Neurosci., 23*(20), 7525–7542.

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nat. Neurosci., 9*(5), 682–689.

Humphries, M. D., Stewart, R. D., & Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J. Neurosci., 26*(50), 12921–12942.

Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science, 19*, 1015–1022.

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

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks, 15*, 549–559.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759–764.

Larsen, T., Leslie, D. S., Collins, E. J., & Bogacz, R. (2010). Posterior weighted reinforcement learning with stimulus uncertainties. *Neural Computation, 22*, 1149–1179.

Law, C. T., & Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nat. Neurosci., 11*(4), 505–513.

Law, C. T., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nat. Neurosci., 12*(5), 655–663.

Lo, C. C., & Wang, X. J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time task. *Nat. Neurosci., 9*, 956–963.

Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cereb. Cortex, 13*(11), 1257–1269.

McClure, S. M., Daw, N. D., & Montague, P. R. (2003). A computational substrate for incentive salience. *Trends Neurosci., 26*(8), 423–428.

McMillen, T., & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology, 50*, 30–57.

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

Nambu, A., & Llinas, R. (1994). Electrophysiology of globus pallidus neurons in vitro. *J. Neurophysiol., 72*(3), 1127–1139.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452–454.

Parent, A., & Smith, Y. (1987). Organization of efferent projections of the subthalamic nucleus in the squirrel monkey as revealed by retrograde labeling methods. *Brain Res., 436*(2), 296–310.

Pasupathy, A., & Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature, 433*(7028), 873–876.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400*, 233–238.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev., 83*, 59–108.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 53*, 195–237.

Ratcliff, R., Gomez, R., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*, 159–182.

Ratcliff, R., & Smith, P. L. (2004). Comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333–367.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience, 89*(4), 1009–1023.

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature, 413*(6851), 67–70.

Roesch, M. R., & Olson, C. R. (2004). Neuronal activity related to reward value and motivation in primate frontal cortex. *Science, 304*(5668), 307–310.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci., 22*(21), 9475–9489.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science, 310*(5752), 1337–1340.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nat. Rev. Neurosci., 2*(1), 33–42.

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

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol., 86*(4), 1916–1936.

Shea-Brown, E., Gilzenrat, M. S., & Cohen, J. D. (2008). Optimization of decision making in multilayer networks: The role of locus coeruleus. *Neural Comput., 20*(12), 2863–2894.

Simen, P. A., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Performance and Psychophysics, 35*, 1865–1897.

Simen, P. A., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks, 19*, 1013–1026.

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

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307*(5715), 1642–1645.

Ungless, M. A., Magill, P. J., & Bolam, J. P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science, 303*(5666), 2040–2042.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychol. Rev., 108*(3), 550–592.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin and Review, 14*, 1011–1026.

Wald, A. (1947). *Sequential analysis*. Hoboken, NJ: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics, 19*, 326–339.

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci., 30*, 31–56.

Wallis, J. D., & Miller, E. K. (2003). From rule to response: Neuronal processes in the premotor and prefrontal cortex. *J. Neurophysiol., 90*(3), 1790–1806.

Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron, 36*(5), 955–968.

Wilson, C. J., Weyrick, A., Terman, D., Hallworth, N. E., & Bevan, M. D. (2004). A model of reverse spike frequency adaptation and repetitive firing of subthalamic nucleus neurons. *J. Neurophysiol., 91*(5), 1963–1980.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681–692.

Zhang, J., & Bogacz, R. (2010). Optimal decision making on the basis of evidence represented in spike trains. *Neural Computation, 22*, 1113–1148.

Received November 30, 2009; accepted September 12, 2010.