Commun Nonlinear Sci Numer Simulat 82 (2020) 105048

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

journal homepage: www.elsevier.com/locate/cnsns

Research paper

Corticostriatal synaptic weight evolution in a two-alternative forced choice task: a computational study

C. Vich[a,1], K. Dunovan[b,1], T. Verstynen[b,*], J. Rubin[c,*]

[a] Dept. de Matemátiques i Informática, Universitat de les Illes Balears, Palma, Illes Balears, Spain
[b] Dept. of Psychology, Carnegie Mellon University and Center for the Neural Basis of Cognition, Pittsburgh, PA, USA
[c] Dept. of Mathematics and Center for the Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Available online 15 October 2019

Keywords:
Dopamine
Reinforcement learning
Decision-making
Spike timing-dependent plasticity
Corticostriatal synapses

ABSTRACT

In natural environments, mammals can efficiently select actions based on noisy sensory signals and quickly adapt to unexpected outcomes to better exploit opportunities that arise in the future. Such feedback-based changes in behavior rely, in part, on long term plasticity within cortico-basal-ganglia-thalamic networks, driven by dopaminergic modulation of cortical inputs to the direct and indirect pathway neurons of the striatum. While the firing rates of striatal neurons have been shown to adapt across a range of feedback conditions, it remains difficult to directly assess the corticostriatal synaptic weight changes that contribute to these adaptive firing rates. In this work, we simulate the evolution of corticostriatal synaptic weights based on a spike timing-dependent plasticity rule driven by dopamine signaling that is induced by outcomes of actions in the context of a two-alternative forced choice task. Our results establish 1) that this plasticity model can successfully learn to select the most rewarding actions available, 2) that in the effective regime plasticity predominantly impacts direct pathway weights, evolving to drive action selection toward a more-rewarded action, and 3) that there can be coactivation of opposing populations within selected action channels, as observed experimentally. The model performance also agrees with the results of behavioral experiments carried out previously in human subjects using probabilistic reward paradigms.

1. Introduction

The flexible range of mammalian behavior in dynamic and often volatile environments suggests that the neural circuits associated with action selection must be highly modifiable. This adaptive behavior requires that the outcomes of past experiences influence neural circuits in a principled way that maximizes the chances of success in the future [1]. A significant body of experimental work has established that corticostriatal synapses represent one site of such plasticity, which is triggered when a behavior followed by an unexpected reward leads to a change in dopamine levels [2–4]. Because the corticostriatal synapses represent a key input pathway to the cortico-basal ganglia-thalamic (CBGT) circuits, these dopaminergic changes have been shown to have a critical impact on global network computations related to action selection [5–11].

* Corresponding authors.
E-mail addresses: timothyv@andrew.cmu.edu (T. Verstynen), jonrubin@pitt.edu (J. Rubin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cnsns.2019.105048
1007-5704/

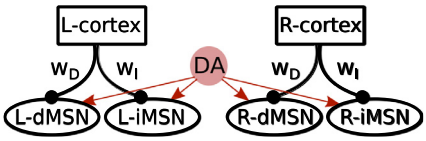

Fig. 1. Spike timing-dependent plasticity (STDP) network. A schematic representation of the implemented neural network model with dopamine (DA) effects on corticostriatal synapses. The *L* and *R* notation denotes the population that influences the choice of the left and right action, respectively. The populations involved are, with $j \in \{L, R\}$, $j - Cortex$: cortical population, $j - dMSN$: direct pathway striatal neurons, and $j - iMSN$: indirect pathway striatal neurons. The strengths of corticostriatal synapses are encoded as weights that evolve over time. Each synapse has its own weight; weights to dMSNs (w_D) obey different plasticity rules than weights to iMSNs (w_I).

Cortical signals to the striatum infiltrate the overall basal ganglia network dynamics via at least two distinct routes, the direct and indirect pathways, each targeted by a corresponding population of striatal medium spiny neurons (MSNs). While these populations are sometimes called D1 (direct) and D2 (indirect) MSNs based on the predominant dopamine receptors that they express, we will refer to these neuron types as dMSNs and iMSNs, respectively. A classic hypothesis posits that the direct pathway provides a "go" signal that permits an action to be implemented, by disinhibiting downstream targets of inhibitory basal ganglia outputs [12]. Different actions may be driven by dMSNs in different channels, and selection of an action involves cortical activation of the corresponding channel. According to this framework, the indirect pathway promotes inhibitory outputs, leading it to be classically referred to as a "no-go" pathway. When an action associated with one channel is selected, the activity of iMSNs in other channels prevents simultaneous activation of competing actions. According to this theory, if they are active after dMSNs, then the iMSNs in the same channel can terminate previously selected actions [13].

Recent experiments, however, have shown that both the dMSNs and the iMSNs linked with a particular action are simultaneously active during action selection [14–17]. This co-activation of dMSN and iMSN populations has challenged the traditional model of a strict isomorphism between dMSN activity and excitation and iMSN activity and inhibition [12]. Indeed, more recent theoretical models have proposed that, within an action channel, the dMSN and iMSNs work in a competitive manner to regulate the certainty of a given action decision [10,18–20]. For example, Dunovan & Verstynen (2016) proposed a Believer-Skeptic framework for understanding CBGT circuit computations [10]. In this model the direct pathway is cast as the Believer, activated by evidence supporting the favorability of a given action, while the indirect pathway serves as the Skeptic, activated by inputs not in favor of that action. The greater the Believer-Skeptic competition within an action channel, the slower the accumulation of evidence in favor of that action. While this viewpoint shares some similarities with the more classic model [12], it allows for simultaneous increases in activity of dMSN and iMSN populations in a single action channel, corresponding to the accumulation of all types of information relating to that action.

Previous work has developed a computational representation of corticostriatal plasticity in the context of action learning and extinction within the full CBGT circuit [21,22]. In this framework, corticostriatal synapses are updated based on a spike timing-dependent plasticity (STDP) rule, determined by the timing of a striatal neuron's spikes relative to the cortical inputs it receives, and on dopamine signals related to reward prediction errors. While the dopamine is shared across neurons and their synapses, the STDP rule sets a synapse-specific eligibility [23–25], such that only those synapses active with the appropriate timing relative to changes in dopamine are modified.

Here we attempt to adapt and update previous models of dopaminergic learning at the corticostriatal synapses to study how, with repeated evidence presentation, dopamine continuously sculpts synaptic weights at dMSNs and iMSNs in order to influence their relative firing patterns and subsequent behavior. We incorporate dynamically evolving dopamine levels in the setting of either constant or probabilistic rewards delivered in a two-alternative forced choice scenario. In our model, spiking activity of striatal neurons in each action channel is driven by ongoing cortical spike trains (Fig. 1), with action selection based on spike patterns at the striatal level resulting from learning-induced weight asymmetries, not from differences in cortical patterns between action channels. With this set-up we asked: Can spike timing-dependent plasticity of corticostriatal synapses, linked to reward-related dopamine levels and learned action values, learn to select more-rewarded actions in ways that mimic the dynamics observed in human participants? If so, what corticostriatal synaptic weight changes support this performance?

Our results yield a positive answer to the first question and generate the prediction that the predominant site of corticostriatal plasticity arises at synapses to dMSNs. Moreover, we observe that the emergent striatal activity patterns produced by our model involve significant spiking in both dMSN and iMSN populations associated with a selected action, consistent with both experimental findings [14–16] and with competing pathway models [19,20] like the Believer-Skeptic hypothesis [10].

The remainder of this paper is organized as follows. In Section 2, we describe our neural model, including cortical spike trains, MSN dynamics, synaptic plasticity, action selection, and reward delivery. We also illustrate how the model effectively functions to update synaptic weights and how it is implemented computationally. In Section 3, we describe the model performance when rewards associated with each action are at a fixed level, when rewards are generally fixed but switch after a specific condition is met, and when rewards are probabilistic. The latter scenario allows us to compare our results to experiments with human subjects. Finally, we conclude with a discussion in Section 4 where we summarize our findings and describe how our model compares to contemporary theoretical models and to emerging empirical observations.

2. Methods

The focus of this work is on a computational model of striatal medium spiny neurons (MSNs) receiving cortical inputs via synapses with plastic weights that determine either a left (L) or a right (R) action decision. In this section we describe the model network (Section 2.1) and how the corticostriatal synapses change according to spike timing-dependent plasticity (STDP), which is driven by phasic reward signals resulting from simulated actions and their consequent dopamine release (Sections 2.2 and 2.3). An example simulation to illustrate the mechanics of the plasticity rule is also presented (Section 2.4).

2.1. Neural model

We consider a computational model of the striatum consisting of two different populations that receive distinct streams of inputs from the cortex (see Fig. 1). We assume that the two cortical input streams are statistically identical, representing equivalent levels of evidence for what the cortical streams encode. Although they do not interact directly, the two striatal populations compete with each other to be the first to select a corresponding action.

Each population contains two different types of units: (i) dMSNs, which facilitate action selection, and (ii) iMSNs, which suppress action selection. Each of these neurons is represented with the exponential integrate-and-fire model, a simplified model that captures the fundamental properties of conductance-based models [26], such that each neural membrane potential obeys the differential equation

$$C\frac{dV}{dt} = -g_L(V - V_L) + g_L\Delta_T e^{(V-V_T)/\Delta_T} - I_{syn}(t) \tag{1}$$

where g_L is the leak conductance and V_L the leak reversal potential. In terms of a neural $I-V$ curve, V_T denotes the voltage that corresponds to the largest input current to which the neuron does not spike in the absence of synaptic input, while Δ_T stands for the spike slope factor, related to the sharpness of spike initialization. $I_{syn}(t)$ is the synaptic current, given by $I_{syn}(t) = g_{syn}(t)(V(t) - V_{syn})$, where the synaptic conductance $g_{syn}(t)$ is impacted by a learning process (see Section 2.2). A reset mechanism is imposed that represents the repolarization of the membrane potential after each spike. Hence, when the neuron reaches a boundary value V_b, the membrane potential is reset to V_r.

The inputs from the cortex to each MSN neuron within a population are generated stochastically, in order to simulate the upstream sensory or planning signals for the targets of either a left or right action. To start, we generate a baseline oscillatory Poisson process $\{X(t_n)\}_n$, which we call the mother train. This process has rate λ such that the spike probability at time point t_n is $P(X(t_n) = 1) \propto \lambda\delta t$, where $\delta t := t_n - t_{n-1}$ is the time step. From this mother train, we generate the daughter trains, each representing the spikes coming to a corresponding MSN neuron. Specifically, each mother spike is transferred to each daughter with probability p, checked independently for each daughter. As discussed below, we take advantage of this construction to assign different p values for different post-synaptic cell types.

Now, the parameters that are relevant to our model are those associated with the daughter trains, namely the rate ν of each daughter train and the pairwise correlation c between daughter trains. We select λ and p to achieve the desired ν and c; to do so, we take the mother train's rate to be $\lambda = \nu/(p * \delta t)$ where

$$p = \nu + c(1 - \nu). \tag{2}$$

Since the mother process generates a spike with probability $\lambda\delta t$ in time step δt (when $A = 0$), and $\lambda\delta t = \nu/p$ is independent of δt, it follows that this choice of λ allows us to maintain our desired daughter spike rate ν, even if we change the time step of our mother Poisson process. Note from Eq. (2) that the way that we can achieve $c = 1$ is to have $p = 1$, such that daughters are fully correlated because they all receive all of the mother spikes. On the other extreme, if no spikes are transferred from the mother train, then this will result in the daughter train having no spikes in it, which means that $p = 0$ should correspond to $\nu = 0$ and $c = 0$, and this is exactly what is given by equation (2). The other way to get $c = 0$ is to have a mother spike rate of one per time step, $\lambda = 1/\delta t$, and to have $p = \nu$, and this is also exactly what emerges from our formulas.

In the STDP network (see Fig. 1, left) for each possible action, we instantiate a corresponding mother train to generate the cortical daughter spike trains for the MSN populations corresponding to that action. Each dMSN neuron or iMSN neuron receives input from a distinct daughter train, with the corresponding transfer probabilities p^D and p^I, respectively. As shown in [27], the cortex to iMSN release probability exceeds that of cortex to dMSN. Hence, we set $p^D < p^I$.

Striatal neuron parameters. We set the exponential integrate-and-fire model parameter values as $C = 1\,\mu F/cm^2$, $g_L = 0.1\,\mu S/cm^2$, $V_L = -65\,mV$, $V_T = -59.9\,mV$, and $\Delta_T = 3.48\,mV$ (see [26]). The reset parameter values are $V_b = -40\,mV$ and $V_r = -75\,mV$. The synaptic current derives entirely from excitatory inputs from the cortex, so the synaptic reversal potential $V_{syn} = 0\,mV$. For these specific parameters, synaptic inputs are required for MSN spiking to occur.

Cortical neuron parameters. To compute p, we set the daughter Poisson process parameter values as $\nu = 0.002$ and $c = 0.5$ and apply Eq. (2). Once the mother trains are created using these values, we set the iMSN transfer probability to $p^I = p$ and the dMSN transfer probability to $p^D = 2/3\,p^I$. We have also tested the learning rule with small oscillations in the cortical spike rate and obtained similar results as in the constant rate case.

Numerical details. The network was integrated computationally using the Runge-Kutta (4,5) method in Matlab (ode45) with the time step $\delta t = 0.01\,ms$. Different realizations lasting $15\,s$ were computed to simulate variability across different subjects in a learning scenario.

Every time that an action is performed (see Sections 2.3 and 2.4), all populations stop receiving inputs from the cortex until 50 *ms* pass without any striatal spikes. During these silent periods, since no MSN spikes occur, no new actions are performed (i.e., they are action refractory periods). After these 50 *ms*, the network starts receiving synaptic inputs again and we consider a new trial to be underway.

2.2. Learning rule

During the learning process, the corticostriatal connections are strengthened or weakened according to previous experiences. In this subsection, we will present equations for a variety of quantities, many of which appear multiple times in the model because they can take different values for different synapses, cells or spike trains. Specifically, there are variables g_{syn}, w for each corticostriatal synapse, A_{PRE} for each daughter train, and A_{POST} and E for each MSN. For all of these, to avoid clutter, we omit subscripts that would indicate explicitly that there are many instances of these variables in the model.

We suppose that the conductance for each corticostriatal synapse onto each MSN neuron, $g_{syn}(t)$, obeys the differential equation

$$\frac{dg_{syn}}{dt} = \Sigma_j w(t_j)\delta(t - t_j) - g_{syn}/\tau_g, \tag{3}$$

where t_j denotes the time of the jth spike in the cortical daughter spike train pre-synaptic to the neuron, $\delta(t)$ is the Dirac delta function, τ_g stands for the decay time constant of the conductance, and $w(t)$ is a weight associated with that train at time t. The weight is updated by dopamine release and by the neuron's role in action selection based on a similar formulation to one proposed previously [22], which descends from earlier work [24]. The idea of this plasticity scheme is that an eligibility trace E (cf. [25,28]) represents a neuron's recent spiking history and hence its eligibility to have its synapses modified, with changes in eligibility following a spike timing-dependent plasticity (STDP) rule that depends on both the pre- and the post-synaptic firing times. Plasticity of corticostriatal synaptic weights depends on this eligibility together with dopamine levels, which in turn depend on the reward consequences that follow neuronal spiking.

To describe the evolution of neuronal eligibility, we first define A_{PRE} and A_{POST} to represent a record of pre- and post-synaptic spiking, respectively. Every time that a spike from the corresponding spike train or cell occurs, the associated variable increases by a fixed amount, and otherwise, it decays exponentially. That is,

$$\begin{aligned} \frac{dA_{PRE}}{dt} &= (\Delta_{PRE}X_{PRE}(t) - A_{PRE}(t))/\tau_{PRE}, \\ \frac{dA_{POST}}{dt} &= (\Delta_{POST}X_{POST}(t) - A_{POST}(t))/\tau_{POST}, \end{aligned} \tag{4}$$

where $X_{PRE}(t)$ and $X_{POST}(t)$ are delta functions with support at times when, respectively, a neuron that is pre-synaptic to the post-synaptic neuron, or the post-synaptic neuron itself, fires a spike. Δ_{PRE} and Δ_{POST} are the fixed increments to A_{PRE} and A_{POST} due to this firing. The additional parameters τ_{PRE}, τ_{POST} denote the decay time constants for A_{PRE}, A_{POST}, respectively.

The spike time indicators X_{PRE}, X_{POST} and the variables A_{PRE}, A_{POST} are used to implement an STDP-based evolution equation for the eligibility trace, which takes the form

$$\frac{dE}{dt} = (X_{POST}(t)A_{PRE}(t) - X_{PRE}(t)A_{POST}(t) - E)/\tau_E. \tag{5}$$

According to this equation, if a pre-synaptic neuron spikes and then its post-synaptic target follows, such that $A_{PRE} > 0$ and X_{POST} becomes non-zero, then the eligibility E increases, while if a post-synaptic spike occurs followed by a pre-synaptic spike, such that $A_{POST} > 0$ and X_{PRE} becomes non-zero, then E decreases. At times without spikes, the eligibility decays exponentially with rate τ_E.

In contrast to some previous work [22], we propose an update scheme for the synaptic weight $w(t)$ that depends on the type of MSN neuron involved in the synapse. It has been observed [29–32] that dMSNs tend to have less activity than iMSNs at resting states, consistent with our assumption that $p^D < p^I$, and are more responsive to phasic increases in dopamine than iMSNs. In contrast, iMSNs are largely saturated by tonic dopamine. In both cases, we assume that the eligibility trace modulates the extent to which a synapse can be modified by the dopamine level relative to a tonic baseline (which we without loss of generality take to be 0), consistent with previous models. Hence, we take $w(t)$ to change according to the equation

$$\frac{dw}{dt} = \alpha_w E f(K_{DA})(w^X_{max} - w), \tag{6}$$

where the function

$$f(K_{DA}) = \begin{cases} K_{DA}, & \text{if the target neuron is a dMSN,} \\ \dfrac{K_{DA}}{c + |K_{DA}|}, & \text{if the target neuron is an iMSN} \end{cases}$$

represents sensitivity to phasic dopamine, α_w refers to the learning rate, K_{DA} denotes the level of dopamine available at the synapses, $w^X_{\max}$ is an upper bound for the weight w that depends on whether the postsynaptic neuron is a dMSN ($X = D$)

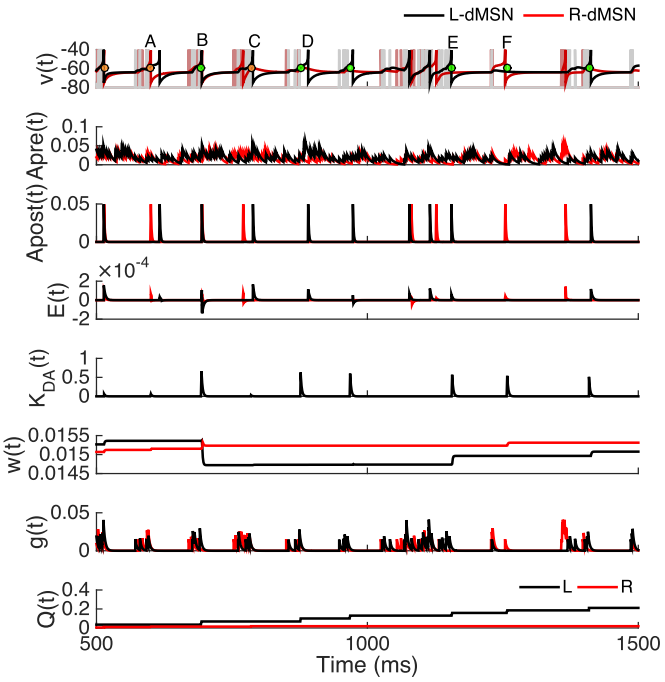

Fig. 2. Evolution of the learning rule variables. Learning-related variables were computed for dMSNs in each action channel, one promoting the *L* action (black, actual reward value 0.7) and one promoting the *R* action (red, actual reward value 0.1). Each panel represents corresponding variables for both neurons except $K_{DA}(t)$, which is common across all neurons. For each example neuron, the top panel shows its membrane potential (dark trace) and the cortical spike trains it receives (light trace with many spikes). This panel also represents the action onset times: green and orange dots if actions *L* and *R* occur, respectively. Different example cases labeled with letters (A,B,C,D,E,F) are described in the text in Section 2.4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

or an iMSN ($X = I$), c controls the saturation of weights to iMSNs, and $|\cdot|$ denotes the absolute value function. Importantly, we follow past work and take $\alpha_w > 0$ for dMSNs and $\alpha_w < 0$ for iMSNs [22]. The form of f, chosen to be an odd function for simplicity, may underestimate iMSN sensitivity to decreases in dopamine (e.g., see [19]); because $\alpha_w < 0$ for iMSNs, this underestimation translates into a weakened increase in weights onto iMSNs, but we shall see that this is not an important factor in our simulation results.

The dopamine level K_{DA} itself evolves as

$$\frac{dK_{DA}}{dt} = \sum_i (DA_{inc}(t_i) - K_{DA})\delta(t_i) - K_{DA}/\tau_{DOP}, \tag{7}$$

where the sum is taken over the times $\{t_i\}$ when actions are performed, leading to a change in K_{DA} that we treat as instantaneous, and τ_{DOP} is the dopamine decay constant. The DA update value $DA_{inc}(t_i)$ depends on the performed action as follows:

$$\begin{aligned} DA_{inc}(t) &= r_i(t) - \max_i\{Q_i(t)\}, \\ Q_i(t+1) &= Q_i(t) + \alpha(r_i(t) - Q_i(t)), \end{aligned} \tag{8}$$

where $r_i(t)$ is the reward associated to action i at time t, $Q_i(t)$ is an estimate of the value of action i at time t such that $r_i(t) - Q_i(t)$ is the subtractive reward prediction error [33], and $\alpha \in [0, 1]$ is the value learning rate. This rule for action value updates and dopamine release resembles past work [19] but uses a neurally tractable maximization operation (see [34,35] and references therein) to take into account that reward expectations may be measured relative to optimal past rewards obtained in similar scenarios [36,37]. In fact, we obtained similar results without the max operation in Eq. 8, but with slower convergence time (data not shown). The evolution of these variables is illustrated in Fig. 2, which is discussed in more detail in Section 2.4.

2.3. Actions and rewards

Actions. Each dMSN facilitates performance of a specific action. We specify that an action occurs, and so a decision is made by the model, when at least three different dMSNs of the same population spike in a small time window of duration Δ_{DA}. This action selection rule is designed to reflect a high enough dMSN rate relative to the iMSN rate in some relevant time window. When this condition occurs, a reward is delivered and the dopamine level is updated correspondingly, impacting all synaptic weights in the network in a way that depends on eligibility as specified in Eq. (6). Then, the spike counting

and the initial window time are reset, and cortical spikes to all neurons are turned off over the next 50 ms before resuming again as usual.

We assume that iMSN activity within a population counters the performance of the action associated with that population [38]. We implement this effect by specifying that when an iMSN in a population fires, the most recent spike fired by a dMSN in that population is suppressed. Note that this rule need not contradict observed activation of both dMSNs and iMSNs preceding a decision [14], see Section 3. We also implemented a version of the network in which each iMSN spike cancels the previous spike from both MSN populations. Preliminary simulations of this variant gave similar results to our primary version but with slower convergence (results not shown).

For convenience, we refer to the action implemented by one population of neurons as "left" or L and the action selected by the other population as "right" or R.

Rewards. In our simulations, we present results from different reward scenarios. In one case, we use constant rewards, with $r_L = 0.7$ and $r_R = 0.1$. In tuning the model, we also considered a regime with reward switches: reward values were as in the constant reward case but after a certain number of actions occurred, the reward-action associations were exchanged. We consider two different switching cases: a single switch, changing the rewards after 20 L actions occurred; and multiple switches, changing the rewards after 15 preferred actions occurred. Finally, we tune and test the learning rule in more challenging contexts, where rewards for both choices could either excede or fall short of their expected values, by comparing model performance with previously obtained experimental data [39]. For this case we implemented probabilistic rewards: every time that an action occurs, the reward r_i is set to be 1 with probability p_i or 0 otherwise, $i \in \{L, R\}$, with $p_L + p_R = 1$ and $p_L > p_R$, keeping the action L as the preferred one. Specifically, we consider the three different cases of $p_L = 0.85$, $p_L = 0.75$, and $p_L = 0.65$ to allow comparison with previous results [39].

2.4. Example implementation

The algorithm for the learning rule simulations is found in Algorithm 1.

Algorithm 1 Dopamine plasticity algorithm.

First, generate cortical mother spike trains and extract daughter trains to be used as inputs to the MSNs from the mother trains.Next, while $t < t_{end}$,

1. use RK45, with step size $dt = 0.01\,ms$, to compute the voltages of the MSNs in the network at the current time t from Eqs.~(1) and (3),
2. determine which dMSNs and iMSNs have reached spike threshold and fired in the current time step
3. update the *action* condition by checking sequentially for the following two events:
   - if any iMSN neuron in population $i \in \{L, R\}$ spikes, then the most recent spike performed by any of the dMSNs of population i is cancelled;
   - for each $i \in \{L, R\}$, count the number of non-cancelled spikes of the dMSNs in the ith population inside a time window consisting of the last $\Delta_{DA}\,ms$; if at least n_{act} non-cancelled spikes have occurred in this window, then action i has occurred and we update DA_{inc} and Q_i according to Eq.~(8),
4. for each MSN, update the support for the delta function $X_{POST}(t)$ to time t if the cell fired a spike in the current time step or else take $X_{POST}(t) = 0$; similarly, for each daughter spike train, update the support for $X_{PRE}(t)$ to time t if a spike occurred in the time step or else take $X_{PRE}(t) = 0$,
5. use RK45, with step size $dt = 0.01\,ms$, to solve Eqs.~(4)-(6) for each synapse, along with Eq.~(7) shared by all synapses, yielding an update of DA and all synaptic weight levels; for neurons that received an input spike, update synaptic conductance using $g(t) = g(t) + w(t)$,
6. set $t = t + dt$.

Fig. 2 illustrates the evolution of all of the learning rule variables over a brief time window. Cortical spikes (thin straight light lines, top panel) can drive voltage spikes of dMSNs (dark curves, top panel), which in turn may or may not contribute to action selection (green – for L – and orange – for R – dots, top panel). Each time a dMSN fires, its eligibility trace will deviate from baseline according to the STDP rule in Eq. (5). In this example, the rewards are $r_L = 0.7$ and $r_R = 0.1$, such that every performance of L leads to an appreciable surge in K_{DA}, with an associated rise in Q_L, but performances of R do not cause such large increases in K_{DA} and Q_R.

Various time points are labeled in the top panel of Fig. 2. At time A, R is selected. The illustrated R-dMSN fires just before this time and hence its eligibility increases. There is a small increase in K_{DA} leading to a small increase in the w for this dMSN. At time B, L is selected. Although it is difficult to detect at this resolution, the illustrated L-dMSN fires just after the action, such that its E becomes negative and the resulting large surge in K_{DA} causes a sizeable drop in w_L. At time C, R is selected again. This time, the R-dMSN fired well before time C, so its eligibility is small, and this combines with the small K_{DA} increase to lead to a negligible increase in w_R. At time D, action L is selected but the firing of the L-dMSN is sufficiently late after this that no change in w_L results. At time E, L is selected again. This time, the L-dMSN fires just before the action

leading to a large eligilibity and corresponding increase in w_L. Finally, at time F, L is selected. In this instance, the R-dMSN fired just before selection and hence is eligible, causing w_R to increase when K_{DA} goes up. Although this weight change does not reflect correct learning, it is completely reasonable, since the physiological synaptic machinery has no way to know that firing of the R-dMSN did not contribute to the selected and rewarded action L.

2.5. Learning rule parameters

The learning rule parameters have been chosen to capture various experimental observations, including some differences between dMSNs and iMSNs. First, it has been shown that cortical inputs to dMSNs yield more prolonged responses with more action potentials than what results from cortical inputs to iMSNs [40]. Moreover, dMSNs spike more than iMSNs when both types receive similar cortical inputs [41]. Hence, the effective weights of cortical inputs to dMSNs should be able to become stronger than those to iMSNs, which we encode by selecting $w^D_{\max} > w^I_{\max}$. This choice is also consistent with the observation that dMSNs are more sensitive to phasic dopamine than are iMSNs [29–32]. On the other hand, the baseline firing rates of iMSNs exceed the baseline of dMSNs [42], and hence we take the initial condition for $w(t)$ for the iMSNs greater than that for the dMSNs.

The relative values of other parameters are largely based on past computational work [22], albeit with different magnitudes to allow shorter simulation times. The learning rate α_w for the dMSNs is chosen to be positive and larger than the absolute value of the negative rate value for the iMSNs. The parameters Δ_{PRE}, Δ_{POST}, τ_E, τ_{PRE}, and τ_{POST} have been assigned the same values for both types of neurons, keeping the relations $\Delta_{PRE} > \Delta_{POST}$ and $\tau_{PRE} > \tau_{POST}$. Finally, the rest of the parameters have been adjusted to give reasonable learning outcomes. This tuning was done by hand. Because we have a fairly complete understanding of the mechanisms that interact to produce the desired plasticity outcomes, we do not expect that there is a completely different parameter regime that would give similar results, except that a simple time rescaling would yield similar behavior but with a different simulation time needed. Our parameter investigations support this claim, but we did not check it systematically. As for local robustness, we systematically varied several of the key parameters in the model. We found that although there was significant variability from trial-to-trial, the average weight values and rates of action selection remained within approximately $\pm 5\%$ of those obtained for our baseline parameter values over the following ranges: $\alpha^D_w \in [60, 80]$, $\alpha^I_w \in [-75, -35]$, $\tau_E \in [1, 9]$, $\Delta_{DA} \in [3, 10]$. In some cases, the robustness extended to the edge of the parameter range explored, so the actual robust range could be even broader.

Parameter values. We use the following parameter values in all of our simulations: $\tau_{DOP} = 2\,ms$, $\Delta_{DA} = 6\,ms$, $\tau_g = 3\,ms$, $\alpha = 0.05$ and $c = 2.5$. For both dMSNs and iMSNs, we set $\Delta_{PRE} = 10$ (instead of $\Delta_{PRE} = 0.1$; [22]), $\Delta_{POST} = 6$ (instead of $\Delta_{POST} = 0.006$; [22]), $\tau_E = 3$ (instead of $\tau_E = 150$; [22]), $\tau_{PRE} = 9$ (instead of $\tau_{PRE} = 150$; [22]), and $\tau_{POST} = 1.2$ (instead of $\tau_{POST} = 3$; [22]). Finally, $\alpha_w = \{80, -55\}$ (instead of $\alpha_w = \{12, -11\}$; [22]) and $w_{\max} = \{0.1, 0.03\}$ (instead of $w_{\max} = \{0.00045, 0\}$; [22]), where the first value refers to dMSNs and the second to iMSNs. Note that different reward values, r_i, were used in different types of simulations, as explained in the associated text.

Learning rule initial conditions. The initial conditions used to numerically integrate the system are $w = 0.015$ for weights of synapses to dMSNs and $w = 0.018$ for iMSNs, with the rest of the variables relating to value estimation and dopamine modulation initialized to 0.

2.6. Definitions of quantities computed from the STDP model

Averaged population firing rate. We compute the firing rate of a neuron by adding up the number of spikes the neuron fires within a time window and dividing by the duration of that window. The averaged population firing rate is computed as the average of all neurons' firing rates over a population, given by

$$\left\langle \frac{\sum_i s_i}{\Delta_t} \right\rangle_n$$

where Δ_t is the time window in ms, s_i is the spike train corresponding to neuron i, and $\langle \cdot \rangle_n$ denotes the mean over the n neurons in the population, The time course of the population firing rate is computed this way, using a disjoint sequence of time windows with $\Delta_t = 500\,ms$.

Action frequency. We compute the rate of a specific action i in a small window of $\Delta_t = 500\,ms$ as the number of occurrences of action i within that window divided by Δ_t.

Mean behavioral learning curves across subjects. The behavioral learning curves indicate, as functions of trial number, the fraction of trials on which the more highly rewarded action is selected. Within a realization, using a sliding trial count window of 5 trials, we computed the fraction of preferred actions selected (number of preferred actions divided by the total number of actions). Then we averaged over N realizations.

Evolution of the mean (across subjects) difference in model-estimated action values. Using N different realizations (simulating subjects in a behavioral experiment), we computed the difference of the expected reward of action L and the expected reward of action R at the time of each action selection (that is, $Q_L(t^*) - Q_R(t^*)$, where t^* is the time of action selection). Notice that $Q_i(t^*)$, for $i \in \{L, R\}$, only changes when an action occurs. Moreover, to average across realizations, we only considered the action number rather than the action onset time.

3. Results

To evaluate how dopaminergic plasticity impacts the efficacy of corticostriatal synapses, we modeled learning using a spike timing-dependent plasticity (STDP) paradigm in a simulation of corticostriatal networks implementing a simple two-alternative forced choice task. In this scenario, one of two available actions, which we call left (L) and right (R), was selected by the spiking of model striatal medium spiny neurons (MSNs; Section 2.3). These model MSNs were grouped into action channels receiving inputs from distinct cortical sources (Fig. 1). Every time an action was selected, dopamine was released, after a short delay, at an intensity proportional to a reward prediction error [Eqs. (7) and (8)]. All neurons in the network experienced this non-targeted increase in dopamine, emulating striatal release of dopamine by substantia nigra pars compacta neurons, leading to plasticity of corticostriatal synapses [Eq. (6); see Fig. 2].

The model network was initialized so that it did not *a priori* distinguish between L and R actions. We first performed simulations in which a fixed reward level was associated with each action, with $r_L > r_R$, to assist in parameter tuning and verify effective model operation. Next, we continued with the constant reward scenario but with reward values exchanged (i.e., L becomes the non-preferred action while R becomes the most rewarded one) after a certain number of actions, to see if the network is capable of learning that the reward switch has occurred and representing how long it lasts. We finally turn to results obtained with probabilistic rewards, as described in the last paragraph in Section 2.3, to compare with data from experiments with human subjects.

3.1. Constant rewards scenario

In this first scenario, where $r_L = 0.7$ and $r_R = 0.1$, a gradual change in corticostriatal synaptic weights occurred (Fig. 3A) in parallel with the learning of the actions' values (Fig. 3B).

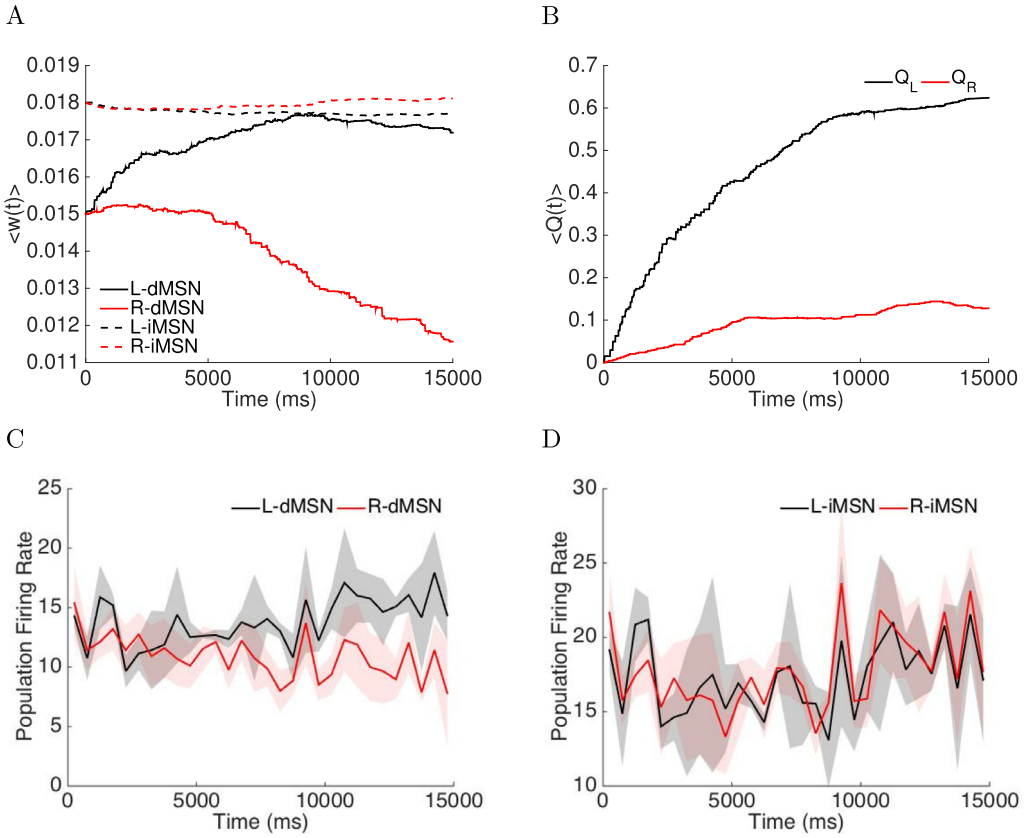

Fig. 3. STDP simulations with constant reward feedback. Time courses of corticostriatal synapse weights and firing rates are shown for simulations performed with a constant (i.e., non-probabilistic) reward schedule, where ($r_L(t) = 0.7$ and $r_R(t) = 0.1$). A: Averaged weights over 7 different realizations and over each of the four specific populations of neurons, which are dMSN selecting action L (solid black); dMSN selecting action R (solid red); iMSN countering action L (dashed black); iMSN countering action R (dashed red). B: Averaged evolution of the action values Q_L (black trace) and Q_R (red trace) over 7 different realizations. C: Firing rates of the dMSN populations selecting actions L (black) and R (red) over time. D: Firing rates of the iMSN populations countering actions L (black) and R (red) over time. Data in C,D was discretized into 50 ms bins. The transparent regions depict standard deviations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

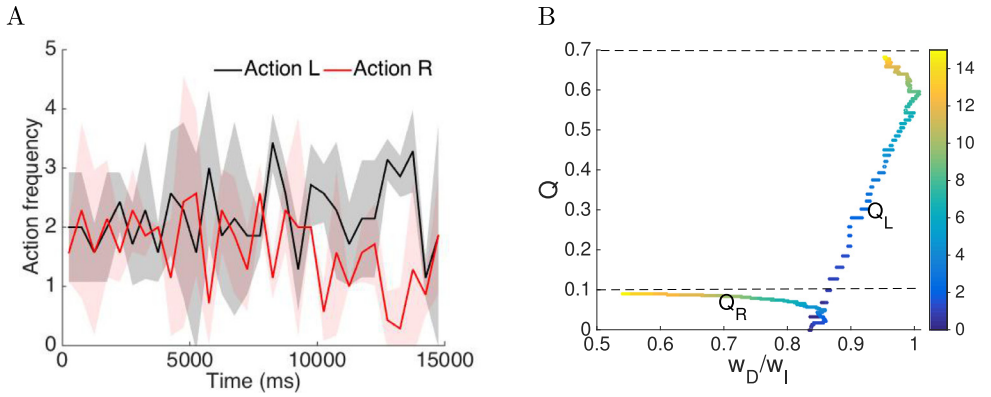

Fig. 4. Constant reward task. A: Frequency of performance of L (black) and R (red) actions is plotted over time (discretized each 50 ms) when the rewards are held constant ($r_L = 0.7, r_R = 0.1$). Both traces are averaged across 7 different realizations. The transparent regions depict standard deviations. B: Estimates of the values of L (Q_L) and R (Q_R) versus the ratio of the corticostriatal weights to those dMSNs that facilitate the action and those iMSNs that interfere with the action. Each trajectory is colored to show the progression of time; each trajectory evolves from the blue points through the green to the yellow. Even without full convergence of the action values Q_R and Q_L to their respective actual reward levels (B), a clear separation of action selection rates emerges (A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These changes in synaptic weights induced altered general MSN firing rates (Fig. 3C,D), reflecting changes in the sensitivity of the MSNs to cortical inputs in a way that allowed the network to learn over time to select the more highly rewarded action (Fig. 4A). That is, firing rates in the dMSNs associated with the more highly rewarded action increased, leading to a more frequent selection of that action. On the other hand, firing rates of the iMSNs remained quite similar (Fig. 3C,D). This similarity is consistent with recent experimental results [43], while the finding that dMSNs and iMSNs associated with a selected action are both active has also been reported in several experimental works [14–16].

In this model, indirect pathway activity counters action selection by cancelling direct pathway spiking (Section 2.3). Based on this cancellation, the ratio of direct pathway weights to indirect pathway weights provides a reasonable representation of the extent to which each action is favored or disfavored. In Fig. 4B, we show how this ratio evolves in parallel with value learning. Here the color code denotes time, and evolution progresses from the blue starting point to the yellow zone corresponding to the end of our simulations. In our simulations, after a long period of gradual evolution of weights and action values, the direct pathway versus indirect pathway weight ratio of the channel for the less favored action started to drop more rapidly, indicating the emergence of certainty about action values and a clearer separation between frequencies with which the two actions were selected (Fig. 4).

3.2. Reward switching scenario

To test whether the network remains flexible after learning a specific action-value relation, we ran additional simulations using a variety of reward schedules in which the reward values associated with the two actions were swapped after the performance of a certain number of actions.

We first performed a simulation in which the rewards associated with the L and R actions were switched only one time after 5 s. In Fig. 5, we can see that when the L-action is rewarded (up to time $t \approx 5s$), the firing rate, action frequency and the action values $Q(t)$ for the L-dMSNs become higher than those for the R-dMSNs, showing a learning of the L action. Up to time 5 s, 20 L actions have been performed and the learning is almost consolidated, since $Q_L(t)$ and $Q_R(t)$ are close to the actual reward values, $r_L = 0.7$ and $r_L = 0.1$, respectively (see top panel of Fig. 5).

At this time, the reward values are swapped. Afterwards, the network is able to learn that the R action elicits the preferable reward. Specifically, as we can see in the top panel of Fig. 5, Q_L and Q_R begin evolving toward the new reward levels, switching their relative magnitudes relatively quickly (i.e., in less than 3s) along the way. Although the weights of corticostriatal synapses to L-dMSNs (R-dMSNs) correspondingly weaken (strengthen), it takes longer, at least 5s, until the R action is reliably performed more frequently than the L action. Thus, the network is able to overcome previously learned contingencies and adaptively learn new ones, yet there is a delay relative to the learning that occurs without the previous bias.

On the other hand, given that the network is capable of learning new optimal actions after the switch, we also wanted to see what happens if the rewards are swapped back and forth before the new learning is consolidates. In Fig. 6, we plot the results of a simulation where the reward values are switched each time that 15 preferred actions take place.

In Fig. 6, after each switch, the estimated action value $Q(t)$ for the (now) non-preferred action starts to gradually fall off, causing small decreases in the weights of synapses to dMSNs associated with that action and in the mean firing rate of the dMSNs. At the same time, the weights of synapses onto the dMSNs that allow the (now) preferred action and their firing rates increase. In contrast to the action value estimates, which switch quickly, the STDP rule yields a delay in switching of relative synaptic weight values onto dMSNs (top panel of Fig. 6), such that the reward changes are not clearly encoded in

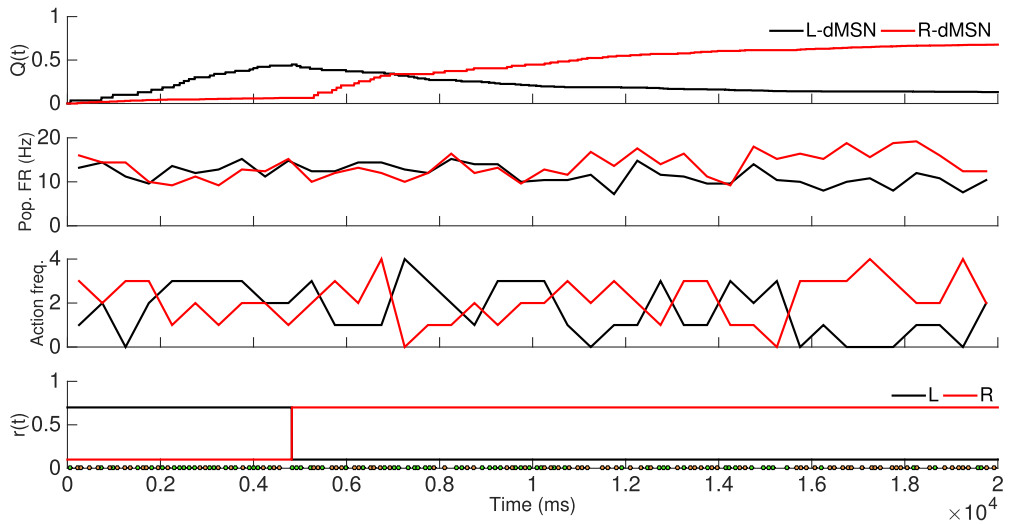

Fig. 5. STDP effects following change in action-value associations. The STDP model was simulated on a task in which the associated rewards for *L* and *R* actions are flipped after an initial learning period. The first three panels represent, from top to bottom, the action values ($Q(t)$), the firing rates of dMSN neurons for each action (*L*, black; *R*, red), and the action frequency for the dMSN population of neurons that produces the *L* action (black) and the *R* action (red). The bottom panel represents the actual reward values for *L* (black) and *R* (red). The reward values switch when 20 *L* actions have occurred. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

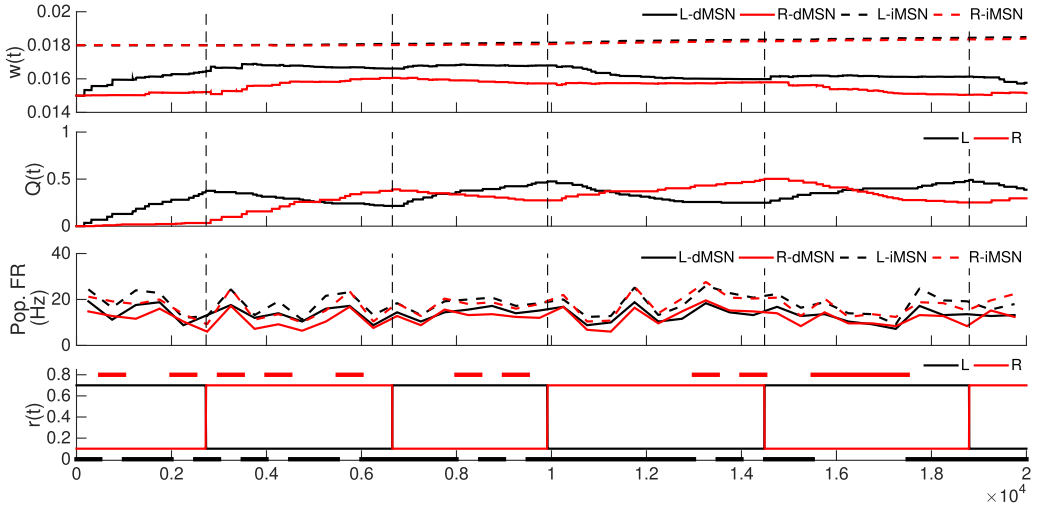

Fig. 6. STDP effects after repeatedly switching the highest rewarded action. The STDP model was simulated on a task in which the associated rewards for *L* and *R* actions are flipped each time that 15 preferred actions have occurred. The first three panels represent, from top to bottom, the weights ($w(t)$) for the different populations averaged over the number of neurons in each population, the estimates of the action values ($Q(t)$), and the firing rates for the different populations averaged over the neurons in each population. The bottom panel represents the actual reward values (thin curves) and the time intervals where an specific action has higher frequency (thick curves). In all panels, black traces refer to the left (*L*) action while red traces indicate the right (*R*) action. In the weight and firing rate plots, the solid lines refer the dMSN neurons while dashed lines refer to iMSN neurons. Vertical dashed lines indicate the times when reward values are switched. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the action selection outcomes (bottom panel of Fig. 6). As weights come closer before the next switch, action selection rates equalize. These results illustrate the lag in the STDP rule, which is advantageous for avoiding changes in action policy due to occasional spurious outcomes but requires repeated exposure to learn new reward contingencies, and suggest that some other plasticity mechanism is likely involved in more rapid or one-shot learning.

3.3. Probabilistic rewards scenario

While our previous simulations show that applying a dopaminergic plasticity rule to corticostriatal synapses allows for a simple network to learn action values linked to reward magnitude, many reinforcement learning tasks rely on estimating

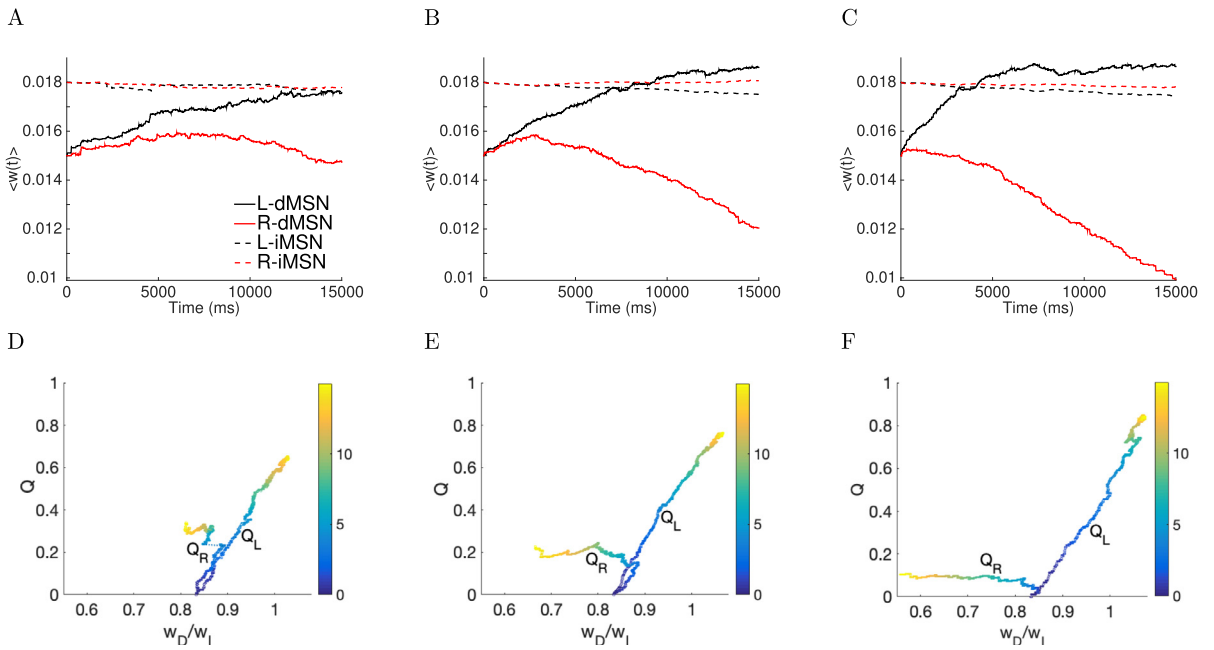

Fig. 7. Corticostriatal synaptic weights with probabilistic reward feedback. First column: $p_L = 0.65$; second column: $p_L = 0.75$; third column: $p_L = 0.85$ case. A, B, and C: Averaged weights over each of four specific populations of neurons, which are dMSN neurons selecting action L (solid black); dMSN neurons selecting action R (solid red); iMSN neurons countering action L (dashed black); iMSN neurons countering action R (dashed red). D, E, and F: Evolution of the estimates of the value L (Q_L) and R (Q_R) versus the ratio of the corticostriatal weights to those dMSN neurons that facilitate the action versus the weights to those iMSN that interfere with the action. As in Fig. 4, each trajectory is colored to show the progression of time, evolving from the blue points through the green to the yellow. Both the weights and the ratios have been averaged over 8 different realizations. The jump in Q_R for $p_L = 0.65$, joined by a horizontal dashed line, comes from the time discretization and averaging. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reward probability (e.g., two-armed bandit tasks). To evaluate the network's capacity to learn from probabilistic rewards, we simulated a variant of a probabilistic reward task and compared the network performance to previous experimental results on action selection with probabilistic rewards in human subjects [39]. For consistency with experiments, we always used $p_L + p_R = 1$, where p_L and p_R were the probabilities of delivery of a reward of size $r_i = 1$ when actions L and R were performed, respectively. Moreover, as in the earlier work, we considered the three cases $p_L = 0.65$ (high conflict), $p_L = 0.75$ (medium conflict) and $p_L = 0.85$ (low conflict).

As in the constant reward case, the corticostriatal synaptic weights onto the two dMSN populations clearly separated out over time (Fig. 7). The separation emerged earlier and became more drastic as the conflict between the rewards associated with the two actions diminished, i.e., as reward probabilities became less similar. Interestingly, for relatively high conflict, corresponding to relatively low p_L, the weights to both dMSN populations rose initially before those onto the less rewarded population eventually diminished. This initial increase likely arises because both actions yielded a reward of 1, leading to a significant dopamine increase, on at least some trials. The weights onto the two iMSN populations remained much more similar. One general trend was that the weights onto the L-iMSN neurons decreased, contributing to the bias toward action L over action R.

In all three cases, the distinction in synaptic weights translated into differences across the dMSNs' firing rates (Fig. 8, first row), with L-dMSN firing rates (D_L) increasing over time and R-dMSN firing rates (D_R) decreasing, resulting in a greater difference that emerged earlier when p_L was larger and hence the conflict between rewards was weaker. Notice that the D_L firing rate reached almost the same value for all three probabilities. In contrast, the D_R firing rate tended to decrease more over time as the conflict level decreased. As expected, based on the changes in corticostriatal synaptic weights, the iMSN population firing rates remained similar for both action channels, although the rates were slightly lower for the population corresponding to the action that was more likely to yield a reward (Fig. 8F).

Similar trends across conflict levels arose in the respective frequencies of selection of action L. Over time, as weights to L-dMSN neurons grew and their firing rates increased, action L was selected more often, becoming gradually more frequent than action R. Not surprisingly, a significant difference between frequencies emerged earlier, and the magnitude of the difference became greater, for larger p_L (Fig. 9).

To show that this feedback learning captured experimental observations, we performed additional probabilistic reward simulations to compare with behavioral data in forced-choice experiments with human subjects [39]. Each of these simulations represented an experimental subject, and each action selection was considered as the outcome of one trial performed by that subject. After each trial, a time period of 50 ms was imposed during which no cortical inputs were sent to striatal

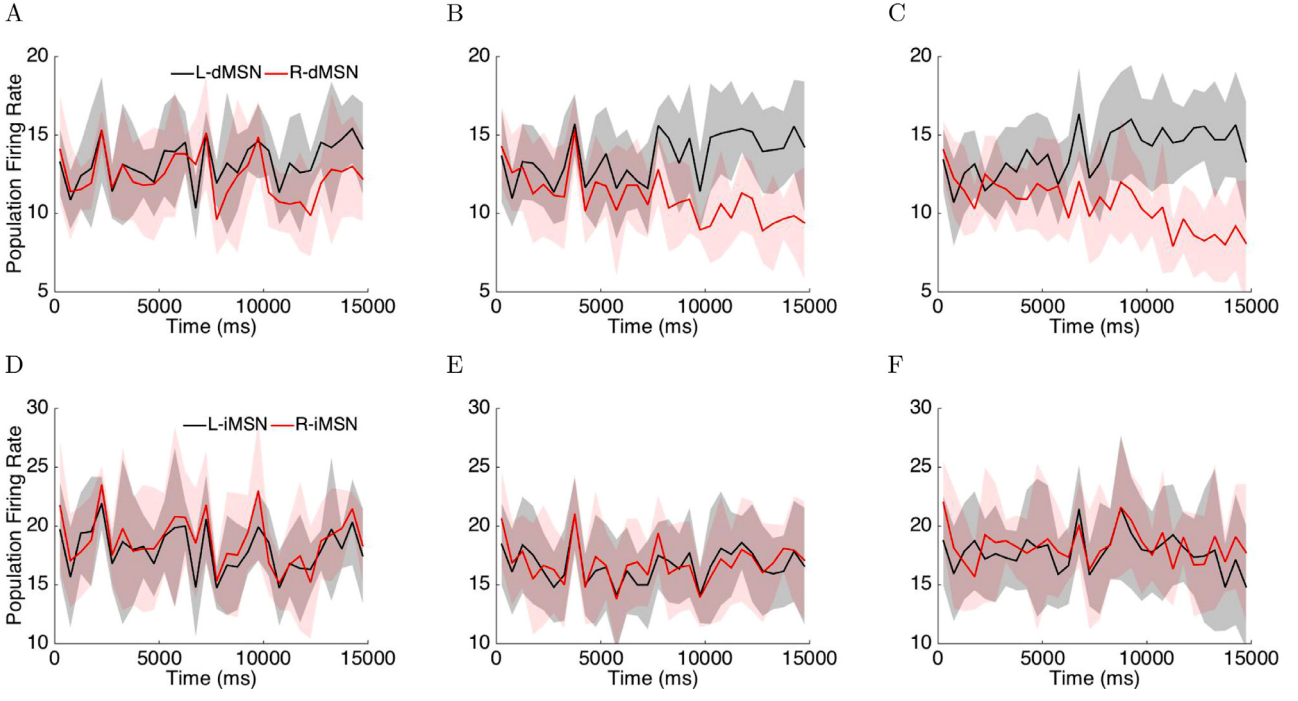

Fig. 8. Firing rates when the reward traces are probabilistic. First column: $p_L = 0.65$; second column: $p_L = 0.75$; third column: $p_L = 0.85$ case. A, B and C: Time courses of firing rates of the dMSNs selecting the L (black) and R (red) actions (50 ms time discretization). D, E, and F: Time courses of firing rates of the iMSNs countering the L (black) and R (red) actions (50 ms time discretization). In all cases, we depict the mean averaged across 8 different realizations, and the transparent regions represent standard deviations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

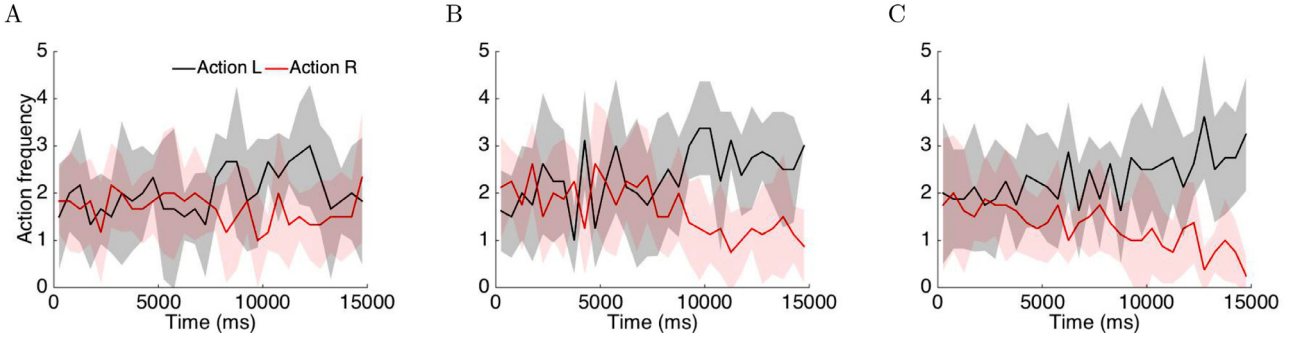

Fig. 9. Action frequencies when reward delivery is probabilistic. All panels represent the number of L (black) and R (red) actions performed across time (discretized each 50 ms) when action selection is rewarded with probability $p_L = 0.65$ (A), $p_L = 0.75$ (B), or $p_L = 0.85$ (C) with $p_L + p_R = 1$. Traces represent the means over 8 different realizations, while the transparent regions depict standard deviations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

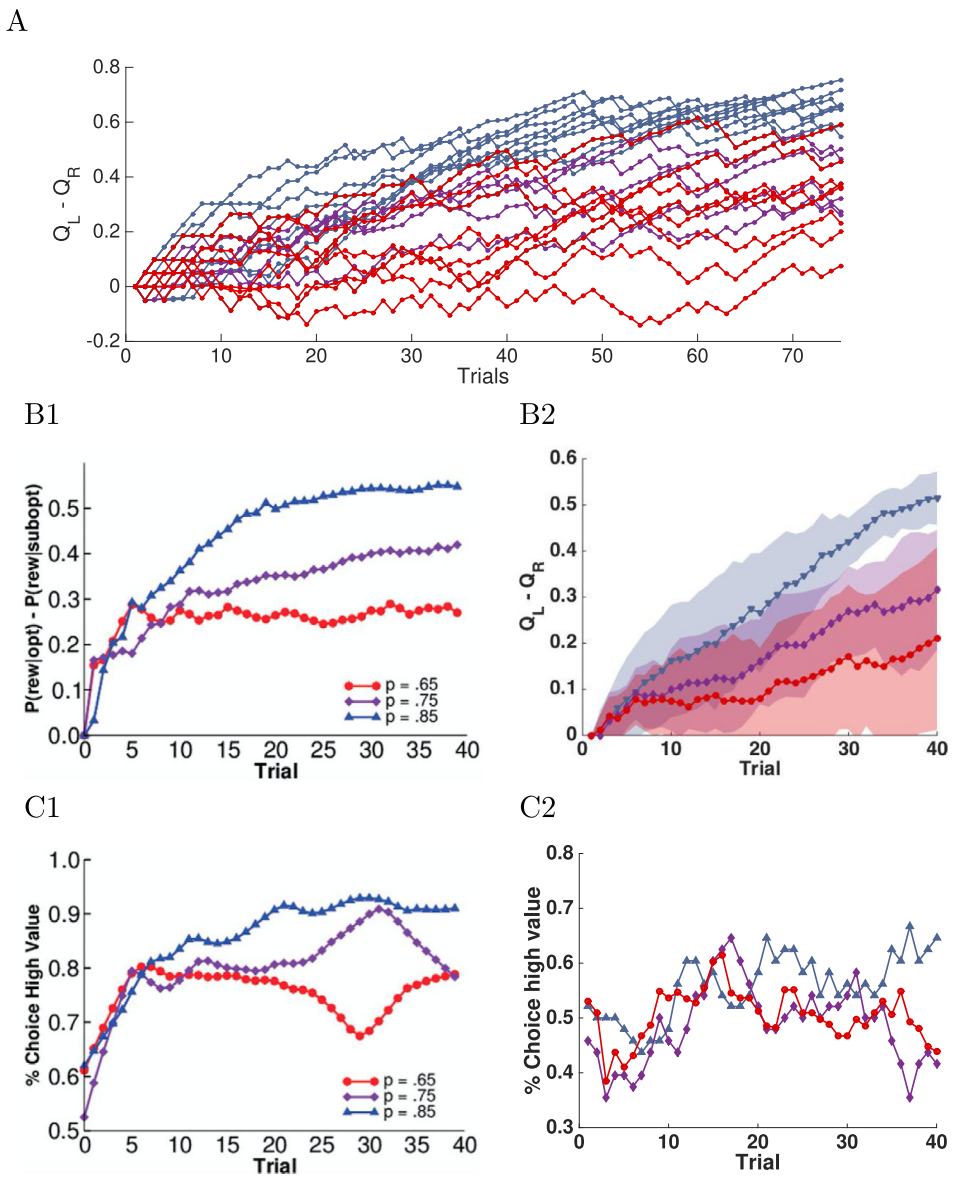

Fig. 10. Relative action value estimates and action selection probabilities. Action value and selection probabilities were estimated over simulated trials given probabilistic reward schedules, with $p_L = 0.65$ (red), $p_L = 0.75$ (purple), $p_L = 0.85$ (blue) and $p_L + p_R = 1$. A: Difference in action value estimates over trials in a collection of individual simulations. B: Means and standard deviations of difference in action value estimates across 8 simulations. C: Percent of trials on which the L action with higher reward probability was selected. B1 and C1 are the results obtained by Frank et al. in [39], collected from 15 human subjects. B2 and C2 are the results obtained using our learning rule, combined across 8 simulations. Note that the y-axis scales differ between panels C1 and C2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

neurons such that no actions would be selected, and then the full simulation resumed. For these simulations, we considered the evolution of the value estimates for the two actions either separately for each subject (Fig. 10A) or averaged over all subjects experiencing the same reward probabilities (Fig. 10B2), as well as the probability of selection of action L averaged over subjects (Fig. 10C2).

The mean in the difference between the action values gradually tended toward the difference between the reward probabilities for all conflict levels. Although convergence to these differences was generally incomplete over the number of trials we simulated (matched to the experiment duration), these differences appear to be close to the actual values for many individual subjects as well as in mean (Fig. 10A,B2). Indeed, visual inspection shows that the levels obtained in our simulations agree well with the behavioral data in [39] (Fig. 10B1) obtained from 15 human subjects, as well as with observations from similar experiments with rats [44], although the initial slopes of our trajectories (Fig. 10B2) underestimate those from the experimental data (Fig. 10B1).

Also as in the behavioral experiments, the probability of selection of the more rewarded action grew across trials for all three reward probabilities and then saturated, with less separation in action selection probability than in action values across different reward probability regimes (Fig. 10C2). Although our actual values for the probabilities of selection of higher value actions did not reach the levels seen experimentally (Fig. 10C1), this likely reflected the non-biological action selection rule in our STDP model (see Section 2.3), and points to the need to incorporate more realistic action selection mechanisms in future work.

4. Discussion

Flexible control of behavior in dynamic environments requires using the outcomes of previous actions to guide how future sensory-driven actions are chosen. In this work, we use a computational model to study how plastic effects at the corticostriatal synapses could contribute to the adaptive decision-making process. In particular, we show how a simple, dopamine-mediated STDP rule can modulate the sensitivity of both dMSN and iMSN populations to cortical inputs in a way that allows for the model network to bias action selection to favor a target that is more likely to lead to a higher reward. The resulting regime emerges through modifications in the ratio of direct and indirect pathway corticostriatal weights within each action channel, a process that is dominated by changes in weights to dMSNs, and features coactivation of opposing dMSN and iMSN populations within the selected channel, as observed experimentally [45].

Experimental results suggest that adaptive decision processes involve a rather complex set of components, including corticostriatal spike timing-dependent plasticity modulated by reward-sensitive dopamine release and the interplay of direct and indirect pathway populations of striatal neurons. Computational modeling provides a way to integrate these complex effects and gain insights into the process of decision-making that can guide future experimental research. Indeed, it was theoretical reasoning that led to the idea of the eligibility trace for dopaminergic learning [24], in which “credit” is assigned to the neurons that contributed to an action and represents a necessary permissive factor for subsequent synaptic modification. This theoretical prediction was recently confirmed to exist experimentally [25,28]. Here we attempt something similar by using computational models that characterize how plasticity modulates interactions within CBGT pathways in order to generate testable predictions about behavior and physiological outcomes in various scenarios involving actions and rewards. Recent experimental results have, in fact, exposed subtleties in the relationship between the direct and indirect pathways, such as co-activation of dMSN and iMSN populations linked with a particular action [45], and the modeling we report here allows us to investigate the compatibility of such findings with the proposed reward-driven learning and action selection framework. This work represents an important first step in showing how the many elements in our model effectively function together to achieve feedback-driven learning and decision-making, which has provided us with important insights that we have already harnessed in a study of the mapping between striatal activity and cognitive concepts of decision-making [10,18,46] and that will be useful for parameter tuning in more complex future studies.

As our starting point, we considered a fairly detailed plasticity system together with a simple action selection rule based on dMSN and iMSN spiking within specified time windows. While this form of action selection is rudimentary, it does incorporate the crucial elements of MSN spike timing and direct/indirect pathway competition, with higher dMSN firing rates translating into more frequent action selection. While a more biologically plausible selection rule would be ideal, implementing eligibility in more complete simulations of basal cortico-basal-ganglia-thalamic circuits, with extensive spiking of multiple competing neural populations during the decision period, is a highly non-trivial challenge that has not yet been addressed in the literature. The use of a simple action selection rule allows us to sidestep this difficulty to achieve a proof of principle that (a) hypothesized plasticity mechanisms based on an RPE-related DA signal, STDP, and eligibility can indeed cause action selection to gradually favor more rewarded outcomes, whether deterministically or probabilistically rewarded, in a way that resembles data from human experiments, and (b) different levels of conflict between probabilistic reward scenarios lead to different ratios of direct and indirect pathway corticostriatal synaptic weights, predominantly through changes in direct pathway weights.

Even with a simple action selection rule, neither of these outcomes was necessarily preordained, and parameter tuning was needed to find a regime that matches experimental results while maintaining robustness to variations in parameter values. For example, it was not a given that the eligibility scheme we used would be compatible with effective learning. Value learning could have led to loss of significant DA signals too soon to allow action selection policy to adapt to available rewards. Chance selection of poorly rewarded outcomes early on could have led to strengthening of pathways favoring those actions (due to small but still positive reward prediction errors) and thus interfered with learning. Alternatively, a stalemate could have resulted: if a left action was selected early in the simulations with both L-dMSNs active (to promote the L action) and R-iMSNs active (to prevent the R action), then reward-evoked DA release would have strengthened synapses to L-dMSNs and weakened those to R-iMSNs. The former would have promoted more L selections, but the latter would have favored R. The outcome of this competition could not have been anticipated *a priori*. As it turned out, in the parameter regime that we identified, these outcomes were avoided, and instead early action selections were due to chance differences in cortical activity levels associated with the two actions. Stochastic elevations in cortical inputs associated with the left action, for example, drove both L-dMSNs and L-iMSNs to exhibit higher firing rates than their R counterparts. For left action selection, the dMSN activity in the left channel prevailed despite the iMSN interference, leading to L-dMSN potentiation and L-iMSN depression due to the STDP rule. Although this is a classic learning mechanism, we emphasize that it was not a foregone conclusion. Moreover, the fact that after plasticity, both dMSN and iMSN neurons associated with a selected action exhibit

elevated firing rates as seen experimentally [45] was not a direct, obvious result of model design. One limitation is that we did not perform a systematic parameter fitting procedure. In our extensive exploration of parameter values, we did not find another parameter regime that produced behavior in agreement with experimental observations, other than a time rescaling that produced similar results more slowly; importantly, we did observe local robustness of our results to variations in key parameters away from our selected values (see Methods).

The model that we consider incorporates many features from the previous literature. The overarching novelty in our work comes from (a) combining this full collection of model components, including action value updates, dopamine-sensitive spike-timing dependent corticostriatal plasticity, synaptic eligibility, and spiking iMSN and dMSN populations that implement action selection, all together in one unified model, and (b) considering action selection across several reward scenarios including settings with probabilistic rewards or with switching of action-reward dependencies. At a more specific level, we note that many aspects of our plasticity model are based on the work of Gurney et al. [21] and Baladron and Hamker [22]. The former work starts from plasticity rules defined for two different levels of DA (high and low) and then produces weight change rules for other DA levels based on interpolation of these extremes. In contrast, we utilize a more direct dependence of plasticity on DA, which significantly simplifies the implementation. We include a DA release with a magnitude that depends on the reward prediction error, which is not a new concept but is not done in [21,22] nor in other recent computational modeling of action selection and effects of dopamine [47,48]. These other papers differ from our work in additional ways as well. Specifically, Mandali et al. focus on the relation between synchrony in the subthalamo-pallidal circuit and the exploration-exploitation tradeoff; they do include plasticity in cortico-striatal synapses, but weight changes are directly proportional to reward prediction errors and do not take into account relative times of cortical and striatal spikes [47]. Sen-Bhattacharya et al. base their action selection modeling on work [5,49] that predates [21]; it includes deterministic dopaminergic modulation of overall synaptic strengths but not synapse-specific or spike-timing dependent plasticity, and it lacks consideration of synaptic eligibility. Finally, Topalidou et al. [50] consider a combined cognitive and motor decision-making model that involves both cortical and BG decision-making circuits. Their work includes simplified Hebbian plasticity of corticostriatal synapses based on reward prediction error without modeling of eligibility, spike-timing dependence, or a dopamine signal.

The learning of values associated with specific actions in our model is treated in a simple, standard way [see Eq. (8)]. The encoding of values in neuronal activity that arises outside of the basal ganglia, yet is accessible to certain basal ganglia components, is consistent with several experimental findings [51–53], although modeling the details of this encoding is outside of the scope of this work. For action values to be learned effectively, actions must be sampled. One interesting scenario arises in the constant reward case when the action that elicits a higher reward (i.e., the optimal action) is switched after learning is consolidated (Figs. 5 and 6). Our use of the maximum possible reward in computing reward prediction error [Eq. (8)] yields rapid changes in dopamine signals after reward switching. This leads to relatively quick adjustments in action values. But the downside of this rapid learning is that the magnitude of the dopamine signal rapidly decays, resulting in slow adjustments of weights and of action selection strategies (e.g,. Fig. 6). Thus, while the network is capable of learning to favor the new optimal action, this occurs with a longer learning time than for the initial, unbiased learning process.

In the probabilistic reward scenario, we compared our results with human experimental data obtained in previous work [39]. For different levels of conflict, or similarity of reward probabilities between the two actions, the mean in the difference between the values assigned to the actions in our model approaches the difference between the reward probabilities, in general qualitative agreement with the experimental results [39]. (Fig. 10A,B). On the other hand, the percentage of trials on which the higher reward action is selected is less directly related to the reward conflict level in both our simulations and the earlier experiments (Fig. 10C). Overall, we find that corticostriatal learning based on reward-related dopaminergic feedback is sufficient to capture the major trends in human performance in a two-alternative forced choice task with probabilistic rewards.

The plasticity model used here makes a very compelling case for how dopamine-mediated STDP at the corticostriatal synapses can naturally modulate both firing rates and action selection in a reinforcement learning context; however, there are several improvements that can and should be the focus of future studies. For example, as discussed above, while our model features rather detailed plasticity mechanisms that build on past modeling studies [21,22,24], we used a phenomenological action selection rule. We leave explicit modeling of the CBGT circuit, in the context of feedback-dependent learning and action selection, for other work [9,46]. In addition, our simulations here largely ignored the role of tonic dopamine in the action selection process [21,54–58], which could reflect motivation or other aspects of reward valuation in the action selection process [58,59]. Future variants of the current model should explore the influence of tonic dopamine so that we can study how both phasic and tonic dopamine mechanisms coexist and contribute to plasticity of corticostriatal connections and subsequent behavior. This addition may help to explain the role of dopamine D2 receptors in reversal learning [60]. Finally, our simulations here were limited to a very simple variant of the two-armed bandit task. While this task is a popular test of learning in the reinforcement learning literature [61], it has limited ecological validity in real world contexts. Future work should explore the performance of dopamine-mediated STDP rules in more complex decision-making contexts that involve more alternatives and more complicated changes in reward dynamics.

Declaration of Competing Interest

None.

Acknowledgments

CV is supported by the Ministerio de Economía, Industria y Competitividad (MINECO), the Agencia Estatal de Investigación (AEI), and the European Regional Development Funds (ERDF) through projects MTM2014-54275-P, MTM2015-71509-C2-2-R and MTM2017-83568-P (AE/ERDF,EU). JR received support from NSF awards DMS 1516288 (CRCNS), 1612913, and 1724240 (CRCNS). TV received support from NSF CAREER award 1351748. The research was sponsored in part by the U.S. Army Research Laboratory, including work under Cooperative Agreement Number W911NF-10-2-0022, and the views espoused are not official policies of the U.S. Government.

References

[1] Bellman RE. Dynamic programming, ser. Cambridge studies in speech science and communication. Princeton: Princeton University Press; 1957.
[2] Schultz W, Apicella P, Scarnati E, Ljungberg T. Neuronal activity in monkey ventral striatum related to the expectation of reward. J Neurosci 1992;12:4595–610.
[3] Reynolds JN, Wickens JR. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw 2002;15:507–21.
[4] Calabresi P, Picconi B, Tozzi A, Filippo MD. Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends Neurosci 2007;30:211–19.
[5] Humphries MD, Stewart RD, Gurney KN. A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. J Neurosci 2006;26:12921–42.
[6] Bogacz R, Gurney K. The basal ganglia and cortex implement optimal decision making between alternative actions. Neural Comput 2007;19:442–77.
[7] Balleine BW, Delgado MR, Hikosaka O. The role of the dorsal striatum in reward and decision-making. J Neurosci 2007;27:8161–5.
[8] Doya K. Modulators of decision making. Nat Neurosci 2008;11:410–16.
[9] Wei W, Rubin JE, Wang XJ. Role of the indirect pathway of the basal ganglia in perceptual decision making. J Neurosci 2015;35:4052–64.
[10] Dunovan K, Verstynen T. Believer-skeptic meets actor-critic: rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. Front Neurosci 2016;10:106.
[11] Nonomura S, Nishizawa K, Sakai Y, Kawaguchi Y, Kato S, Uchigashima M, Watanabe M, Yamanaka K, Enomoto K, Chiken S, Sano H, Soma S, Yoshida J, Samejima K, Ogawa M, Kobayashi K, Nambu A, Isomura Y, Kimura M. Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. Neuron 2018;99:1302–1314.e5.
[12] Mink JW. The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 1996;50:381–425.
[13] Nambu A, Tokuno H, Hamada I, Kita H, Imanishi M, Akazawa T, Ikeuchi Y, Hasegawa N. Excitatory cortical inputs to pallidal neurons via the subthalamic nucleus in the monkey. J Neurophysiol 2000;84:289–300.
[14] Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, Costa RM. Concurrent activation of striatal direct and indirect pathways during action initiation. Nature 2013;494:238–42.
[15] Tecuapetla F, Matias S, Dugue GP, Mainen ZF, Costa RM. Balanced activity in basal ganglia projection pathways is critical for contraversive movements. Nat Commun 2014;5:4315.
[16] Tecuapetla F, Jin X, Lima SQ, Costa RM. Complementary contributions of striatal projection pathways to action initiation and execution. Cell 2016;166:703–15.
[17] Parker JG, Marshall JD, Ahanonu B, Wu Y-W, Kim TH, Grewe BF, Zhang Y, Li JZ, Ding JB, Ehlers MD, et al. Diametric neural ensemble dynamics in parkinsonian and dyskinetic states. Nature 2018;557:177.
[18] Dunovan K, Lynch B, Molesworth T, Verstynen T. Competing basal ganglia pathways determine the difference between stopping and deciding not to go. Elife 2015;4:e08723.
[19] Mikhael JG, Bogacz R. Learning reward uncertainty in the basal ganglia. PLoS Comput Biol 2016;12:e1005062.
[20] Bariselli S, Fobbs W, Creed M, Kravitz A. A competitive model for striatal action selection. Brain Res 2019;1713:70–9.
[21] Gurney KN, Humphries MD, Redgrave P. A new framework for cortico-striatal plasticity: behavioural theory meets in vitro data at the reinforcement-action interface. PLOS Biol 2015;13:1–25.
[22] Baladron J, Nambu A, Hamker FH. The subthalamic nucleus-external globus pallidus loop biases exploratory decisions towards known alternatives: a neuro-computational study. Eur J Neurosci 2019;49:754–67.
[23] Miller R. Cortico-striatal and cortico-limbic circuits: a two-tiered model of learning and memory functions. In: Information processing by the brain: Views and hypotheses from a cognitive-physiological perspective; 1988. p. 179–98.
[24] Izhikevich E. Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex 2007;17:2443–52.
[25] Shindou T, Shindou M, Watanabe S, Wickens J. A silent eligibility trace enables dopamine-dependent synaptic plasticity for reinforcement learning in the mouse striatum. Eur J Neurosci 2019;49:726–36.
[26] Fourcaud-Trocmé N, Hansel D, van Vreeswijk C, Brunel N. How spike generation mechanisms determine the neuronal response to fluctuating inputs. J Neurosci 2003;23:11628–40.
[27] Kreitzer AC, Malenka RC. Striatal plasticity and basal ganglia circuit function. Neuron 2008;60:543–54.
[28] Fisher SD, Robertson PB, Black MJ, Redgrave P, Sagar MA, Abraham WC, Reynolds JNJ. Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. Nat Commun 2017;8:334.
[29] Dreyer JK, Herrik KF, Berg RW, Hounsgaard JD. Influence of phasic and tonic dopamine release on receptor activation. J Neurosci 2010;30:14273–83.
[30] Richfield E, Penney J, Young A. Anatomical and affinity state comparisons between dopamine d1 and d2 receptors in the rat central nervous system. Neuroscience 1989;30:767–77.
[31] Gonon F. Prolonged and extrasynaptic excitatory action of dopamine mediated by d1 receptors in the rat striatum in vivo. J Neurosci 1997;17:5972–8.
[32] Keeler J, Pretsell D, Robbins T. Functional implications of dopamine d1 vs. d2 receptors: a 'prepare and select' model of the striatal direct vs. indirect pathways. Neuroscience 2014;282:156–75.
[33] Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. Arithmetic and local circuitry underlying dopamine prediction errors. Nature 2015;525:243.
[34] Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 2007;10:1615–24.
[35] Kozlov AS, Gentner TQ. Central auditory neurons display flexible feature recombination functions. J Neurophysiol 2014;111:1183–9.
[36] Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 2012;482:85–8.
[37] Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 2006;9:1057–63.
[38] Roseberry TK, Lee AM, Lalive AL, Wilbrecht L, Bonci A, Kreitzer AC. Cell-type-specific control of brainstem locomotor circuits by basal ganglia. Cell 2016;164:526–37.
[39] Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D. fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. J Neurosci 2015;35:485–94.
[40] Flores-Barrera E, Vizcarra-Chacón B, Tapia D, Bargas J, Galarraga E. Different corticostriatal integration in spiny projection neurons from direct and indirect pathways. Front SystNeurosci 2010;4:15.

[41] Escande MV, Taravini IRE, Zold CL, Belforte JE, Murer MG. Loss of homeostasis in the direct pathway in a mouse model of asymptomatic parkinson's disease. J Neurosci 2016;36:5686–98.
[42] Mallet N, Ballion B, Moine CL, Gonon F. Cortical inputs and GABA interneurons imbalance projection neurons in the striatum of parkinsonian rats. J Neurosci 2006;26:3875–84.
[43] Donahue CH, Liu M, Kreitzer A. Distinct value encoding in striatal direct and indirect pathways during adaptive learning. bioRxiv 2018. doi:10.1101/277855.
[44] Tort ABL, Komorowski RW, Manns JR, Kopell NJ, Eichenbaum H. Theta–gamma coupling increases during the learning of item–context associations. Proc Natl Acad Sci 2009;106:20942–7.
[45] Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, Costa RM. Concurrent activation of striatal direct and indirect pathways during action initiation. Nature 2013;494:238.
[46] Dunovan K, Vich C, Clapp M, Verstynen T, Rubin J. Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. PLoS Comput Biol 2019;15:e1006998.
[47] Mandali A, Rengaswamy M, Chakravarthy VS, Moustafa AA. A spiking basal ganglia model of synchrony, exploration and decision making. Front Neurosci 2015;9:191.
[48] Sen-Bhattacharya B, James S, Rhodes O, Sugiarto I, Rowley A, Stokes AB, Gurney K, Furber SB. Building a spiking neural network model of the basal ganglia on spinnaker. IEEE Trans CognitDev Syst 2018;10:823–36.
[49] Gurney KN, Prescott TJ, Redgrave P. A computational model of action selection in the basal ganglia. i. a new functional anatomy. Biol Cybern 2001;84:401–10.
[50] Topalidou I, Cooper K, Pereira L, Ailion M. Dopamine negatively modulates the NCA ion channels in c. elegans. PLOS Genet 2017;13:1–31.
[51] Palminteri S, Boraud T, Lafargue G, Dubois B, Pessiglione M. Brain hemispheres selectively track the expected value of contralateral options. J Neurosci 2009;29:13465–72.
[52] Phelps EA, Sokol-Hessner P. Chapter 9 - social and emotional factors in decision-making: appraisal and value. In: Dolan R, Sharot T, editors. Neuroscience of preference and choice. San Diego: Academic Press; 2012. p. 207–23. doi:10.1016/B978-0-12-381431-9.00019-X.
[53] Wallis LJ, Viranyi Z, Müller CA, Serisier S, Huber L, Range F. Aging effects on discrimination learning, logical reasoning and memory in pet dogs. AGE 2016;38:6.
[54] Beeler JA, Daw N, Frazier CRM, Zhuang X. Tonic dopamine modulates exploitation of reward learning. Front Behav Neurosci 2010;4:170.
[55] Humphries MD, Khamassi M, Gurney K. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. Front Neurosci 2012;6:9.
[56] Schultz W. Dopamine reward prediction-error signalling: a two-component response. Nat Rev Neurosci 2016;17:183.
[57] Silva JAd, Tecuapetla F, Paixão V, Costa RM. Dopamine neuron activity before action initiation gates and invigorates future movements. Nature 2018;554:244.
[58] Berke JD. What does dopamine mean? Nat Neurosci 2018;21:787–93.
[59] Niv Y, Daw ND, Dayan P. How fast to work: response vigor, motivation and tonic dopamine. In: Advances in neural information processing systems; 2006. p. 1019–26.
[60] Kwak S, Huh N, Seo J-S, Lee J-E, Han P-L, Jung MW. Role of dopamine d2 receptors in optimizing choice strategy in a dynamic and uncertain environment. Front BehavNeurosci 2014;8:368.
[61] Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT Press; 1998.